electronics

MDPI

Review

Quality Assessment of Artificial Intelligence Systems: A Metric-Based Approach

Oleksandr Gordieiev [1], Daria Gordieieva [2,*], Austen Rainer [2], Anatoliy Gorbenko [1] and Olga Tarasyuk [3]

[1] School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds LS1 3HE, UK
[2] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT9 5BN, UK
[3] School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK
* Correspondence: d.gordieieva@qub.ac.uk

Abstract

This paper addresses the growing need for reliable methods to evaluate the quality of artificial intelligence (AI) systems as they become widely used in both critical domains and everyday applications. The study aims to develop a metric-based approach to assessing AI system quality by harmonising product quality and quality in use models in line with updated international standards. To achieve this, the authors analyse existing ISO/IEC 25000 series standards, identifying inconsistencies between older and newer versions, and propose an updated quality model that incorporates both perspectives. Building on guidance documents, international standards, and contemporary research, the study introduces a set of metrics designed to measure new subcharacteristics of AI system quality, particularly where standardised metrics have not yet been developed. The proposed approach bridges the gap between established quality models (ISO/IEC 25010:2023, ISO/IEC 25019:2023, ISO/IEC 25059:2023) and standardised measurement practices (ISO/IEC 25023:2016, ISO/IEC 25022:2016), enabling more consistent and practical evaluation of AI systems. These metrics can be applied by researchers and practitioners to improve the quality of AI systems, enhance their reliability, and reduce risks associated with insufficient quality. Future work will focus on empirical validation of the proposed approach to confirm its applicability and usefulness across diverse AI applications.

Keywords: artificial intelligence systems; quality assessment; quality models; quality models in use; metric-based approach

Academic Editor: George A. Tsihrintzis

Received: 16 December 2025
Revised: 31 January 2026
Accepted: 2 February 2026
Published: 5 February 2026

1. Introduction

1.1. Motivation

It can be considered that the preliminary stage in the formation of artificial intelligence (AI) spans the period from 1936, when Alan Turing formalised the concept of "computability" [1], to 1950, when he published his paper "Computing Machinery and Intelligence" [2]. The formal beginning of artificial intelligence as a distinct field of mathematics and engineering is generally attributed to 1956, when John McCarthy introduced the term "artificial intelligence" during the Dartmouth Summer Research Project on Artificial Intelligence [3]. It is also important to acknowledge the substantial contribution of Norbert Wiener [4] to the foundations of artificial intelligence through his pioneering work in the field of cybernetics, particularly in the development of control theory and the study of

feedback mechanisms. These contributions have significantly influenced the conceptual and technical underpinnings of intelligent systems. Since then, the field has continued to evolve, with both mathematical models (methods) and engineering systems becoming increasingly complex.

Today, a wide variety of artificial intelligence models have been developed, serving diverse purposes, and hundreds of these models are already deployed in real-world applications across multiple domains—including medicine, defence, engineering, education, economics, and beyond [5]. In parallel, computational capabilities have significantly advanced, now reaching the exaflop scale in supercomputers performing tasks involving artificial intelligence. This trend, on the one hand, has led to a growing number of incidents associated with the insufficient quality of AI systems, and on the other, it has increased the demands placed on quality evaluation and assurance for such systems. The observed tendency also aligns with the Empirical Law of Software Reliability [6–8], which can be extended to AI systems: the more complex the software system, the more errors it is likely to contain.

The motivation for writing this paper stemmed from empirical studies that have demonstrated the adverse consequences of poor-quality AI systems. The most well-known failure cases are presented in Table 1. Usually, AI system malfunctions are associated with insufficient estimation accuracy (e.g., in grading or credit scoring), erroneous classification (e.g., of human images or movements), incorrect risk forecasting (including housing price predictions), and flawed or inappropriate recommendations (e.g., in medical or legal contexts).

Table 1. The most well-known incidents involving artificial intelligence systems.

