ISO/IEC JTC 1/SC 42

ISO/IEC CD TR 42106(en)

Secretariat: ANSI

Information technology — Artificial intelligence — Overview of differentiated benchmarking of AI system quality characteristics

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

Published in Switzerland

Contents

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 documents 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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on 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 the following URL: 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 atwww.iso.org/members.html.

Information technology — Artificial intelligence — Overview of differentiated benchmarking of AI system quality characteristics

1 Scope

This document provides an overview of conceptual frameworks for graded *benchmarking* of AI system quality characteristics. The aim is to examine the feasibility of using differentiated *benchmarking* of quality characteristics based on the complexity and context of use of an AI system.

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 22989:2022, *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 TR 24028:2020, *Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence*

ISO/IEC TR 24030:2024, *Information technology - Artificial intelligence (AI) - Use cases*

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

ISO/IEC 29155-1:2017, *Systems and software engineering - Information technology project performance benchmarking framework - Part 1: Concepts and definitions*

ISO 41011:2024, *Facility management — Vocabulary*

3 Terms and definitions

The Terms and definitions clause is a mandatory element of the text.

For rules on the drafting of the Terms and definitions, refer to the ISO/IEC Directives, Part 2:2018, Clause 16.

To insert a new terminological entry, go to the *Structure* tab and click on *Insert Term entry*.

For the purposes of this document, the terms and definitions given in ISO/IEC 22989:2022, ISO/IEC 23053:2022 and the following apply.

ISO and IEC 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
benchmark
reference point against which comparisons can be made

Note 1 to entry: In the context of ISO/IEC 42106, an AI system quality characteristic is the object of comparison

3.2
benchmarking
activity of comparing objects of interest to each other or against a benchmark to evaluate characteristic(s)

Note 1 to entry: In the context of ISO/IEC 42106, the object of interest is an AI system quality characteristic.

4 Overview of relevant benchmarking methods

4.1 Review of benchmarking definitions

Searching '*benchmarking* (3.2)' from the following standards terminology databases, there are 76 results from OBP; 0 results from IEV, 0 results from ITU-T. By deleting not relevant terms and definitions and merging the same definitions together, 14 definitions are collected (see Annex A).

— ISO Online browsing platform (OBP): available at http://www.iso.org/obp

— IEC Electropedia (IEV): available at http://www.electropedia.org/

— ITU-T (ITU-T) Terms and Definitions available at: https://www.itu.int/br_tsb_terms/#/

These 14 definitions include several instances that define *benchmark* (3.1) and *benchmarking* (3.2) as a pair, with the definition of *benchmarking* (3.2) relying on the paired definition of *benchmark* (3.1) (e.g. The pairs {D3, D5}, {D13, D14}, {D4, D2}). A clustered view of the objects of interest for each of these definitions of *benchmark* (3.1)/*benchmarking* (3.2) is given in Table 1.

Table 1 — Objects and characteristics relevant for **benchmark** and **benchmarking**

Code	**Object**	**Description of characteristics**	Sources
1	a reference point/tool/method (metric against; any standard or reference; point of fixed location; permanent mark) (benchmark)	comparisons can be made ; process , performance or q uality can be measured; others can be measured	D4, D5, D6, D10, D11, D12, D13
2	activity of comparing, evaluating and analysis (activity of comparing or evaluate ; comparative evaluation or analysis; activity of measurement and analysis)	objects of interest to each other or against a benchmark , characteristic; similar operational practices; an organization can use to search for and compare practices inside and outside the organization, with the aim of improving its performance'; similar operational practices	D3, D7, D8, D9, D14

3	process of comparing processes, performances or quality against practices	the same nature, under the same circumstances and with similar measures	D2
4	single value (benchmark)	used for orientation	D1

Core concepts about "benchmarking" are reflected in the repetition of words and phrases across these 14 definitions. Among these, "comparisons" (10 times), “performances” (7 times), "can be measured" (6 times), and "practices" (5 times) are the most frequently used core concepts.

The terms “process” (4 times),“organization”(4 times)， "a reference point" (3 times), “metric against" (3 times)， “evaluate (3 times)”,and “standard” (3 times) also contribute to an understanding of the basic concept of "benchmarking”.

4.2 Types of benchmarking

From the review of uses of *benchmarking* (3.2) in the standardization literature, it is evident that there are primarily two types of *benchmarking* (3.2) in existing definitions of ISO deliverables in terms of different focuses on activity (ISO/IEC 29155-1) or processes (ISO 41011) as objects of *benchmarking* (3.2). With regards to activity as *benchmarking* (3.2) object, the focus lies on comparing objects of interest to each other or against *benchmark* (3.1) to evaluate criteria or characteristics (ISO/IEC 29155-1:2017). Such activities have characteristics of similar operation practice, similar attributes, processes or performance that are comparable. The *benchmark* (3.1) refers to reference point or metric against which comparisons can be made. Reference point can be tool for performance improvement through systematic search and adaptation of leading practice, can be standard against which results can be measured or evaluated, can be method for comparing the performance of the leading organizations in a market segment or procedure, problem, or test that can be used to compare systems or components to each other or to a standard.

With regards to process as the benchmarking object, the focus lies on comparing processes, performances and /or quality against practices of the same nature, under the same circumstances and with similar measures. Its special considerations are the systematic process for the identification of, becoming acquainted with and for adoption of successful practices of benchmarking partners. Such concept is use in domain of facility management (ISO 41011:2024).

Within this document, we use the concept of *benchmarking* (3.2) to focus on the activity of comparing objects of interest against *benchmark* (3.1) to evaluate characteristics, and the concept of *benchmark* (3.1) to focus on a reference point to which comparisons can be made. Such concepts are used widely in domain of information technology project performance *benchmarking* (3.2) framework of systems and software engineering.

Therefore, the definitions of *benchmark* (3.1) and *benchmarking* (3.2) given in this document are adapted from ISO/IEC 29155-1:2017, 3.2 and ISO/IEC 29155-1:2017, 3.3 respectively, for reflecting the emphasis on product benchmarking most clearly.

4.3 Metrics, measures and criteria

AI system performance is measured using a vast array of quantitative metrics. Additionally, a number of measures, such as loss functions, are relevant for measuring performance during training, but not as finally reportable metrics of the system's performance. Finally, some criteria are used for model size determination, model selection, model training time etc., but are not directly reported as performance indices.

We review metrics, measures and criteria used for common machine learning tasks. The given list is not comprehensive, but is intended to provide a useful overview of tasks and their corresponding measures.

