Technical Specification

ISO/IEC TS 25058

Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems

Ingénierie des systèmes et des logiciels — Exigences et évaluation de la qualité des systèmes et des logiciels (SQuaRE) — Lignes directrices pour l'évaluation de la qualité des systèmes d'intelligence artificielle (IA)

First edition
2024-01

Reference number
ISO/IEC TS 25058:2024(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/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

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

Introduction

An artificial intelligence (AI) system can be challenging to evaluate. Consequently, the impact of an AI system with poor quality can be considerable since it can be developed to facilitate the automation of critical actions and decisions.

The purpose of this document is to guide AI developers performing a quality evaluation of their AI systems. This document does not state exact measurements and thresholds, as these vary depending on the nature of each system. Instead, it specifies comprehensive guidance that covers the relevant facets of an AI system's quality for successful quality evaluation.

Testing is within the scope as far as each characteristic and sub-characteristic is verified by testing strategies, but details of testing methods and measurements are covered elsewhere, for example in the ISO/IEC/IEEE 29119 series.

Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems

1 Scope

This document provides guidance for evaluation of artificial intelligence (AI) systems using an AI system quality model.

The document is applicable to all types of organizations engaged in the development and use of AI.

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 TS 4213, *Information technology — Artificial intelligence — Assessment of machine learning classification performance*

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

ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

ISO/IEC 25059:2023, *Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems*

ISO/IEC/IEEE 29119-1, *Software and systems engineering — Software testing — Part 1: General concepts*

ISO/IEC/IEEE 29148, *Systems and software engineering — Life cycle processes — Requirements engineering*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC TS 4213, ISO/IEC 22989, ISO/IEC 23053, ISO/IEC 25059, ISO/IEC/IEEE 29119-1 and ISO/IEC/IEEE 29148 apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

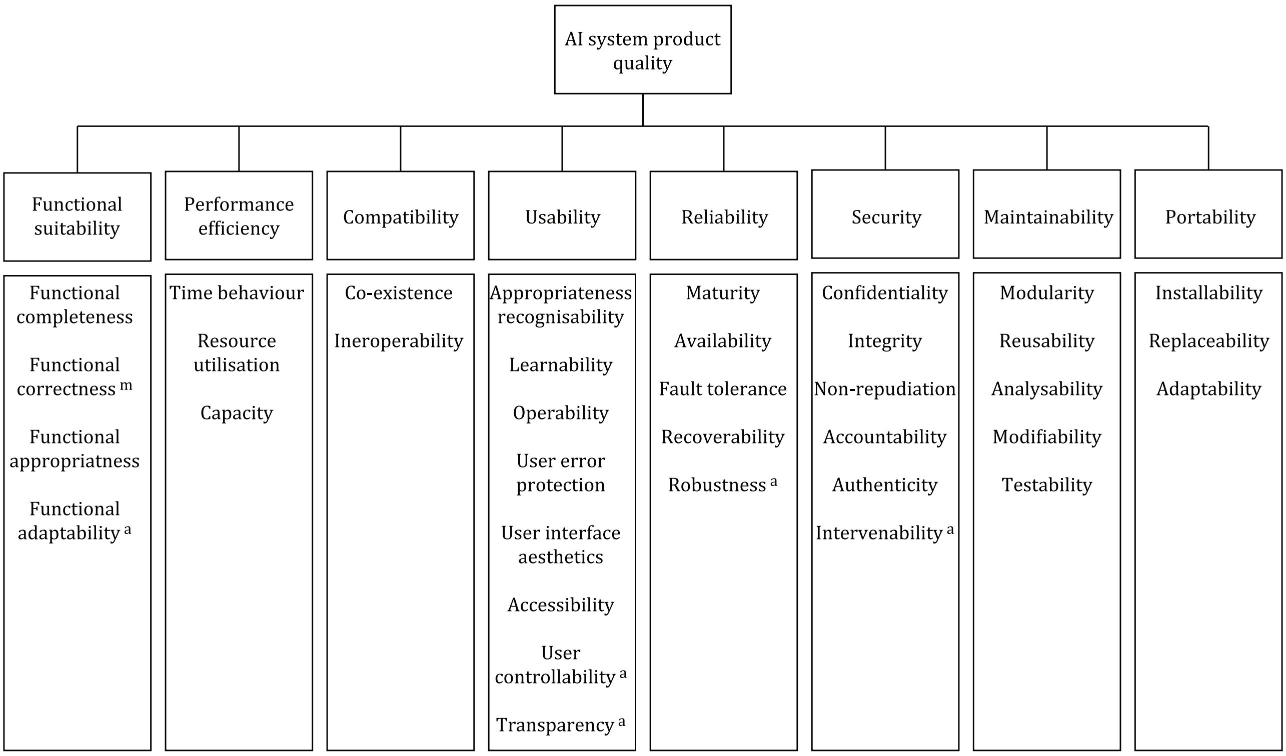
— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

4 Overview

To ensure that relevant facets of an AI system's quality are covered by the quality evaluation guidance, this document references Systems and software Quality Requirements and Evaluation (SQuaRE) product quality and quality in use models' characteristics for an AI system (see ISO/IEC 25059). The product quality and quality in use models' characteristics, as applicable to a general system, apply to an AI system. Several sub-characteristics have been added, and some have different meanings or contexts.