#	Name (Country, Year)	Description	Participants	Consequences
1	Inaccurate algorithms for grading school examinations (A-levels/GCSE) (UK, 2020) [9,10]	The grade moderation algorithm was used to normalise results, as traditional A-level and GCSE examinations were cancelled due to the COVID-19 pandemic. It downgraded the grades of many pupils, exacerbating inequality. The results were later replaced with teacher-assessed grades	Ofqual (the examinations regulator in England), the Department for Education, schools, and university applicants	As a result of the algorithm's implementation, approximately 40% of grades were downgraded compared to teacher assessments. This led to protests, the withdrawal of the algorithm, and the publication of analytical reports and academic studies
2	Biased assessment of recidivism risk using the COMPAS algorithm (USA, 2016) [11]	COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a proprietary recidivism risk assessment algorithm widely used by courts in the United States. It exhibited racial disparities in its evaluation of the likelihood of reoffending	Equivant (ex-Northpointe), USA courts, researchers	The algorithm influenced decisions regarding parole, bail, and the length of sentences. For African Americans, the risk was higher. Broad academic debate, the reconsideration of approaches to fairness metrics, and the growing caution of courts and regulators towards AI "black boxes" followed
3	AI Solution for health care IBM Watson for Oncology (USA, 2013–2020) [12]	Low accuracy of recommendations in clinics, limited data, and issues with clinical integration	IBM, MD Anderson Cancer Center, and other clinics	IBM closed its Watson Health division, and industry lessons on the clinical validation of AI
4	Inaccurate housing price forecasting by Zillow's "Zillow Offers" service (USA, 2021) [13]	The algorithms overestimated housing prices, particularly under the sharp market volatility of 2020–2021	Zillow Group, home sellers/buyers	Business closure, losses—the company lost about $500 million and laid off around 2000 employees; the incident became a case study

Table 1. *Cont.*

#	Name (Country, Year)	Description	Participants	Consequences
5	Google Photos scandal with the "gorilla" label and the subsequent blocking of the category (USA, 2015) [14]	In a photo with two Black individuals, the algorithm applied the label "gorillas". A mistaken racist classification	Google Inc.	Reputational damage, the emergence of the notion of an "AI bias incident" as a distinct category of problems, and the rise in initiatives on AI fairness and incident databases
6	Allegations of gender bias in credit scoring in Apple Card (USA, 2019) [15]	Goldman Sachs launched the Apple Card. Women received significantly lower credit limits than their husbands/partners, even when they had a better credit history or a joint account	Apple, Goldman Sachs, and New York State regulators	Regulatory investigations, debate on the transparency of credit models
7	Inaccurate responses by Air Canada's chatbot (Canada, 2024) [16]	In 2022, a passenger consulted Air Canada's chatbot on the airline's website to clarify the rules on compensation in the event of the death of a close relative. Air Canada refused to provide compensation, claiming that the chatbot had "made a mistake" and was "not an official source."	Air Canada, passenger, British Columbia Civil Resolution Tribunal	Court ruling: Air Canada bears full legal responsibility for the responses of its own chatbot; the customer was awarded compensation This case set a precedent: companies cannot invoke an "AI error" as an external problem—everything stated by their AI system is equated with official information. The case is widely discussed by legal experts as an example that the era of the "irresponsible chatbot" has come to an end
8	"Fake" references from generative AI for a lawyer (USA, 2023) [17]	A lawyer filed a motion with non-existent references to precedents generated by AI	R. Mata, Avianca, Inc., U.S. District Court for the Southern District of New York (SDNY)	Fines and requirements for the verification of AI-generated materials in legal practice
9	Fatal collision involving an Uber self-driving car with a system developed by the Advanced Technologies Group (ATG) (USA, 2018) [18]	The system misclassified the pedestrian, while the emergency brakes had been disabled. At the same time, the driver was distracted by a phone	Uber Advanced Technologies Group (ATG), National Transportation Safety Board (NTSB), and victim Elaine Herzberg	Death of a person, NTSB report, and tightening of safety requirements for autonomous vehicle testing
10	Systematic errors in facial recognition (2018) [19]	The study revealed significant differences in gender classification accuracy in commercial facial recognition systems from IBM, Microsoft, and Face++. Errors were minimal for light-skinned men (up to 0.8%) but increased sharply for dark-skinned women, ranging from 20% to 35%	Joy Buolamwini and Timnit Gebru evaluated systems from Microsoft, IBM, and Face++	Sparked global debate on algorithmic fairness; led to improvements in commercial facial recognition systems; influenced policy discussions and regulation efforts

The consequences of the incidents presented cause both material and immaterial harm and range from socially significant damage (for example, erroneous assessments, unfair credit decisions, or misinformation provided by customer-facing chatbots) to threats to human life and health (for example, fatal accidents involving autonomous vehicles). Such diversity of consequences reflects not only the wide range of application domains in which AI systems are used, but also their specific technical properties, which distinguish them

from traditional computer systems and significantly complicate the tasks of ensuring and assessing their quality.