Table 2 — Metrics, measures and criteria for common machine learning tasks

Task	Measures	**Criteria**	References
Classification	Accuracy Cross Validation Precision, Recall, Confusion Matrix ROC Curve	Model size	Data Mining Practical Machine Learning Tools and Techniques(Ian H. Witten,Eibe Frank) - Ch 1.5(Generalization as search), - - Ch 6.2 (Classification rule) Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow by O'REILLY - Ch 3 Classification
Regression	Root mean squared error (RMSE)	Neural Nets - Early Stopping - Weight Decay - Model Averaging	Section 5.1(5.1 Quantitative Measures of Performance, Applied Predictive Modeling Springer)
Prediction Ranking	Spearman's rank correlation		Applied Predictive Modeling Springer - Sec 5.1 Quantitative Measures of Performance,) - Sec 7.1 Neural Networks
Localisation (Bounding box around an object)	Intersection over Union (IoU)		Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow by O'REILLY - Ch14 Deep Computer Vision Using Convolutional Neural Networks (Object Detection)
Object Detection	Mean Average Precision (mAP)	Early stopping, Validation loss monitoring	Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow by O'REILLY - Ch14 Deep Computer Vision Using Convolutional Neural Networks (Object Detection)
Image - Semantic segmentation	Mean Intersection over Union (mIoU)		Advanced Deep Learning with TensorFlow 2 and Keras (Packt Publishing) - Sec 12.5 Semantic Segmentation Validation
Time Series Forecasting	MSE, MAE, MAPE(percentage error), Mean Absolute Scaled Error(MASE)		Forecasting: Principles & Practice(Rob Hyndman) - 2.6 Evaluating forecast accuracy

Task	Measures	**Criteria**	References
POS tagging	Accuracy		Speech and Language Processing(Daniel Jurafsky) - 8.6 Evaluation of Named Entity Recognition
Named Entity Recognition	Precision,Recall,F1 Score		Speech and Language Processing(Daniel Jurafsky) - 8.6 Evaluation of Named Entity Recognition
Dependency Parsing	Labelled attachment score (LAS), unlabeled attachment score (UAS),label accuracy score (LS)		Speech and Language Processing(Daniel Jurafsky) - Sec 18.4 Evaluation
Information Retrieval	PR curve, interpolated Precision, Mean Average Precision		Speech and Language Processing (Daniel Jurafsky) 14.1 InformationRetrieval
Summarisation	ROGUE(F1 score from the n-gram precision and recall)		Natural Language Processing with Transformers - Ch 6 Summarisation

Mathematical descriptions of measures used in this Table are given in Informative Annex C.

It is notable that, whereas AI system quality encompasses multiple dimensions, as listed in ISO/IEC 25059:2023, existing metrics heavily prioritize measurement of functional suitability, to the exclusion of several other important characteristics, such as reliability, maintainability, usability and security. We review the consequences of this imbalance further below in Clause 5.4.

5 Benchmarking AI systems

5.1 Benchmarking AI systems quality

Benchmarking the quality characteristics of AI systems is crucial for several reasons. Firstly, it allows us to measure and compare the performance of different AI models objectively, providing valuable insights into their strengths and weaknesses. By *benchmarking* (3.2) factors such as accuracy, efficiency, reliability, and robustness, stakeholders can identify areas for improvement and innovation, driving advancements in AI technology. Additionally, *benchmarking* (3.2) facilitates standardization and *benchmarking* (3.2) facilitates standardization and transparency within the AI ecosystem, enabling stakeholders to make informed decisions about which models are most suitable for their specific needs. Furthermore, *benchmarking* (3.2) helps to establish benchmarks against which future AI systems can be evaluated, fostering a continuous cycle of improvement and innovation.

Several methods exist for *benchmarking* (3.2) AI systems, each tailored to measure specific quality characteristics and performance metrics. 2937 provides formal methods for performance *benchmarking* (3.2) of hardware-related metrics of AI server systems, emphasizing the measurement of training time, power consumption, and inference latency. To measure the functional correctness and suitability of AI systems, the general approach is the use of standardized datasets and evaluation metrics, where AI models are tested on established *benchmark* (3.1) datasets, such as ImageNet for image classification or MNIST for handwritten digit recognition. These datasets come with predefined training and testing splits, enabling consistent evaluation across different models. Another method involves organizing competitions and challenges, such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) or the Common Objects in Context (COCO) challenge, where researchers and developers submit their AI models to compete against each other on specific tasks. These competitions provide a platform for rigorous evaluation and comparison of AI systems in real-world scenarios. Most relevant for this report, standardization organizations like the National Institute of Standards and Technology (NIST) and the AI Benchmarking Initiative (AIBench) have developed standardized methodologies and benchmarks for evaluating AI systems in specific domains, promoting transparency and reproducibility in AI research.

While current approaches for benchmarking AI systems are valuable, they also have several limitations. An important limitation is dataset bias, where the performance of AI models can be skewed due to biases present in the training data. This can lead to over-fitting to specific datasets and poor generalization to real-world scenarios. Leader-board competitions, the most common form of accuracy *benchmarking* (3.2) for AI systems, are particularly sensitive to dataset decay, and require careful handling to prevent over-fitting to held out data [4]. Another challenge is the proliferation of evaluation metrics across domains, making it difficult to compare the performance of AI models across different tasks. Benchmark datasets and competitions typically focus on narrow tasks or domains, limiting the scope of evaluation and potentially overlooking important aspects of AI systems, such as ethical considerations and societal impact. Moreover, the reproducibility of *benchmarking* (3.2) results can be challenging, particularly when details about model architectures, hyper-parameters, and training procedures are not adequately documented.

5.2 Context-of-use

Software quality standards have historically emerged from hardware quality standards, which in turn were created to address simple mechanical systems. For such simple systems, component reliability tends to correlate well with nearly all desirable quality metrics, such as functional correctness, safety and resilience. For the most part, since software systems also have conceptually enumerable input-output characteristics, standardization approaches rooted in reliability engineering have translated well to them.

However, this historical provenance of software quality standards systematically under-emphasises the role of context-of-use on the quality characteristics of software-based systems. This is a significant limitation, as the context-of-use offers considerable information about the possible hazards of a system's use, which is necessary to design appropriate requirements for the system. As Nancy Leveson observes, "System and software requirements development are necessarily a system engineering problem, not a software engineering problem." [5].

It is therefore necessary to treat AI systems from a sociotechnical perspective, ensuring that the degree of quality assurance is aligned with the degree of quality expected of the system based on the context-of-use.