Figures 1 and 2 illustrate an AI system's product quality and quality in use models' characteristics and sub-characteristics. Please note that some sub-characteristics have been added or modified from the SQuaRE quality models for general systems, as an AI system differs from a general system and software.

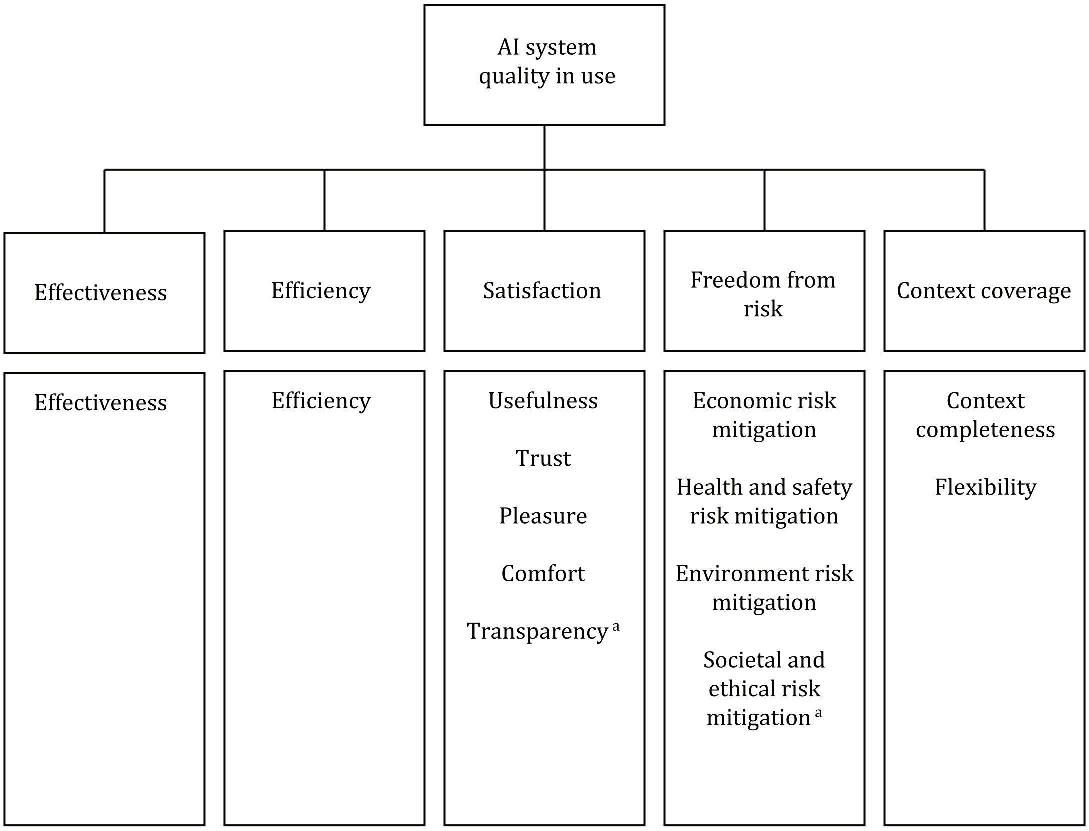


[a] New sub-characteristic.

[m] Modified sub-characteristic.

SOURCE ISO/IEC 25059:2023, Figure 1.

Figure 1 — AI system product quality model



[a] New sub-characteristic.

SOURCE ISO/IEC 25059:2023, Figure 2.

Figure 2 — AI system quality in-use model

5 Quality evaluation methodology

Quality evaluation guidance is defined by relevant quality model sub-characteristics.

All the sub-characteristics from the SQuaRE product quality and quality in use models are covered in this document.

The guidance in this document should complement the SQuaRE quality evaluation process described in ISO/IEC 25040 for AI systems.

6 Functional suitability

6.1 Functional completeness

Quality of the functional completeness sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.2.1.

6.2 Functional correctness

Quality of the functional correctness sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.2.2.

Functional correctness should be evaluated with the proper key performance indicators (KPIs) and measurements.

Measurements and key performance indicators should be established to measure the capability of an AI system to do a specific task and to evaluate the amount of unpredictability of the system.

The right evaluation measurements should be used to measure functional correctness based on an AI system's problem type and the stakeholders' objectives. For a list of typical evaluation measurements, refer to ISO/IEC 23053:2022, 6.5.5.

Functional correctness should also be evaluated using functional testing methods, such as:

- metamorphic testing: technique that establishes relationships between inputs and outputs of the system;
- expert panels: technique used when an AI system is built to replace the judgement of experts, which consists of establishing a panel to review the test results;
- benchmarking an AI system: technique used when an AI system is replacing existing approaches or when a similar AI system can be used as a benchmark;
- testing an AI system's behaviours against various scenarios or test cases defined by stakeholders;
- testing in a simulated environment: technique used when an AI system is characterized by physical action on the environment;
- field trials: technique used when there is a potential difference or evolution between testing environments and actual operation conditions;
- risk management: testing AI system behaviour against identified risk scenarios.