Although a detailed analysis of the incidents presented in Table 1 is beyond the scope of this study, it should be noted that the examples provided make it possible to identify a number of characteristic features of AI systems that may lead to potential negative consequences. First, in many cases, AI systems may function incorrectly outside the conditions represented in the training and test data, which leads to a loss of quality under real-world operating conditions. Second, AI systems are characterised by opaque decision-making mechanisms, which complicate the interpretability of their results, the identification of the causes of incorrect behaviour, and the implementation of corrective actions. Third, the behaviour of AI systems may change over time due to model retraining or changes in input data, which may result in the emergence of new, previously unidentified quality issues in such systems.

Unlike traditional computer systems, in AI systems, hazardous consequences may arise not only as a result of classical failures, but also due to incorrect functionality, including erroneous classification or incorrect risk assessment. This is particularly critical for safety-critical systems, in which such errors may directly lead to negative incidents, including threats to human life and health. At the same time, for general-purpose AI systems, negative consequences more often occur as immaterial harm, for example, moral, psychological, or social harm. Thus, the assessment of AI system safety should be regarded as an integral aspect of comprehensive quality assessment. In addition to safety-related challenges, the presented examples of incidents show that issues related to functional correctness, robustness, and transparency of such systems pose additional challenges in the deployment and use of AI systems.

For a more in-depth analysis of incidents related to the use of artificial intelligence, it is recommended to refer to open sources such as the AI Incident Database [20]. A detailed examination of such cases lies beyond the scope of this paper. The examples provided (Table 1) demonstrate that the current level of quality of AI systems remains insufficient, which may lead to socially significant undesirable consequences. In this regard, a comprehensive assessment of the quality of such systems becomes a primary task, as a necessary condition for their reliable and safe implementation.

1.2. Background

Standards of the ISO/IEC 25000 series are widely used for assessing the quality of traditional software. However, these standards cannot be directly applied to AI systems, owing to the specific characteristics of such systems. AI systems directly depend on the quality and quantity of data used for model training, validation, and testing. Moreover, AI systems are probabilistic and are characterised by adaptiveness during operation, the opacity of decision-making (the so-called "black box" effect), as well as the potential capacity for self-learning and behavioural change over time. These features complicate the application of traditional quality metrics, which are designed to measure quality attributes and to support evaluation procedures oriented towards static and predictable components, and highlight the necessity of developing specialised approaches to the measurement of AI system quality.

When discussing the quality of AI systems, quality assessment should be divided into:

- assessment of data quality;
- assessment of the quality of algorithms and models;
- assessment of the quality of AI-based systems as a whole.

This paper is devoted to the assessment of the quality of AI systems; therefore, a metric-based approach to evaluating the quality of such systems will be considered further,

based on a quality model from two perspectives: product quality and quality in use. The product quality model describes quality from the perspective of the information and communication technology (ICT) product itself, while the quality in use model describes quality from the perspective of the end user.

In July 2023, IEC and ISO published a new standard, ISO/IEC 25059 [21], which provides a standardised framework for evaluating AI system quality, addressing AI-specific characteristics. As beforethe release of this standard, which introduced the AI product quality model and the AI quality in use model, the scientific literature -continues to present various extended quality models for AI systems, as well as studies examining individual characteristics/subcharacteristics of AI system quality [22,23]. Broadly, scientific research in this field can be divided into several groups:

- quality models for AI systems;
- quality models for AI systems for specific application domains;
- studies devoted to particular (sub)characteristics of AI system quality.

Quality models of AI systems may be developed in various ways, such as by adapting an existing (traditional) quality model to the specific characteristics of AI, extending it with new characteristics and subcharacteristics, generating a model based on the results of a systematic literature review, expert analysis, or empirical data obtained from stakeholder surveys or interviews. Most of the proposed AI system quality models are derived from the software quality model of ISO/IEC 25010 [24,25], extended with new quality characteristics and subcharacteristics considered by the authors to be specific to AI systems, and/or refined interpretations of existing characteristics [26–29]. For example, Kuwajima and Ishikawa [27] propose adapting the ISO/IEC 25000 quality model to the unique nature of machine learning by extending the existing (sub)characteristics and adding new quality (sub)characteristics in line with the ethical guidelines of the European Commission for trustworthy AI. Other methods of constructing quality models also exist. For instance, Indykov et al. [30], based on a systematic literature review, proposed a quality model for machine learning (ML) systems, which includes 11 frequently mentioned attributes, and compared it with the relevant standards ISO/IEC 25010 and ISO/IEC 25059. Siebert et al. [31] constructed a quality model for an ML system based on an industrial use case of Fujitsu's Accounting Center. Nakamichi et al. [32] proposed a requirements-driven method to determine the quality characteristics of machine learning software systems, identifying 34 ML systems quality characteristics, where 18 characteristics are specific to ML systems, while the others are for enterprise software systems.