5.3 Complex adaptive systems

Annex B in NIST AI 100-1 summarizes key aspects in which risks from AI systems are different from risks from traditional software systems. Directly relevant to this report, it is argued that:

a) Data used in model training is not always representative of the context-of-use of the system.

b) It is possible that real ground truth data does not exist, or is not available.

c) Data distributions could drift over time, and become detached from the original context in which the system was trained.

d) Use of pre-trained models limits controllability of data quality and bias mitigation strategies

In addition to these risks to system correctness, multiple additional sociotechnical considerations apply for other quality characteristics of AI systems, such as:

e) Humans interacting with AI systems can change their behaviours to work around the narrow intelligence of such systems replacing human operators.

f) AI systems can be subjected to data poisoning and spoofing attacks, reducing their effectiveness, when deployed.

g) Human operators working alongside AI decision support systems can become overconfident and accept system suggestions by default

h) Human operators working alongside AI decision support systems could mistrust and ignore AI system suggestions

i) AI system integration into legacy IT systems could expand the cybersecurity threat envelope of the existing system in ways that are difficult to detect with an audit of the two systems in isolation.

This list of considerations is not comprehensive, and is presented primarily to emphasize the thematic point that AI systems must be validated with a sociotechnical systems approach, accounting for the fact that they interact with users and third parties in complex ways, and that other entities adapt to being interacted with by AI systems in ways that are not always foreseeable. Thus, new methods and approaches for *benchmarking* (3.2) complex and adaptive systems could be proposed.

Editor's Note: DW to make a contribution looking in more detail at what the challenges of benchmarking complex, adaptive systems actually are, partially bringing in perspectives from control theory.

5.4 Limitations in benchmarking AI systems

While *benchmarking* (3.2) is already a challenging activity for simpler systems, requiring standardization of multiple facets of data, processes and measurements, *benchmarking* (3.2) AI systems poses novel challenges that must be addressed with care. In particular, it is challenging to benchmark AI systems because

— AI systems are applied in a variety of domains and contexts of use, each with different sources of risk and uncertainty. Benchmarking such systems can either be adaptive to these differences, or be sufficiently comprehensive to address all of them. For example, object detection models are frequently benchmarked using mean average precision (mAP) across object classes, advised for instance in 2937. However, there are several contexts-of-use, e.g. object recognition for driverless vehicles wherein misidentification of some classes of objects, e.g. pedestrians at risk, is of greater importance than other objects, e.g. street signs. In such contexts, mAP may well exaggerate the functional suitability of the system, since low importance classes are more likely to natively be encountered in the data environment than high importance classes.

— Designers report AI system performance using a variety of metrics, with comparisons across metrics not possible. For example, predictive models for healthcare domains frequently report performance in terms of F1 score or area under the ROC curve. However, such measures assume the availability of infinite clinical resources to act upon model predictions. In reality, clinicians may only be able to act upon a limited number of inputs from such predictive models, thus favoring evaluation metrics drawing upon the recommender systems literature, such as mean reciprocal rank, top-k precision etc. These metrics are mutually incommensurable, making it difficult to assess the true value of such systems in use[7].

— Benchmark datasets can contain noise, imbalance, and bias in unknown quantities. Performance evaluations inherit these problems in the form of fragility, inaccuracy and algorithmic bias respectively. Examples of AI algorithms perpetuating societal and demographic biases abound in the academic literature, and a vast literature on fairness in machine learning has emerged in response to this problem [8].

— Evaluating very large models requires specialized techniques and infrastructure, which are not equally accessible under resource constraints. Particularly for large language models, the computational and

energy requirements necessary to train models are very large, and inaccesible to most public institutions[9]
.

— Many AI applications involve interaction with humans, and the nature of this interaction changes reflexively as humans adapt to the use of the system. For example, automation of actions in cockpits has been shown to be associated with atrophy of flying skills in human pilots [10], and similar deficits are anticipated in the use case of driverless cars[11]. Benchmarking AI systems in such contexts requires consideration of human factors and user experience, which adds considerable complexity to any possible evaluation.

These problems are heavily inter-linked, and load heavily on the fact that modern AI systems are developed using very large datasets and very large models, with downstream sociotechnical considerations not clearly known at the time of system *benchmarking* (3.2). While it is possible to develop comprehensive *benchmarking* (3.2) standards that accommodate the large scale and complexity of AI systems in use, the application of such standards would necessarily require high levels of expertise and resource allocations. This would inevitably create a large compliance burden on organizations and other stakeholders in the AI ecosystem.

Alternatively, it is possible to conceive of approaches for differentiated *benchmarking* (3.2) of AI systems, such that quality characteristics of systems are benchmarked at different levels, and with different degrees of standardization, adaptive to sociotechnical consideration of where such a system lies on a spectrum of potential for harm. In this way, compliance burden would rationally scale with the harm potential of AI systems, thus simultaneously enabling innovation while maintaining safety.

6 Approaches for differentiated benchmarking

6.1 Management frameworks

A management systems standard (MSS) is a set of guidelines and criteria to help organizations implement effective management practices. These standards provide a framework for organizations to structure their processes, improve efficiency, meet regulatory requirements, and achieve specific objectives.

ISO/IEC 42001 is an example of an MSS specifying requirements for establishing, maintaining and improving AI management systems within organizations. The target for this standard are organizations producing AI-based products and services.

MSS are likely to be centrally important in managing the development and deployment of AI systems. To some extent MSS can provide guidance with respect to differentiated *benchmarking* (3.2) by pointing organizations to specific processes that require amplification in certain contexts-of-use. Therefore, this clause reviews MSS relevant for AI systems.

The AI Risk Management Framework (AI RMF) developed by the US National Institute of Standards and Technology (NIST) is another representative exampleNIST AI 100-1. This framework utilizes a descriptive methodology, offering flexibility in implementation. It focuses on assessing hazards, exposures, and vulnerabilities associated with AI systems, allowing organizations to manage risks effectively across various use cases and sectors.

ALTAI, developed by the European Commission's High-Level Expert Group on AI, is a procedural framework released in June 2019 and updated until July 2020. It covers all principles and stages of AI implementation, offering a region-agnostic and sector-agnostic perspective. ALTAI follows a procedural approach, emphasizing trustworthiness in AI systems. It assesses hazards, exposures, and vulnerabilities to ensure the ethical and trustworthy deployment of AI technologies.

The Algorithm Impact Assessment Tool (AIA) is a Canadian government initiative established in 2019 and updated until November 2022. While specific focus areas are not explicitly mentioned, AIA addresses planning, requirements analysis, design, and testing stages. The framework takes a procedural approach, ensuring that AI implementations are region-agnostic and sector-agnostic. AIA assesses hazards, exposures, and vulnerabilities without specifying particular domains.