Functional correctness evaluation techniques should be performed on different and representative datasets.

The best machine learning (ML) model should be selected using the appropriate evaluation measurements against a validation dataset. The simple ML model validation technique uses only one validation dataset. However, a cross-validation technique is suggested when possible.

In a separate back-testing phase, the selected ML model should be tested once again with new data (the testing dataset) for consistency.

Training, validation and testing datasets should all be built with different data.

Validation and testing datasets should all be built with representative subsets of the actual operation conditions.

The ML model should be tested against datasets with known cohorts to identify positive or negative bias creep.

The final settings to tune the ML model (e.g. the cut-off threshold in classification) should be defined together with the business users.

The functional correctness should be evaluated on production data for monitoring purposes.

Product deployment should take place after the back-testing phase.

6.3 Functional appropriateness

The quality of the functional appropriateness sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.2.3.

6.4 Functional adaptability

An AI system should have a mechanism to adapt dynamically to changes in the production data, by using one of the following;

- deploying a continuous or reinforcement learning modelling approach;
- implementing an automated retraining workflow.

NOTE Functional adaptability does not necessarily adapt to changes of the system objectives as these changes potentially transform the functional state of the system.

The organization should develop an adaptation system to generate a feedback loop. This managed system comprises four essential functions: monitor; analyse; plan; execute.

- The monitor function tracks the managed system and the environment in which the system operates and updates the knowledge.
- The analyse function uses the up-to-date knowledge to evaluate the need for adaptation, exploiting rigorous analysis techniques or simulations of runtime ML models.
- The plan function selects the best option based on the adaptation goals and generates a plan to adapt the system from its current configuration to the new configuration.
- The execute function implements the adaptation actions of the plan with relevant intervention.

Functional adaptability should be evaluated using measurements, key performance indicators and functional testing methods, as documented in 6.2, to measure the adaptability of an AI system to a new dataset.

The organization should take into consideration resource trade-offs when selecting the best ML model for deployment, as the most accurate ML model can be prohibitively expensive to computationally evaluate. Refer to 7.2 for more details.

7 Performance efficiency

7.1 Time behaviour

Quality of the time behaviour sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.3.1.

The organization should calculate time behaviour during the training, evaluation and inference workflows under normal conditions as part of normal workflows in production, using the production environment, infrastructures and computing resources, as time behaviour depends on resource utilization. Refer to 7.2 for guidance on resource utilization.

The organization should consider an AI system adaptability mechanism while measuring the process duration. For example, a system that consists of a sequence of retraining, evaluation and inference should measure the duration of the entire sequence of workflows.

The organization should test and assess potential conflicts between computational resources. For example, if the training and inference workflows use the same computational resources or if multiple inferences happen simultaneously, this can negatively affect time behaviour of an AI system.

The organization should test and assess timing between data collection, data transformation and other data-dependent AI system workflows. For example, AI system inference cannot be processed if the required input data are not collected and transformed beforehand.

7.2 Resource utilization

Quality of the resource utilization sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.3.2.

The organization should allocate the appropriate resources during the training process of an ML-based AI system. Factors such as computational resource types, time requirements, data quantity, ML model type and the quantity of hyperparameters can impact the resources required to complete this process.

An AI system can also get its input from other AI systems, which should be considered as this can create additional dependencies on utilized resources.

The organization should consider available resources through a system adaptability strategy (e.g. incremental learning, active learning, online learning or retraining strategy). For example, an adaptability strategy consisting of retraining an AI system hourly takes few resources during inference but several resources during the retraining. Consider as part of this adaptability strategy that inference and retraining can happen simultaneously.

7.3 Capacity

Quality of the capacity sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.3.3.

8 Compatibility

8.1 Co-existence

Quality of the co-existence sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.4.1.

8.2 Interoperability

Quality of the interoperability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.4.2.

9 Usability

9.1 Appropriateness recognizability

Quality of the appropriateness recognizability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.5.1.

9.2 Learnability

Quality of the learnability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.5.2.

Additionally, measurements commonly used to evaluate learnability of an AI system refer to but are not limited to the following:

— Task performance, for example, learning how to input the variables needed in the user interface for an AI system to return a ML model prediction. The measurements evaluate the ability to perform a task in the expected time, percentage of users who completed the task, characteristics of the users who completed the task and others.

— Usage, for example, learning the commands available to operate an AI system. The measurements evaluate the success rate of the command use, the number of commands used in a time interval and others.

— Cognitive process, for example, the thinking time required for a user to understand an ML model inference.

- User feedback: The opinion of the user after learning or not learning to use an AI system. For example, does the system contain any counterintuitive parameters or functionalities?

The learnability measurements are susceptible to the tester or the user capabilities and experience. For example, users with a technical background or general AI understanding can learn how to perform a task and use commands in shorter times, while users with a non-technical background need a good user interface to reach the learnability goals. A diverse group of users is important to accurately evaluate learnability.

An evaluation of learnability measures can be made and can include the following:

- User testing: users can rate the degree to which they can learn how to use an AI system. For example, are the user guidelines clear? Do I understand the variables required to run a specific AI system? Are the default fields comprehensive and based on real scenarios? Do the outputs of an AI system correspond to the outputs described in the user guidelines?
- Other learnability evaluation tools: the learning curve, which is the relationship between time and task repetition, provides information on a) first-use learnability; b) how quickly users improve with repetition (steepness of the curve); c) how the productivity of the user can improve if they learn how to use the system appropriately (efficiency of the ultimate plateau).

