FINAL
DRAFT

TECHNICAL
SPECIFICATION

ISO/IEC DTS 25058

ISO/IEC JTC **1**/SC **42**

Secretariat: **ANSI**

Voting begins on:
2023-09-28

Voting terminates on:
2023-11-23

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

Reference number
ISO/IEC DTS 25058:2023(E)

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 through proper 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 in 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:

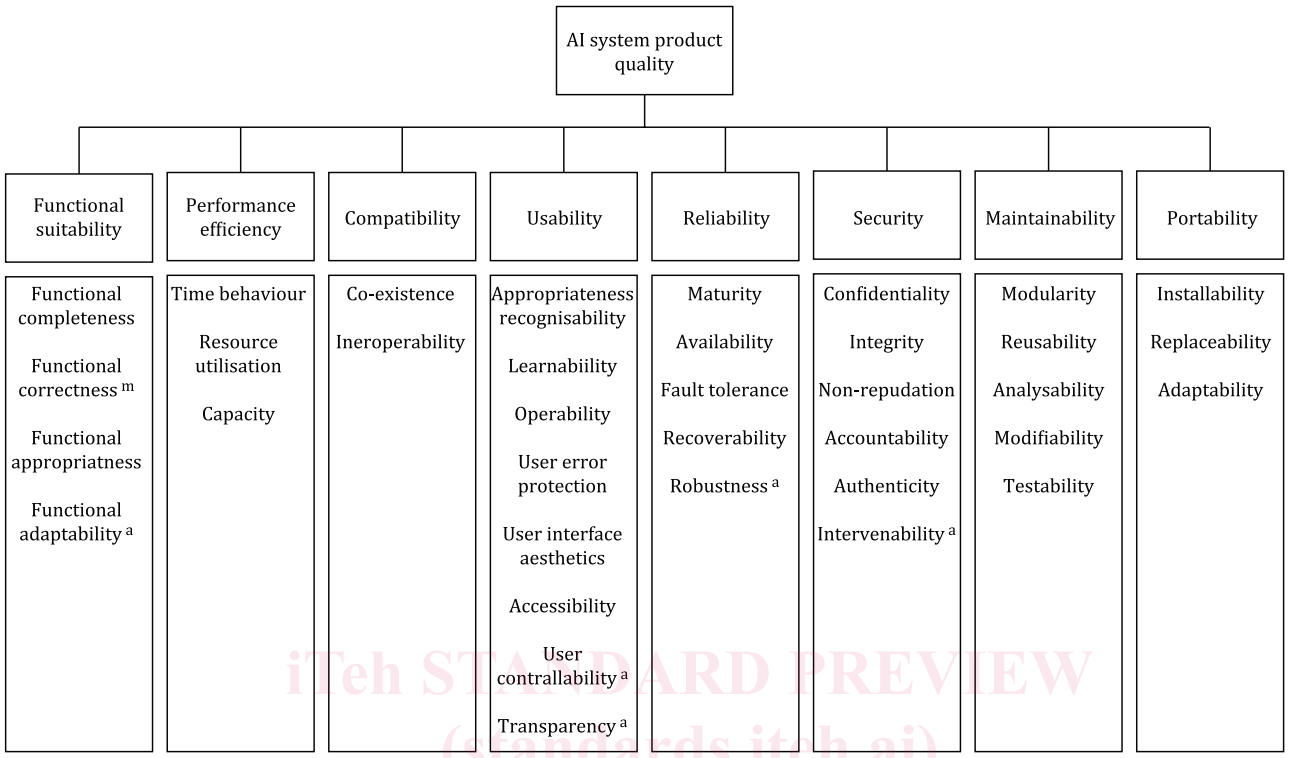
— ISO Online browsing platform: available at https://www.iso.org/obp

— 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.

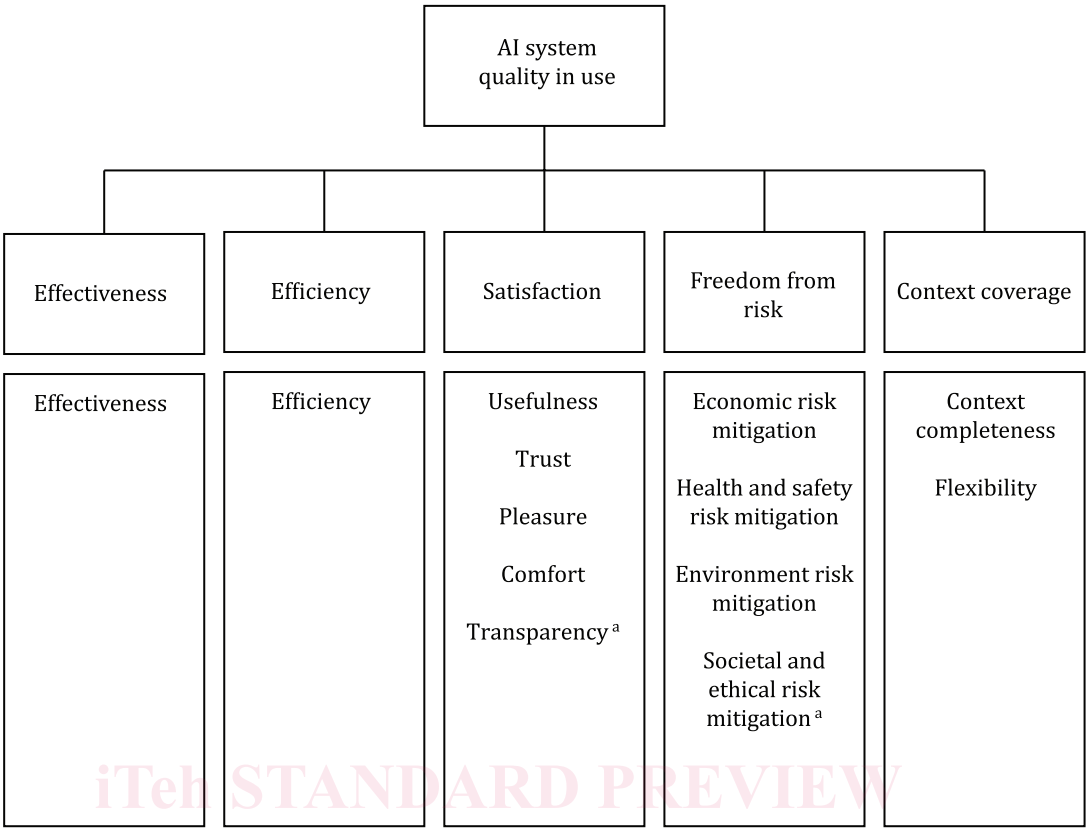


a New sub-characteristics.

m Modified sub-characteristics.

SOURCE ISO/IEC 25059:2023, Figure 1.

Figure 1 — AI system product quality model



[a] New sub-characteristics.

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 K-fold 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.