The Recommended Practices for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being, a standard introduced by the US Institute of Electrical and Electronics Engineers (IEEE) in May 2020, follows a descriptive approach. While specific focus areas are not explicitly mentioned, the practices are designed to be region-agnostic and sector-agnostic. The framework provides guidance on assessing hazards, exposures, and vulnerabilities associated with autonomous and intelligent systems, emphasizing their impact on human well-being.

MSS are workhorses of standardization activities. They enable organizations to address aspects in their management processes without having to rework their internal organizational vocabularies, foregrounding the voluntary nature of standardization. MSS also permit organizations to address all aspects of their workflows, including qualitative elements that are difficult to address with measurements.

However, MSS, for all their salutary properties, cannot by themselves produce trustworthy products. They must be supplemented by standards that document technical aspects of system development, testing and *benchmarking* (3.2). Additionally, it can be helpful to provide guidance for which technical benchmarks apply to which system in which context-of-use.

6.2 Classification-based frameworks

Classification is a very common form of standardization activity [13]. The standardization of IT systems, in particular, seems to lend itself well to classification-based frameworks, as is evidenced by NIST's cybersecurity framework subcategories, which enables an organization to standardize processes relevant for its specific needs [14]. The judgment of relevance provides the source of differentiation in the standardization process in such frameworks, with the most common frame of relevance judgment being risk or impact assessment.

A number of frameworks for risk or impact assessment of artificially intelligent systems pre-exist. Some of these frameworks use risk-based classification to differentiate *benchmarking* (3.2) treatment for various AI products. Some such frameworks are reviewed below.

The German Data Ethics Commission has created a guidance document describing five criticality classes for AI systems depicting harm for, i.e. the physical as well as psychical well-being, finance, date, manipulation of information as well as a negative form of nudging. Based on this guidance, regulation classes for AI systems can vary depending on the jurisdiction and specific regulations in place. The document describes these five regulation classes with corresponding duties for responsible parties, such as providers and manufacturers as well as concerns which justify the placement of an AI system into a specific class.

Class 1 - No or Minimal Potential for Harm:

Duties: correctness checks, transparency, system analyses in cases of suspicion.

Concerns: potential for unexpected or unintended consequences.

Class 2 - Low Risk:

Duties: risk assessment, transparency obligations, and basic safety standards.

Concerns: undue risks to individuals or society.

Class 3 - Moderate Risk:

Duties: oversight, risk assessments, third-party audits, and adherence to specific industry standards.

Concerns: harm to individuals, privacy violations, diffusion in accountability, fairness in AI decision-making.

Class 4 - High Risk:

Duties: thorough risk assessments, continuous monitoring, and robust fail-safes, independent audits, compliance with strict safety and security standards, and regular reporting to regulatory authorities.

Concerns: significant harm to individuals, society, or critical infrastructure as well as negative ethical, legal, and social implications, including, i.a., discrimination, bias, and transparency.

Class 5 - Forbidden:

Duties: extraction of product from market by supervision authorities.

Concerns: extreme potential for harm, including threats to human life, national security, or global stability. Immediate detection of product of such classes, as well clarity in accountability are indispensable.

The EU AI Act also adopts a risk-based classification approach, categorizing AI systems into different risk levels based on their potential impact on rights, safety and societal values. High risk systems are expected to be subject to stricter requirements and oversight, and providers of high risk systems are expected to comply with additional requirements related to data quality, documentation, transparency and traceability throughout the AI system's life-cycle. The Act also allows for conformity assessment to verify compliance with the requirements set forth in the regulation.

The Automated Decision-Making Systems in the Public Sector: An Impact Assessment Tool for Public Authorities, developed by Algorithm Watch in June 2021, is designed for public authorities in Germany. It provides guidelines for assessing hazards, exposures, and vulnerabilities associated with AI implementations in the public sector. This framework follows a tiered procedural approach, wherein systems that exceed a set threshold score in a first-level checklist during evaluation are taken through a more extensive set of controls than systems that do not.

Introduced in 2018 by the NL ECP, Platform for the Information Society, the Artificial Intelligence Impact Assessment framework provides guidelines for assessing the impact of AI technologies. The framework is designed to be region-agnostic and sector-agnostic. This framework also follows a tiered procedural approach, offering organizations guidance on assessing hazards, exposures, vulnerabilities, and mitigation risks associated with AI implementations based on the perceived criticality of the deployment.

The Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration, developed by the European Law Institute (ELI), was introduced in January 2022. The model rules are designed to be region-agnostic and applicable to public sectors. This framework also follows a tiered procedural approach, guiding organizations in conducting impact assessments related to hazard, exposure, and vulnerability associated with algorithmic decision-making systems based on the perceived criticality of the deployment.

Classification-based approaches to risk and impact assessment have the advantage of being commonly known, easily reproducible, easily documentable, and intuitive to work with for regulatory and governance bodies. However, there are also several limitations to such approaches.

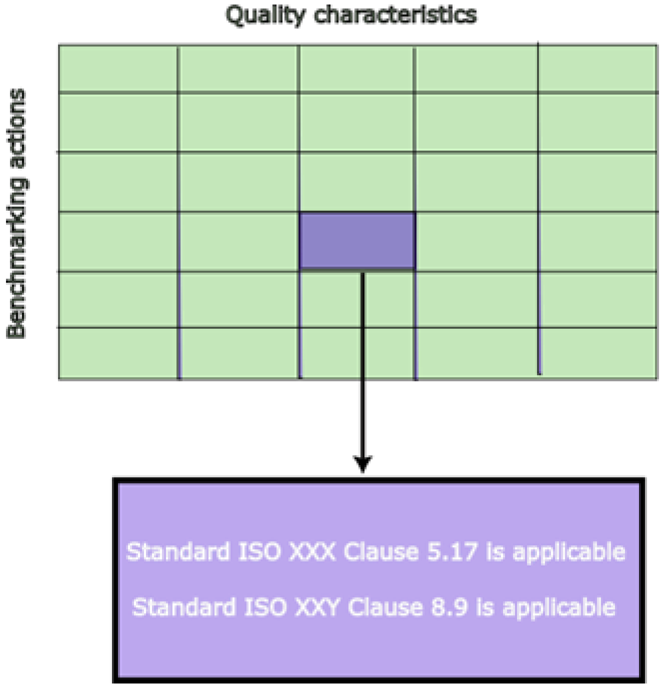
For example, risk matrix based approaches to risk assessment presume a utilitarian view of risk, such that there exists an implicit acceptance of severe risks provided the likelihood of such risks is acceptably low [15]. It is, however, well known that people systematically underestimate the probability of unlikely events in decisions they make from experience [16]. Therefore, analysts' likelihood estimates inevitably understate expected risk when risk matrix approaches are applied in the determination of risk [17].