9.3 Operability

Quality of the operability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.5.3.

Operability of an AI system can also be measured by the following:

- Peer review: technical peers evaluate an AI system based on individual experiences and lessons learnt. For example, an ML engineer can identify that the time between having an ML model inference and delivering the result to the user be improved. The ML engineer can suggest new ML model deployment options.
- User testing: users will be asked to rate the degree to which an AI system is meeting their expectations with efficiency and efficacy. For example, is the ML model providing accurate classification of the entries? Is it comparable to human classification?
- Other evaluation methods linked to the measurements of reusability and reliability: an AI system’s consistency when exposed to different environments and users can help meet the expectations of the user. For example, measuring inference delivery times of an ML-based AI system when exposed to different variables – users, processing power and others.

9.4 User error protection

Quality of the user error protection sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.5.4.

9.5 User interface aesthetics

Quality of the user interface aesthetics sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.5.5.

9.6 Accessibility

Quality of the accessibility sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.5.6.

9.7 User controllability

The organization should use the controllability framework according to ISO/IEC TS 8200 for an AI system designed to be controllable. For an evaluation, the following should be considered and prepared in advance:

— AI system that is runnable and with control functionalities implemented according to requirements;

— test items that are designed specifically for the test of control functionalities;

— toolkits capable of issuing controllability instructions, receiving or observing system appearance or internal parameter changes, as well as computing concerned measures (e.g. response latencies and stabilities).

The evaluation of an AI system's controllability should include the following:

— To check whether an AI system's control functionalities meet requirements, each of the required control functionalities, as well as their useful combinations or workflows, should be tested.

— To qualify the level of controllability of an AI system, specified control functionalities in the level where the system is designed, implemented or required to be in should be tested.

A specific set of test items should be prepared according to the evaluation objectives.

Sub-processes of control, such as control transfer, engagement and disengagement of control and uncertainty handling of control transfer, as well as the actual control, should be tested. For each control functionality, the test item can check the following. Stakeholders can apply a subset of these according to specific concerns:

— Correctness of a control functionality: an AI system that can learn or execute as expected when receiving correct control instruction and remain in its current state when receiving an incorrect control instruction.

— Duration of a control functionality: length of time needed by a sub-process or the entire process to complete a control functionality. The durations of important sub-processes (e.g. transfer of control and engagement of control) should be checked for an AI system that operates in risky environments.

— Reliability of a control functionality: degree to which the control functionality can behave consistently, particularly in those cases when system faults or external unpredicted incidences can happen. To test this, fault injection mechanisms can be selected and applied.

— Number of operations needed by a control functionality: total number of operations a controller should carry out, such that the entire control functionality process can complete and return results. All operations needed by all sub-processes count.

For further details, refer to ISO/IEC TS 8200.

9.8 Transparency

The presentation of information to stakeholders should be open, comprehensive and understandable.

The organization should be able to understand, trace and document all privacy-relevant data processing considerations, including the legal, technical and organizational.

An AI system should have clear owners who are accountable for meeting the expected benefits and communicating about the system's outcomes to the stakeholders.

The relevant characteristics used in an AI system should be comprehensive, accessible, clear and understandable to the stakeholders. For example, considering that only a finite number of variables are used as input to an AI system, this limitation should be communicated clearly to avoid misunderstandings.

The organization should communicate the risks of an AI system's output (i.e. predictions, decisions or activities) affecting society, the economy or the environment in a clear, accurate, timely, honest and complete manner.

The organization should communicate an AI system's output (i.e. predictions, decisions or activities) to relevant stakeholders in a comprehensive, accessible and understandable manner.

Appropriate information about an AI system and its level of quality should be made available to relevant stakeholders.

10 Reliability

10.1 Maturity

Quality of the maturity sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.6.1.

10.2 Availability

Quality of the availability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.6.2.

10.3 Fault tolerance

Quality of the fault tolerance sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.6.3.

10.4 Recoverability

Quality of the recoverability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.6.4.

10.5 Robustness

The robustness of an AI system should be measured under normal operational conditions. These conditions are dependent on the AI system's domain, as described in ISO/IEC 24029-2:2023, 5.2. The bounded domain should:

- be determined by a set of attributes that are clearly defined;
- be sufficient for the AI system to conduct one or more given tasks as intended;
- use training data that is representative of data expected to be used for inference.

The following actions should be considered to measure the AI system's robustness against the bounded domain:

- assess training, validation and testing dataset to ensure they cover normal operational conditions;
- develop specific test scenarios to test the system's performance under a wide range of normal operational conditions;
- use simulation as a test data generator to address the full range of operation;
- consider regularization techniques, data augmentation or introduction of random noise to maximize robustness of an AI system under normal operating conditions;
- evaluate functional correctness (see 6.2 for the recommended guidance on functional correctness and ISO/IEC TR 24029-1:2021, 5.2 for robustness metrics).