There are emerging studies in which AI system quality models are adapted for specific application domains, although they remain limited in number. For example, Kelly et al. [33] proposed an extended quality model for AI products for safety-critical applications. Their model was expanded through: the addition of quality characteristics and subcharacteristics absent from the ISO/IEC 25059 quality model; alignment with the updated version of ISO/IEC 25010:2023 [24], to which ISO/IEC 25059 had not previously been updated; consideration of the requirements of the AI Act; and the inclusion of quality characteristics and subcharacteristics related to the safety of AI systems, based on industry standards such as ISO PAS 8800 for automotive systems [34].

A separate group of studies is devoted to examining various quality characteristics and subcharacteristics inherent to AI systems by means of different research methods. A popular method is conducting systematic literature reviews to identify specific quality characteristics of AI systems [23,35,36]. Qualitative interviews and surveys are also employed. For example, Habibullah et al. [37] conducted a qualitative study using interviews and a survey of experts from industry and academia to identify non-functional requirements for machine learning.

Most of the mentioned models were developed before the publication of the ISO/IEC 25059 standard, initiating a discussion among researchers and practitioners on the need to develop a standardised, technology-oriented quality model for AI systems. Some of the main characteristics and subcharacteristics proposed in the author-defined models were subsequently incorporated into the quality models presented in ISO/IEC standards. However, with the publication of the latest standards dedicated to quality models for software and AI systems [21,24] new challenges have emerged related to inconsistencies between the characteristics and subcharacteristics of the quality models defined in these standards.

The main limitation of the proposed AI system quality models is the insufficient degree of development of methods for measuring the proposed quality (sub)characteristics. In particular, there is a limited number of studies addressing how quality attributes can be measured [23,37,38]. Researchers note that the focus should not only be on selecting terms for AI system quality (sub)characteristics and prioritising different characteristics for AI systems in various domains, but also on the development of specific metrics and methods for their measurement [30], emphasising the abstractness of existing models, which makes them less valuable for practitioners [31]. The literature also highlights that, due to the non-deterministic behaviour of AI systems, there are certain challenges in measuring quality (sub)characteristics, which are explained by the following: the difficulty of quantitatively assessing some quality (sub)characteristics, the absence of baseline measurement indicators, the complexity of the AI ecosystem, data quality, testing costs, result bias, and the dependence of quality assessment on the domain of the AI system [37].

In the literature, some authors propose only quality models for AI systems without providing metrics for assessing (sub)characteristics. Refs. [26,28,30,33,39] considering the development of metrics as a direction for future research. In most cases, when studying AI system quality attributes, authors focus on measuring one or several software product quality attributes that they regard as the most important, and propose approaches to measuring these attributes while taking into account the specific features of AI systems [29,32,35,38,40–42]. There are also studies examining the problem of evaluating trustworthy AI [43–46], in which the authors suggest various metrics for the technical characteristics of AI systems considered within the trustworthy AI framework, such as robustness, transparency, safety, and others; however, the formulas for calculating these metrics are often not provided. Some authors propose methods for adapting standardised metrics for traditional software to the nature of AI systems [27], but often do not provide metrics for measuring the new characteristics they introduce. At the same time, there exist author-developed AI system quality models that are not based on ISO/IEC 25010 but, for example, on an industrial use case, in which their own quality measures are proposed for quality (sub)characteristics that are not aligned with standardised (sub)characteristics [31]. Since the publication of ISO/IEC 25059, new studies have emerged in which authors propose metrics for newly standardised AI system quality characteristics [47], as well as works in which authors attempt to map existing metrics for evaluating AI systems in a specific domain (e.g., generative AI systems) to standardised quality characteristics [36].