Additionally, classification is inherently a unidimensional approach to standardization, which is reasonable in cases where the dimension along which risk or impact is expected to vary is clearly understood, but not in cases where the dimensionality of risk variance itself is complex and multi-dimensional. Risk-based classification of AI systems fundamentally inherits this defect.

6.3 Levels of specification

Management systems standards can specify the set of standardization options available to someone seeking to benchmark AI system quality, but not technical guidance with respect to specific actions they can take to attain their purpose. Risk-based classification schemes can guide them towards specific actions appropriate for a relevant context-of-use, but they do so based on a relatively unidimensional evaluation of context and risk. The complex sociotechnical nature of AI system deployments can be suitable by some flexible generalization of existing classification-based approaches.

Software specifications offer guidance for ensuring that programs actually do implement the logic they are expected to implement. These specifications can vary in their level of detail, with standard modules specified mostly as flowcharts, and critical subunits specified with additional information about data ranges, exception possibilities, etc. Repurposing this software idiom for the task of *benchmarking* (3.2) AI systems, levels of specification can be created to specify the set of *benchmarking* (3.2) actions appropriate for AI systems with different potentials for harm.


Quality characteristics
Benchmarking actions
Standard ISO XXX Clause 5.17 is applicable
Standard ISO XXY Clause 8.9 is applicable

A levels-based approach to *benchmarking* (3.2), as illustrated in the diagram above, would specify the set of benchmarks or *benchmarking* (3.2) procedures necessary to establish quality characteristics for AI systems with a particular level of harm potential. The levels approach, therefore, is a generalization of the classification-based approach, seeking to map AI systems with different harm potentials to benchmarks targeted at an enumerated set of quality characteristics.

The construction of AI specification levels requires (a) the design of property-action matrices, mapping software quality characteristics for AI systems to standardization schemes and products, (b) modelling the ecosystem of AI applications at a general level to identify common patterns of standardization needs for different stakeholders, and (c) characterising property-action matrices particularized for systems with different levels of complexity. Specification levels themselves emerge as nominal categories describing the specific property-outcome matrix to be applied while specifying or designing any given system. Indicative sample property action matrices are presented in Informative Annex B.

By associating commonly occurring patterns of quality characteristic requirements within AI systems with a rubric of specification *levels,* stakeholders will be able to communicate the trustworthiness of AI systems to observers succinctly, yet accurately. At the same time, the development of such specification levels could also ease the development and execution of regulatory frameworks within specific application domains.

However, the definition of such specification levels necessarily involves discretization of an intrinsically continuous spectrum of potential for harm, which is not necessarily possible or desirable in all contexts-of-use for AI systems, and, like all discretization exercises, could require the definition of arbitrary boundaries.

7 Feasibility analysis

7.1 Case Example 1: On-job training recommendation system

A recommender system, VTrain, (use case 23; ISO/IEC TR 24030:2024) draws upon real-life data about on-job training from about 120k employees across a catalog of about 5000 courses. The core AI elements of this system are algorithms that seek to cluster sequences of courses taken by other employees to suggest courses from the same cluster as courses already taken by an employee to them. The performance of the algorithm is measured by back-tested prediction accuracy, viz. the number of predicted courses actually taken by employees in test set.

The organization self-identifies different sources of bias potentially built into the model via its training set, and the stress of inappropriately directing employees to take certain courses through an uninterpretable model as possible threats and vulnerabilities of the system. From the perspective of the AI software quality model ISO/IEC 25059:2023, these risks are respectively to the 'functional suitability' and 'usability' quality characteristics of the AI system. The suitability concerns identified by the organization would, in the terms of the quality model, primarily affect the functional 'appropriateness' of the model, given its native biases. The usability concerns identified by the organization would, in terms of the quality model, primarily affect the user controllability and transparency of the system.

As an example of how management frameworks may offer guidance for differentiated benchmarking of such a system, we consider the guidance available in ISO/IEC 42001. Clause 6.1.2 in ISO/IEC 42001 describes risk assessment processes to be followed by organizations. We may presume that the risks identified above would have emerged from a similar process. Clause 6.1.3 , in turn, outlines risk treatment processes, which primarily consists of the development of a list of controls to be implemented as part of the treatment process. An informative reference list of controls is provided for additional guidance in Annex A of 42001.

However, this list of controls does not provide guidance useful for producing quantitative benchmarks helpful for addressing the specific risks documented for this case example. For example, with respect to the concern for possible bias in data used for training the algorithms, the most relevant control advised in the standard relates to 'assessing AI system impact on individuals and groups of individuals', which simply exhorts organizations to assess and document the potential impacts of the system to individuals or groups of individuals throughout the system's life cycle. While doing so is certainly necessary for understanding the nature of bias present in the system's algorithms, it is not sufficient. In particular, the standard is silent on any specific benchmarks that may be identified for the specific system quality characteristic, e.g. group-specific precision or F1 scores.

The Annex also helpfully suggests that organizations document information about the data resources utilized for the AI system, which again, is a necessary ingredient in understanding the source of potential bias in the system's algorithms, but is by itself insufficient at measuring or characterizing the bias via benchmarking.

As an example of how classification-based frameworks may offer guidance for differentiated benchmarking for such a system, we consider the example of the German Data Ethics Commission's framework. As a tool intended for use within a private organization, for recommending (but not coercing) training modules for employees, it is reasonable to presume that this system would be considered either a Class 1 (no harm) or Class 2 (minimal harm) system, thereby imposing duties of transparency, correctness checks and system analysis of problematic cases, on the system providers in the former case, and a risk assessment mandate in the latter case. We may presume that the risk assessment process administered given a Class 2 adjudication by the framework would lead to a similar process as advised in Clause 6.1.2 of ISO/IEC 42001 or similar advice from a similar management standard.

However, the deficit in guidance observed upon above for management frameworks would then be inherited by this approach, thereby implying that current standardization approaches produce necessary, but not sufficient methodology for appropriate differentiated benchmarking of AI systems such as the one presented in this case example.

7.2 Case Example 2: User intent recognition

An organization is using an AI chatbot to offer online customer service, relying on AI algorithms to correctly recognize the problem-solving intent of customers based on text inputs, and offering actionable solution templates (use case 43; ISO/IEC TR 24030:2024) . The performance of the system is measured using the accuracy of intent recognition, the rate at which questions asked by customers are satisfactorily addressed by the system, and satisfaction ratings obtained from customers about system performance.