An AI system should endure long-tail, black swan and abnormal events. The following actions should be considered:

- perform test scenarios of long-tail and abnormal events, benchmarks and stress-test of an AI system;

- identify breaking points of an AI system;
- decide whether an AI system is robust enough for potential future needs based on the analysis results.

An AI system should handle diverse perceptible and unforeseen attacks. The following actions should be considered:

- Address failures and errors according to best practice, for example, through hardening, testing and verification of an AI system's stability using techniques such as metamorphic testing, data augmentation, generative adversarial networks, adversarial training, adversarial example generation or adversarial example detection.
- Apply specific countermeasures in the field of machine learning, such as anomaly detection of an AI system's input and output and flagging novel misuses to further mitigate functional risks.

An AI system should generate representative outputs under abnormal environmental conditions by:

- leveraging confidence score or confidence intervals to decide whether to act on the output generated or if a backup workflow should be triggered;
- calibrating confidence score or confidence intervals to be representative of the uncertainty relative to the output being true by using calibration measurements validated on testing data;
- integrating a backup workflow, such as a manual queue, a heuristics model, a statistical model or a separate AI system to overwrite outcomes from an AI system that are considered abnormal or too uncertain based on their confidence scores or confidence intervals;
- implementing a robust backup workflow that should be operational when an AI system fails to generate outputs.

An AI system's hardware should be robust to common causes of failure. From the point of view of common cause failures at the hardware level, there are no differences between conventional and AI-based hardware. A list of relevant common cause failures can be found in International Standards such as IEC 61508-2 and ISO 26262-11.

An AI system should adapt to evolving environments. Refer to 6.4 for the recommended guidance on functional adaptability.

11 Security

11.1 Confidentiality

Quality of the confidentiality sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.7.1.

The organization should evaluate an AI system's processes for ensuring data (both for input and output data) privacy or confidentiality to assess if they can be reverse engineered and would compromise identity protection or data integrity.

The organization should establish a mitigation plan, including depersonalization, synthetic data, dimension reduction or other mechanisms if a confidentiality risk is established.

11.2 Integrity

Quality of the integrity sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.7.2.

The organization should evaluate an AI system to assess if the system's integrity can be corrupted by erroneous or adversarial (data poisoning) input data.

The organization should establish a mitigation plan, including fraud detection, anomaly detection or other mechanisms if an integrity risk is identified.

11.3 Non-repudiation

Quality of the non-repudiation sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.7.3.

11.4 Accountability

Quality of the accountability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.7.4.

11.5 Authenticity

Quality of the authenticity sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.7.5.

11.6 Intervenability

Intervenability is a condition of controllability according to ISO/IEC TS 8200. Refer to 9.7 for appropriate guidance.

12 Maintainability

12.1 Modularity

Quality of the modularity sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.8.1.

12.2 Reusability

Quality of the reusability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.8.2.

The organization should store and abstract knowledge for an ML-based AI system during system creation for future use according to ISO/IEC 23053:2022, 7.7. An AI system consists of an input, weighting and output component. An AI system embeds knowledge in its learned weights, but the meaning of those weights requires contextualization by both the input and output components. Knowledge is accessed through the output components.

A reusable AI system should be designed to avoid the consequences of using embedded system weights in conceptual domains the training data does not cover. It also requires a pipeline able to differentiate between versions of training data. Possible approaches for changing the target domain of an AI system include the following:

- Retraining: for an ML-based AI system, the system is retrained in its entirety using a fresh training set inclusive of the new conceptual domain. This is the ‘brute force’ approach to managing target domain in an AI system’s usage.
- Transfer learning: for an ML-based AI system, ML model weights within the system are carefully tuned on a small set of data comprising the new conceptual domain to “reuse” weights already in the system to identify useful features.
- Reframing: the system’s individual components are altered to allow the system to process an unfamiliar domain. This can include adjustments to input component embeddings, output component normalization or adjustments, or switching off system components entirely.

- Revision: AI system core functionality is extended to encompass the new conceptual domain. This approach is more suited to simpler systems rather than more complex architectures such as neural networks.

An AI system should not be used for purposes it was not designed for, without taking additional steps for allowing system flexibility. This can involve additional feature engineering, retraining and revalidation to adapt the system to the new domain. A reusable AI system should be designed in a way to allow flexibility in how it interprets new information with limited manual tuning from designers. This requires a detailed examination of how information is represented within the system when it is first deployed.

Validation that an AI system is reusable for another concept should include considerations of the functional correctness guidance and measurements. Refer to 6.2 for guidance on functional correctness.

12.3 Analysability

Quality of the analysability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.8.3.

The organization should implement a mechanism for the logging of an AI system's predictions and explanations.

12.4 Modifiability

Quality of the modifiability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.8.4.

The organization should design an AI system adaptability mechanism in a way that ensures proper adaptation of the ML model without introducing defects. The development team should ensure that every new version of the ML model or system is increasing or at least maintaining its level of quality. Conditional workflows should be established if normal retraining is degrading the system quality.

Refer to 6.4 for guidance on functional adaptability.

12.5 Testability

Quality of the testability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.8.5.

The organization should implement a complete and efficient testing strategy with every new update to the system to avoid introducing defects.

The organization should implement complete, proactive and continuous monitoring capabilities.