In most of the reviewed works, it is noted that research on the quality issues of AI systems and the measurement of quality (sub)characteristics is still at an early stage and requires further investigation. Despite the existence of various metrics, the current literature lacks a standardised measurement approach for the newly standardised quality (sub)characteristics inherent to AI systems, which significantly complicates their practical application and the comparability of the obtained assessments.

1.3. Aim, Objectives, and Structure

The aim of this paper is to develop a metric-based approach to assessing the quality of artificial intelligence systems on the basis of standardised quality models.

To achieve this aim, the paper addresses the following objectives:

1. To analyse existing international standards regulating quality models and procedures for measuring quality (sub)characteristics of AI systems, with a focus on identifying inconsistencies between the new and outdated versions of the ISO/IEC 25000 series standards.
2. To develop an updated quality model for AI systems from two perspectives—product quality and quality in use—aligning it with the updated version of international standards.
3. To develop a justified set of metrics for the evaluation of quality characteristics and subcharacteristics of AI systems in accordance with the current versions of international standards in the field of AI systems and existing best practices in AI system quality assessment.

The rest of this paper is organised as follows. Section 2 presents the development of a standardised technology-oriented quality model for AI systems. Section 3 introduces metrics for product quality assessment. Section 4 presents metrics for assessment of quality in use of AI systems. Section 5 discusses the findings and the direction of future work. The conclusions are in Section 6.

2. Development of a Standardised Technology-Oriented Quality Model for Artificial Intelligence Systems

2.1. ISO/IEC Standards Overview

The object of the study is information systems with artificial intelligence. Artificial intelligence systems are engineered systems that generate outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives [48]. Since this paper considers the assessment of the quality of such systems and the quality of their use, we present the corresponding definitions.

AI product quality is the degree to which an AI system meets specified requirements for quality when used under specified conditions, including both general software characteristics (e.g., reliability, security, usability) and AI-specific characteristics (e.g., robustness, transparency, or intervenability). **Quality in use of an AI system** is the degree to which an AI system enables users to achieve their goals effectively, safely, efficiently, and with trust in a specified context of AI system use. The quality of AI systems is described by means of a quality model, a defined set of characteristics, and the relationships between them, which provides a framework for specifying quality requirements and evaluating quality [24]. The quality model is represented by two models: the product quality model (from the perspective of the ICT product itself) and the quality in use model (from the perspective of the end user). Let us consider them in more detail.

The new standard ISO/IEC 25059 [21] for quality of AI systems describes two models: a product quality model and a quality-in-use model. Evolutionarily, these models are a development of the traditional models of software quality and quality in use, presented within the ISO/IEC 25000 series of standards, with certain modifications that take into account the specific characteristics of AI. The AI system quality model and the quality-in-use model are two level-quality models that include quality characteristics and subcharacteristics specific to AI, as well as those already defined in the software quality standard—ISO/IEC 25010. However, the authors of the standard ISO/IEC 25059:2023 [21] based it not on the new (latest) standards ISO/IEC 25010:2023 “Product Quality Model” [24] and ISO/IEC 25019:2023 “Quality in use Model” [49], but on their earlier version ISO/IEC 25010:2011 “System

and Software Quality Models" [25], which is no longer in use. These gaps (1 and 2) are demonstrated in Figure 1a. At present, the ISO/IEC 25059 standard is under revision.