The organization identifies instances of high semantic ambiguity, and the presence of code-mixed expressions, i.e. expressions involving use of multiple languages, as potential system vulnerabilities. From the perspective of the AI software quality model ISO/IEC 25059:2023, these risks are respectively to the 'functional suitability' and 'usability' quality characteristics of the AI system. The suitability concerns identified by the organization would, in the terms of the quality model, primarily affect the functional 'completeness' of the model, given its limitations in handling code-mixed queries. The usability concerns identified by the organization would, in terms of the quality model, primarily affect the 'user error protection' of the system.

As an example of how management frameworks may offer guidance for differentiated benchmarking of such a system, we consider the guidance available in ISO/IEC 42001. Clause 6.1.2 in ISO/IEC 42001 describes risk assessment processes to be followed by organizations. We may presume that the risks identified above would have emerged from a similar process. Clause 6.1.3 , in turn, outlines risk treatment processes, which primarily consists of the development of a list of controls to be implemented as part of the treatment process. An informative reference list of controls is provided for additional guidance in Annex A of 42001.

However, this list of controls does not provide guidance useful for producing quantitative benchmarks helpful for addressing the specific risks documented for this case example. For example, with respect to the concern for error correction when system behaves erratically under conditions of high semantic ambiguity, the most relevant control advised in the standard relate to 'AI system verification and validation', which simply exhorts organizations to define and document verification and validation measures for the AI system and specify criteria for their use. While doing so is certainly necessary for understanding the system's behavior for known instances of semantic ambiguity, it is not sufficient. In particular, the standard is silent on any specific benchmarks that may be identified for the specific system quality characteristic, e.g. Average number of prompts needed to recover from ambiguity errors.

The Annex also helpfully suggests that organizations document information about the data resources utilized for the AI system, which again, is a necessary ingredient in understanding the speed of recovery from errors in the system's algorithms, but is by itself insufficient at measuring or characterizing this quality characteristic via benchmarking.

As an example of how classification-based frameworks may offer guidance for differentiated benchmarking for such a system, we consider the example of the German Data Ethics Commission's framework. As a tool intended for use in a purely informative function for a company's customers, it is reasonable to presume that this system would be considered either a Class 1 (no harm) or Class 2 (minimal harm) system, thereby imposing duties of transparency, correctness checks and system analysis of problematic cases, on the system providers in the former case, and a risk assessment mandate in the latter case. We may presume that the risk assessment process administered given a Class 2 adjudication by the framework would lead to a similar process as advised in Clause 6.1.2 of ISO/IEC 42001 or similar advice from a similar management standard.

However, the deficit in guidance observed upon above for management frameworks would then be inherited by this approach, thereby implying that current standardization approaches produce necessary, but not sufficient methodology for appropriate differentiated benchmarking of AI systems such as the one presented in this case example.

7.3 Case Example 3: Generation of clinical pathways

An AI system provider proposes the use of a temporal data mining techniques to develop a set of clinical pathways to be used for scheduling management of clinical care. The system clusters sequences of nursing orders, obtained from a hospital information management system, into a low-dimensional representation. It then uses clinical metadata to identify important features from these low-dimensional clusters, which enables the system to identify groups of activities relevant for clinical workflows in the hospital, along with their expected time courses. The system is expeceted to recommend appropriate interventions to smoothen a patient's journey through the hospital system. Its performance is measured using the complexity of the clinical pathways created, and patients' length of stay in the hospital.

The provider self-identifies the possible presence of biases within and across hospital as a significant source of risk for their system. From the perspective of the AI software quality model ISO/IEC 25059:2023, this risk would broadly affect the 'functional suitability', 'usability', 'reliability' and 'portability' quality characteristics of the AI system. The suitability concerns identified by the organization would, in the terms of the quality model, primarily affect the functional 'adaptability' of the model, given the heterogeneity of workflows within and across hospitals. The usability concerns identified by the organization would, in terms of the quality model, primarily affect the 'learnability' of the system. The reliability concerns identified by the organization would, in terms of the quality model, primarily affect the 'robustness' of the system. The portability concerns identified by the organization would, in terms of the quality model, primarily affect the 'adaptability' of the system.

As an example of how management frameworks may offer guidance for differentiated benchmarking of such a system, we consider the guidance available in ISO/IEC 42001. Clause 6.1.2 in ISO/IEC 42001 describes risk assessment processes to be followed by organizations. We may presume that the risks identified above would have emerged from a similar process. Clause 6.1.3 , in turn, outlines risk treatment processes, which primarily consists of the development of a list of controls to be implemented as part of the treatment process. An informative reference list of controls is provided for additional guidance in Annex A of 42001.

However, this list of controls does not provide guidance useful for producing quantitative benchmarks helpful for addressing the specific risks documented for this case example. For example, with respect to the concern for adaptability of the system to patient- and locale-specific heterogeneities, the most relevant control advised in the standard relate to 'assessing AI system impact on individuals and groups of individuals', which simply exhorts organizations to assess and document the potential impacts of the system to individuals or groups of individuals throughout the system's life cycle. While doing so is certainly necessary for understanding the system's behavior with respect to patients with different clinical needs, it is far from sufficient. In particular, the standard is silent on any specific benchmarks that may be identified for the specific system quality characteristic, e.g. satisfaction ratings with treatment measured across hospitals.

The Annex also helpfully suggests that organizations document information about the data resources utilized for the AI system, which again, is a necessary ingredient in understanding the speed of recovery from errors in the system's algorithms, but is by itself insufficient at measuring or characterizing this quality characteristic via benchmarking.

As an example of how classification-based frameworks may offer guidance for differentiated benchmarking for such a system, we consider the example of the German Data Ethics Commission's framework. As a tool intended for use in guiding clinical decisions with respect to patients, it is reasonable to presume that this system would be considered either a Class 4 (high risk) or Class 5 (extreme risk, forbidden) system, thereby imposing duties of transparency, correctness checks and system analysis of problematic cases, on the system providers in the former case, and market exclusion requirements for regulators in the latter case. We may presume that the risk assessment process administered given a Class 2 adjudication by the framework would lead to a similar process as advised in Clause 6.1.2 of ISO/IEC 42001 or similar advice from a similar management standard.

However, the deficit in guidance observed upon above for management frameworks would then be inherited by this approach, thereby implying that current standardization approaches produce necessary, but not sufficient methodology for appropriate differentiated benchmarking of AI systems such as the one presented in this case example.