13 Portability

13.1 Adaptability

Quality of the adaptability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.9.1.

13.2 Installability

Quality of the installability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.9.2.

13.3 Replaceability

Quality of the replaceability sub-characteristic should be measured against quality measures according to ISO/IEC 25023:2016, 8.9.3.

14 Effectiveness

Quality of the effectiveness sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.2.

15 Efficiency

Quality of the efficiency sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.3.

16 Satisfaction

16.1 General

Quality of the satisfaction characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.4.1.

16.2 Usefulness

Quality of the usefulness sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.4.2.

16.3 Trust

Quality of the trust sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.4.3.

16.4 Pleasure

Quality of the pleasure sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.4.4.

16.5 Comfort

Quality of the comfort sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.4.5.

16.6 Transparency

Refer to 9.8 for the recommended guidance on transparency.

17 Freedom from risk

17.1 General

AI risks should be identified, quantified or qualitatively described and prioritized against risk criteria and objectives relevant to the organization, according to a risk management framework for AI systems.

Risks arising from the development, the use of AI systems and their management should be managed according to ISO/IEC 23894.

17.2 Economic risk mitigation

Quality of the economic risk mitigation sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.5.2.

17.3 Health and safety risk mitigation

The organization should measure the quality of the health and safety risk mitigation sub-characteristic against quality measures according to ISO/IEC 25022:2016, 8.5.3.

Achieving health and safety risk mitigation should be an ongoing activity across the total product life cycle of a system, including early concept, requirements engineering, design, development, quality management, supply, acquisition, maintenance and decommissioning (ISO/IEC 22989). The requirements phase is especially important in designing safer AI systems.[18],[21]

Other standards and methods address health and safety in different ways and can complement the requirements and guidance provided in this document.

Addressing health and safety risks requires the balancing of often competing considerations in context, including with respect to other quality characteristics, such as security.

Several other quality characteristics detailed in this document can have a significant positive or negative bearing on the safety of an AI system; these include usability (Clause 9), learnability (9.2), operability (9.3), reliability (Clause 10) and functional suitability (Clause 6).

Consequently, AI system designers and users should carefully consider the relative importance of their chosen quality characteristics.

To effectively mitigate safety risks, an organization should:

- identify health and safety requirements of their systems according to intended use;[21]
- apply health and safety risk mitigation methods at the system level;[18],[20]
- identify software quality requirements derived from, and traceable to, the system level health and safety AI system requirements;[20],[23]
- identify and address health and safety risks relating to human, procedural and technical elements of their systems;[18]
- implement and operate an effective organizational safety management system covering the entire life cycle of the AI system;[18],[21]
- integrate its software management system into the organization’s AI quality management system (see ISO/IEC 42001);
- state a commitment to improving health and safety risk mitigation;
- ensure that the domain and requirements engineering activities have been undertaken by a sufficiently interdisciplinary team of experts;
- identify:
  - the environment in which their systems are intended to be used;
  - the goals of its AI system;
  - safety constraints (behavioural boundaries) within which its AI systems can operate safely;
- implement control measures to ensure that health and safety constraints are not violated;
- implement measures for:
  - the identification and prevention of unwanted losses;
  - the elimination, removal and reduction of hazards, including through system design, programming language and development method selection;[20]
- learn continuously about health and safety risk mitigation;

- identify and respond to both leading and lagging health and safety indicators;
- undertake a rigorous quality evaluation of the AI system requirements analysis to ensure that it includes:
    - the identification of health and safety goals;
    - the identification of safety constraints;
    - the identification of component interaction failure modes (this includes individual component failure modes) in the form of social, procedural and technical forms;
- undertake human factors engineering in systems requirements analysis and in system design and operation;
- list detailed assumptions about the specified intended domain and intended operational use of the AI system;

NOTE 1 An organization can include divergence of listed assumptions from the domains and use in which it plans to use AI systems as a source of risk to be addressed, and as a system selection factor.

- undertake hazard analyses, including techniques that identify the various forms of bias, such as cognitive biases in the intended users of the system;
- identify and address vulnerabilities to inherent biases in the hazard-analysis techniques employed (see ISO/IEC TR 24027 and References [21] and [25]);
- design AI systems to operate in safe states and, where possible, be able to recover from unsafe states.

An organization should prevent the following situations:

- absence of meaningful oversight over the selection, design, development, operation, maintenance or decommissioning of the AI system (ISO/IEC 38507);
- absence of specific safety goals and requirements;
- safety requirements developed after the AI system has been designed;
- documented assumptions that do not hold true for the intended use of the AI system;
- absence of documented assumptions;
- conflation of other quality characteristics with that of safety, for example:
    - redundancy;
    - robustness;
    - useability;

NOTE 2 As indicated in the introduction to this clause, these (and other) quality characteristics can have a significant bearing on the health and safety relating to an AI system, but they do not, in themselves, constitute safety. That is, an AI system that is reliable is not necessarily a safe system (indeed a less-reliable system can be safer than a reliable system in some cases).

- hazards that have been identified but then argued away as being unlikely rather than designed out or, where that is not possible, mitigated to the greatest extent possible;
- adoption of reductionist hazard-analysis methods for AI system risk acceptability.