Annex A
(informative)

Definitions of benchmarking

Table A.1

No	Definition	Source
D1	single value representing an accepted reference value derived either from comparisons among participants or from literature, used for orientation Note 1 to entry: The benchmark may be determined collaboratively or individually. Note 2 to entry: By clustering , different benchmarks can occur for different peer groups. (benchmark)	ISO 17258, 3.2 ISO 24523, 3.2 ；ISO 24513, 3.7.1.1.2 ；ISO/TR 24514, 3.1
D2	process of comparing processes, performances or quality against practices of the same nature, under the same circumstances and with similar measures	ISO 41011, 3.8.5.1
D3	activity of comparing objects of interest to each other or against a benchmark to evaluate characteristic(s) activity of comparing objects or practices of interest to each other or against a benchmark to evaluate criteria (or characteristic)	ISO/IEC 29155-1:2011; , 3.2 ISO/IEC 29155-1:2017, 3.3; ISO/IEC/IEEE 24765:2017, 3.363 ISO/IEC 18520:2019, 3.1.1
D4	reference point or metric against which process , performance or quality can be measured (benchmark)	ISO 41011:2024, 3.8.5 ISO 14031:2021, 3.4.8
D5	reference point against which comparisons can be made (benchmark)	ISO/IEC 29155-1 ,2.1 ISO 17258, 3.1 ISO/IEC 29155-1, 3.2; ISO 14031, 3.4.8; EN ISO 14050, 3.2.15 ISO 21678, 3.2; ISO 21931-1, 3.2.16 ISO/IEC/IEEE 24765, 3.362
D6	any standard or reference by which others can be measured	ISO/TS 18667, 3.1.1

D7	comparative evaluation or analysis of similar operational practices	ISO/TR 24514, 3.1 EN ISO 14644-16, 3.3.1 ISO 10010, 3.4
D8	activity of measurement and analysis that an organization can use to search for and compare practices inside and outside the organization, with the aim of improving its performance	ISO 10014, 3.8 ISO 30400, 3.1.18
D9	comparing attributes, processes or performance between organizations	ISO 30400, 3.17 ISO 32210, 3.34
D10	tool for performance improvement through systematic search and adaptation of leading practices	ISO 20468-1, 3.1.2
D11	standard against which results can be measured or assessed (benchmark)	ISO/IEC 25010, 4.3.2; ISO/IEC/IEEE 24765, 3.362
D12	method for comparing the performance of the leading organizations in a market segment	ISO 13053-2, 2.1
D13	procedure, problem, or test that can be used to compare systems or components to each other or to a standard(benchmark)	ISO/IEC/IEEE 24765, 3.362
D14	the comparison of actual or planned practices, such as processes and operations, to those of comparable organizations to identify best practices, generate ideas for improvement, and provide a basis for measuring performance	ISO/IEC/IEEE 24765, 3.363

Annex B
(informative)

Sample levels of specification

Complementary to the aims of ISO/IEC 25059:2023, which defines quality characteristics broadly desirable for all AI systems, this sample definition of a levels of specification workflow focuses on mapping minimal subsets of quality characteristics that are desirable for AI systems with different contexts-of-use. In particular, the following quality characteristics defined in ISO/IEC 25059:2023 are used as the 'properties' used to define property-action matrices in this project:

Table B.1

Functional Suitability	Performance Efficiency	Compatibility	Usability
Reliability	Security	Maintainability	Portability

AI systems' stakeholders have been defined in ISO/IEC 22989:2022 (Section 5.17), and their roles in the AI ecosystem have been additionally delineated in ISO/IEC TR 24028:2020 (Section 8.1). By providing a framework for defining differentiated levels of specifications, commensurate with the expected context-of-use of AI systems, such a standardization approach is expected to rationalize the compliance and regulatory burden on AI providers, while maintaining system trustworthiness for AI customers and AI partners (as defined in ISO/IEC 22989:2022).

Table B.2 — Property-action matrix for low-risk systems

		Properties							
		Functional Suitability	Performance Efficiency	Compatibility	Usability	Reliability	Security	Maintainability	Portability
	Model Description	ISO/IEC AWI TR 5469, **ISO/IEC CD 42001**			**ISO/IEC AWI TS 6254**	**ISO/IEC AWI TS 6254**			
	Model Evaluation	**ISO/IEC 4213**				**ISO/IEC 4213**			
	Data integrity								
	Data stratification								
	Baseline definition								
	Transparent Evaluation								
	Decision-making oversight								
Actions	Failure correlation modelling								

Annex C
(informative)

Descriptions of measures

1. Accuracy: Measures the proportion of correct predictions made by the model, commonly used in classification tasks.

2. Precision: Measures the proportion of true positive predictions among all positive predictions made by the model, indicating the model's ability to avoid false positives.

3. Recall: Measures the proportion of true positive predictions among all actual positive instances, indicating the model's ability to capture all positive instances.

4. F1 Score: Harmonic mean of precision and recall, providing a balance between the two metrics in classification tasks.

5. Specificity: Measures the proportion of true negative predictions among all actual negative instances, indicating the model's ability to avoid false negatives.

6. ROC Curve (Receiver Operating Characteristic Curve): Graphical representation of the trade-off between true positive rate and false positive rate across different classification thresholds.

7. AUC (Area Under the ROC Curve): Quantitative measure of the model's ability to distinguish between classes, with a higher value indicating better performance in classification tasks.

8. Mean Absolute Error (MAE): Average of the absolute differences between predicted and actual values, commonly used in regression tasks.

9. Mean Squared Error (MSE): Average of the squared differences between predicted and actual values, providing a measure of prediction accuracy in regression tasks.

10. Root Mean Squared Error (RMSE): Square root of the MSE, providing a measure of the typical error magnitude in regression tasks.

11. Cross-Entropy Loss: Measures the difference between predicted and actual probability distributions, commonly used in classification tasks.

12. KL Divergence (Kullback-Leibler Divergence): Measures the difference between two probability distributions, commonly used in model training and evaluation.

13. IoU (Intersection over Union): Measures the overlap between predicted and ground truth bounding boxes or segmentation masks, commonly used in object detection and segmentation tasks.

14. BLEU Score (Bilingual Evaluation Understudy Score): Measures the quality of machine-translated text by comparing it to reference translations, commonly used in natural language processing tasks.

15. Perplexity: Measures the uncertainty of a language model, with lower values indicating better performance in natural language processing tasks.

16. Top-k Accuracy: Measures the proportion of correct predictions when considering the top k most likely classes, providing a measure of the model's top-ranking accuracy in classification tasks.