Hazards should be designed out of:

- use of parameters or methods that can’t be verified (e.g. probability calculations) for safety risk decision-making (i.e. which risks to treat);
- a focus on risk identification and reduction rather than on hazard identification and removal;

- a focus or identification of user or operator error;
- claims or statements that the AI system is, or components of the system are, “safe”;

NOTE 3 For the purpose of this document, safety is considered to be an emergent property of a socio-technical system (with human, procedural and technical components) that arises from the interaction of its component parts. As such, it is not possible to evaluate a system as being safe or unsafe.

- use of language in documentation indicating adoption of linear-causality loss (accident) models, such as ‘chain of events’ and ‘root cause’.

17.4 Environmental risk mitigation

Quality of the environmental risk mitigation sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.5.4.

An organization should identify potential direct and indirect harms and benefits related to an AI system regarding environmental considerations, such as energy consumption, greenhouse gas emissions or water consumption, as described in Reference [27]. In addition to the typical environmental risks considered for a computer system, an AI development team should address the following elements:

- The computational resources, databases and applications used for the development and the operation of an AI system.
- The sophistication of functional requirements or functional correctness objectives, as they can require the usage of a large AI system that has a more negative environmental impact. Levels of computer resources utilized during training, inference and functional adaptability should be assessed, since the more resources, and the longer these resources are utilized, the more negative environmental impacts are caused by an AI system.
- The AI system’s adaptability strategy, as it can lead to additional environmental risks. For example, a system which is adapted continuously can consume more energy than a system that is adapted less frequently. Similarly, an AI system that gets adjusted using continuous learning is more efficient than full retraining.

The risk assessment and mitigation plan should also address indirect risks, such as the following:

- The resources needed for the manufacture of hardware used for the development or operation of an AI system (including the use of general-purpose hardware versus hardware tailored for specific types of AI system). An organization should consider the environmental risks across their entire supply chain.
- The impacts of the AI application, as the application of AI and its structural and behavioural effects can cause both positive and negative indirect environmental impacts. For example, an AI system that optimizes mining operations can exacerbate the negative environmental impacts of the mining, extractive and manufacturing sectors, or an AI system that generates product recommendations used in e-commerce can increase consumption in unsustainable ways. Also, an AI system that optimizes routes can have a positive impact by reducing energy consumption, or an AI system that predicts demand of perishable goods can reduce waste.

The identified risks should be further studied to the extent that mitigation actions can become part of the system requirements.

17.5 Societal and ethical risk mitigation

An organization that develops, produces, deploys or uses AI systems should manage societal and ethical risks by using ISO/IEC 23894, which provides guidance on key societal or ethical considerations.

An organization should identify and address societal concerns and ethical considerations throughout the life cycle of AI systems as part of a risk mitigation process. Please refer to ISO/IEC TR 24368, ISO/IEC TR 24027 and ISO/IEC TS 12791[1] for additional information.

[1] Under preparation. Stage at the time of publication: ISO/IEC DTS 12791:2023.

18 Context coverage

18.1 General

Though seemingly in the similar context of use, AI systems are at risk of unexpected behaviour, so measuring the context coverage of AI systems needs different considerations than conventional systems.

Since the performance of an AI system changes throughout its life cycle, the development team should continuously measure its context coverage.

18.2 Context completeness

Quality of the context completeness sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.6.2.

One significant pitfall of an AI system is its weak explainability and interpretability capabilities, as the outcomes of an AI system are generated from its training data and are not as explicit as in software development in which the code is producing the outcomes. Therefore, this requires a more detailed description of context of use for an AI system than a conventional system. The description of the context of use can be formed as a set of significant life cycle factors with specified estimates of the statistical characteristics (probability density functions) of each factor throughout the life cycle of an AI system.

During an AI system quality assessment, the organization should consider significant life cycle factors of either one or a combination of circumstances, opportunities and individual preferences.

These specified estimates can be derived from envisaged operating conditions, technical specifications, experience and other means. Different significant factors can be defined on various scales (e.g. numerical scale, interval scale, scale of categories, scale of relationships). While estimating statistical characteristics of significant factors, the cross-correlation features of these factors should be considered.

For example, in a facial recognition AI system, significant life cycle factors can include:

- perspective angles of the face (three angles);
- gender, age and race of the person;
- presence of the interfering objects and relative face area blocked by the interfering objects (e.g. glasses, hat, beard);
- time interval between the stages of training and operation of the ML model;
- level and gradient of illumination on the face;
- characteristics of the sensor used.

The completeness of the list of significant life cycle factors and accuracy of the estimates of the statistical characteristics of these factors should define the quality of context of use description.

The description of the context of use can be used for generating test datasets for assessment of AI system characteristics. In this case, the quality of the description of the context of use determines the representativeness of the measured characteristics of an AI system.

If $\bar{F}=\{f_i\}$, the set of the significant life cycle factors $f_i, i=1..N$ for a specific AI system; $D(\bar{F})$ is the probability density function of $\bar{F}$.

If significant life cycle factors are statistically independent, then function $D(\bar{F})$ can be represented as $D(\bar{F})=\Pi_i^N d_i(f_i)$, where $d_i(f_i)$ is the probability density function of f_i.