17. Execution Time: Measures the time taken for the model to make predictions or perform inference on a given dataset, indicating the model's efficiency.

18. Sensitivity: Measures the proportion of true positive predictions among all actual positive instances, commonly used in medical diagnostics and anomaly detection tasks.

19. Precision-Recall Curve: Graphical representation of the trade-off between precision and recall across different classification thresholds, providing insights into the model's performance at different operating points.

20. Average Precision (AP): Computes the average precision across all recall values, providing a single scalar metric for precision-recall curves in classification tasks.

21. Mean IoU (Intersection over Union): Average of IoU scores computed across multiple instances or classes, commonly used in image segmentation tasks.

22. Dice Coefficient: Measures the similarity between two samples, commonly used in image segmentation tasks.

23. Cohen's Kappa: Measures inter-rater agreement for categorical data, correcting for agreement occurring by chance, commonly used in classification tasks.

24. AUC-PR (Area Under the Precision-Recall Curve): Quantitative measure of the model's ability to balance precision and recall across different thresholds in classification tasks.

25. Spearman Rank Correlation Coefficient: Measures the strength and direction of association between two ranked variables, commonly used in correlation analysis.

26. Sørensen-Dice Index: Measures the spatial overlap between two samples, often used in image segmentation tasks.

27. R2 Score (Coefficient of Determination): Measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s), commonly used in regression tasks.

28. Cohen's Kappa (Weighted): A variant of Cohen's Kappa that considers weighted agreement for categorical data, useful when categories have different levels of importance.

29. Adjusted Rand Index: Measures the similarity between two sets of clustering results, correcting for chance agreement, commonly used in clustering tasks.

30. Jaccard Index (Intersection over Union): Measures the similarity between two sets, computed as the size of the intersection divided by the size of the union, commonly used in clustering and similarity analysis.

31. Hamming Loss: Measures the fraction of labels that are incorrectly predicted, averaged over all samples and classes, commonly used in multi-label classification tasks.

32. Mean Average Precision (mAP): Computes the average precision across multiple classes or categories, providing a single metric for multi-class classification tasks.

Bibliography

[1] ISO/IEC 29155-1:2017, *Systems and software engineering - Information technology project performance benchmarking framework - Part 1: Concepts and definitions*

[3] 2937:2022, *IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems*

[4] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. Advances in Neural Information Processing Systems, 28.

[5] Leveson, N. (2020). Are you sure your software will not kill anyone?. Communications of the ACM, 63(2), 25-28.

[6] NIST AI 100-1:2023, *Artificial Intelligence Risk Management Framework*

[7] Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine, 6(1), 135.

[8] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys (CSUR), 55(3), 1-44.

[9] Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., ... & Gadepally, V. (2023, September). From words to watts: Benchmarking the energy costs of large language model inference. In 2023 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1-9). IEEE.

[10] Casner, S. M., Geven, R. W., Recker, M. P., & Schooler, J. W. (2014). The retention of manual flying skills in the automated cockpit. Human factors, 56(8), 1506-1516.

[11] Casner, S. M., Hutchins, E. L., & Norman, D. (2016). The challenges of partially automated driving. Communications of the ACM, 59(5), 70-77.

[12] ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*

[13] Hert, C. A. (1994). Information Technology Standardization: A Classification Process?. Advances in Classification Research Online, 95-110.

[14] Shen, L. (2014). The NIST cybersecurity framework: Overview and potential impacts. Scitech Lawyer, 10(4), 16.

[15] Cox, Anthony (2008), What's Wrong with Risk Matrices, Risk Analysis 28(2):497–512.

[16] Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological science, 15(8), 534-539.

[17] Yoo, S., Gregorian, D., Kopeikin, A., Leveson, N. (2023). Improving the Risk Matrix. In: Karakoc, T.H., Yilmaz, N., Dalkiran, A., Ercan, A.H. (eds) New Achievements in Unmanned Systems. ISUDEF 2021. Sustainable Aviation. Springer, Cham.

[18] ISO 17258:2015, *Statistical methods — Six Sigma — Basic criteria underlying benchmarking for Six Sigma in organisations*

[19] ISO 24523:2017, *Service activities relating to drinking water supply systems and wastewater systems — Guidelines for benchmarking of water utilities*

[20] ISO 24513:2019, *Service activities relating to drinking water supply, wastewater and stormwater systems — Vocabulary*

[21] ISO/TR 24514:2018, *Activities relating to drinking water and wastewater services — Examples of the use of performance indicators using ISO 24510, ISO 24511 and ISO 24512 and related methodologies*

[22] ISO 14031:2021, *Environmental management — Environmental performance evaluation — Guidelines*

[23] EN ISO 14050:2020, *Environmental management - Vocabulary (ISO 14050:2020)*

[24] ISO 21678:2020, *Sustainability in buildings and civil engineering works — Indicators and benchmarks — Principles, requirements and guidelines*

[25] ISO 21931-1:2022, *Sustainability in buildings and civil engineering works — Framework for methods of assessment of the environmental, social and economic performance of construction works as a basis for sustainability assessment — Part 1: Buildings*

[26] ISO/IEC/IEEE 24765:2017, *Systems and software engineering - Vocabulary*

[27] ISO/TS 18667:2018, *Space systems — Capability-based Safety, Dependability, and Quality Assurance (SD&QA) programme management*

[28] ISO/TR 24514:2018, *Activities relating to drinking water and wastewater services — Examples of the use of performance indicators using ISO 24510, ISO 24511 and ISO 24512 and related methodologies*

[29] EN ISO 14644-16:2019, *Cleanrooms and associated controlled environments - Part 16: Energy efficiency in cleanrooms and separative devices (ISO 14644-16:2019)*

[30] ISO 10010:2022, *Quality management — Guidance to understand, evaluate and improve organizational quality culture*

[31] ISO 10014:2021, *Quality management systems — Managing an organization for quality results — Guidance for realizing financial and economic benefits*

[32] ISO 30400:2022, *Human resource management — Vocabulary*

[33] ISO 30400:2016, *Human resource management — Vocabulary*

[34] ISO 32210:2022, *Sustainable finance — Guidance on the application of sustainability principles for organizations in the financial sector*

[35] ISO 20468-1:2018, *Guidelines for performance evaluation of treatment technologies for water reuse systems — Part 1: General*

[36] ISO/IEC 25010:2011, *Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models*

[37] ISO 13053-2:2011, *Quantitative methods in process improvement — Six Sigma — Part 2: Tools and techniques*