The specified contexts of use (specified conditions) for an AI system are defined as a range of values, $\Omega_1 \ni \bar{F}$, and contexts beyond the specified conditions of use are defined as a range of values, $\Omega_0 \ni \bar{F}$, in such a way that:

$$\Omega_1 \cap \Omega_0 = \emptyset, \quad \Omega_1 \cup \Omega_0 = \Omega$$

where Ω is the range of all possible values of significant life cycle factors $\bar{F}$.

$\bar{F}$ takes discrete values in such a way that:

$$L_1 = |\Omega_1| \text{ and } L_0 = |\Omega_0|$$

where L_1 and L_0 are the cardinality of the sets Ω_1 and Ω_0.

If an AI system function of quality Q($\bar{F}$) is defined so that:

$$Q(\bar{F}) = \begin{cases} 1, \text{if AI system quality satisfied to requirements specified by the user} \\ 0, \text{if AI system quality not satisfied to requirements specified} \end{cases}$$

then the context coverage can be calculated as the mean of the quality value in all contexts using the formula:

$$C = \frac{\sum_{\bar{F} \in \Omega}^{L} Q(\bar{F}) D(\bar{F})}{\sum_{\bar{F} \in \Omega}^{L} D(\bar{F})} = \sum_{\bar{F} \in \Omega}^{L} Q(\bar{F}) D(\bar{F})$$

where $L = L_1 + L_0$ is the cardinality of the set Ω.

C varies from 0 to 1, where 0 means an unusable AI system and 1 means an AI system with proper quality in all probable contexts of use.

The flexibility can be calculated as the mean of the quality value beyond the specified context using the formula:

$$C_0 = \frac{\sum_{\bar{F} \in \Omega_0}^{L_0} Q(\bar{F}) D(\bar{F})}{\sum_{\bar{F} \in \Omega_0}^{L_0} D(\bar{F})} = \frac{\sum_{\bar{F} \in \Omega_0}^{L_0} Q(\bar{F}) D(\bar{F})}{1 - D_1}$$

The context coverage, context completeness and flexibility should all be linked by the following formula:

$$C = D_1 C_1 + (1 - D_1) C_0$$

Typically for an AI system well adapted to the specified context of use, $C_0 < C < C_1$.

18.3 Flexibility

Quality of the flexibility sub-characteristic should be measured against quality measures according to ISO/IEC 25022:2016, 8.6.3.

Bibliography

[1] ISO/IEC 5338, *Information technology — Artificial intelligence — AI system life cycle processes*

[2] ISO/IEC TS 8200, *Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems*

[3] ISO/IEC TS 12791, *Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks*

[4] ISO 21448:2022, *Road vehicles — Safety of the intended functionality*

[5] ISO/IEC 23894, *Information technology — Artificial intelligence — Guidance on risk management*

[6] ISO/IEC TR 24027, *Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*

[7] ISO/IEC TR 24029-1:2021*Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview*

[8] ISO/IEC 24029-2:2023, *Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*

[9] ISO/IEC/TR 24368, *Information technology — Artificial intelligence — Overview of ethical and societal concerns*

[10] ISO/IEC 25022, *Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — Measurement of quality in use*

[11] ISO/IEC 25023:2016, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of system and software product quality*

[12] ISO/IEC 25040, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Evaluation process*

[13] ISO 26262-11, *Road vehicles — Functional safety — Part 11: Guidelines on application of ISO 26262 to semiconductors*

[14] ISO/IEC/IEEE 29119 (all parts), *Software and systems engineering — Software testing*

[15] ISO/IEC 38507, *Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations*

[16] ISO/IEC 42001, *Information technology — Artificial intelligence — Management system*

[17] IEC 61508-2, *Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 2: Requirements for electrical/electronic/programmable electronic safety-related systems*

[18] Dekker, S. *Foundations of Safety Science. A Century of Understanding Accidents and Disasters*. CRC Press, 2019

[19] Dekker, S. *Drift into Failure. From Hunting Broken Components to Understanding Complex Systems*. CRC Press, 2011

[20] Dobbe, R. *System Safety and Artificial Intelligence*. 2011. Available from: https://arxiv.org/pdf/2202.09292.pdf

[21] Leveson, N. *Engineering a Safer World: Systems Thinking Applied to Safety*. Boston, MA. MIT Press, 2012

[22] Meadows, D., Wright, D., eds. *Thinking in Systems: A Primer*. Chelsea Green Publishing Co., 2008

[23] Rierson, L. *Developing Safety-Critical Software. A Practical Guide for Aviation Software and DO-178C Compliance*. CRC Press, Taylor & Francis Group LLC, 2013

[24] Leveson, N. High-pressure steam engines and computer software. *Computer*. 1994, 27(10), 65–73. doi: 10.1109/2.318597

[25] Leveson, N. *Improving the Standard Risk Matrix: Part 1*. 2019

[26] Leveson, N. The Therac-25: 30 Years Later. *Computer*. 2017, 50(11), 8–11

[27] OECD. Measuring the environmental impacts of artificial intelligence compute and applications. *OECD Digital Economy Papers*. 2022, 341. doi: 10.1787/7babf571-en

ICS 35.080

Price based on 20 pages

iso.org