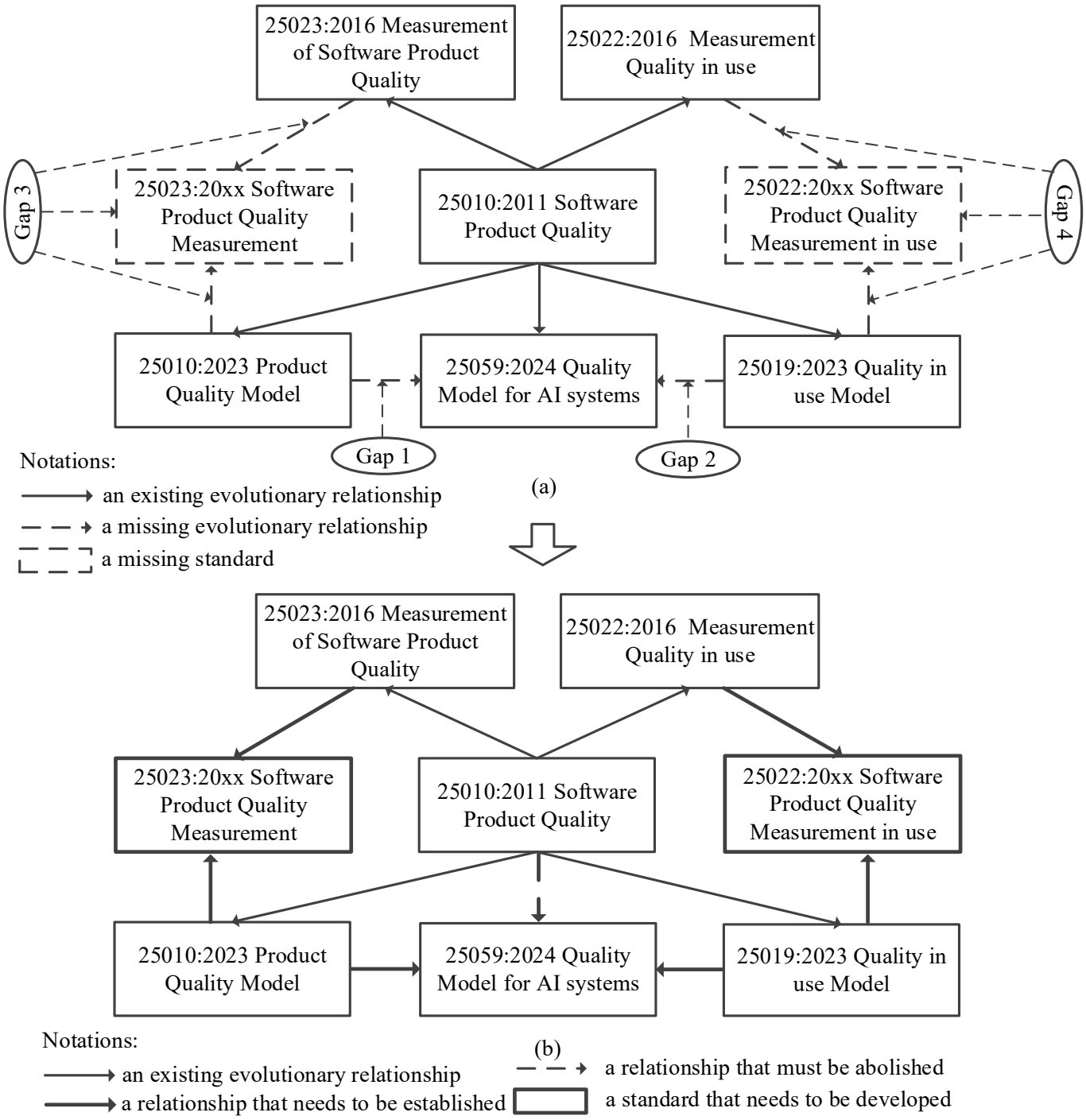

Figure 1. Evolutionary connections between ISO/IEC standards with gaps (**a**) and without gaps (potential alignment) (**b**).

The ISO/IEC 25059 standard contains only quality models and does not include metrics for measuring individual quality (sub)characteristics. In January 2024, the ISO/IEC TS 25058 standard [50] was published, which provides recommendations for the evaluation of AI systems using the AI system quality model and describes specific measures and indicators, although it does not provide concrete metrics. At the same time, it is based on the methodologies set out in a number of traditional ISO/IEC stan-

dards on product quality and quality in use measurement, that is, on the corresponding pair of standards: ISO/IEC 25023:2016 "Measurement of Software Product Quality" [51] and ISO/IEC 25022:2016 "Measurement of Quality in use" [52]. However, evolutionary linkages between the pairs of standards ISO/IEC 25023:2016 "Measurement of Software Product Quality" [51] and ISO/IEC 25010:2023 "Product Quality Model" [24], as well as ISO/IEC 25022:2016 "Measurement of Quality in use" [52] and ISO/IEC 25019:2023 "Quality in use Model" [49], are absent. The reason for this mismatch between standards is the significant delay in the release of new versions of the standards 25023:20xx "Measurement of Software Product Quality" and ISO/IEC 25022:20xx "Measurement of Quality in use". The next two gaps—Gap 3 and Gap 4—indicate the misalignments between these standards. Potential evolutionary linkages between standards without gaps are shown in Figure 1b.

As a result of the mismatch between the standards that describe quality models, part of the modified characteristics and subcharacteristics from the latest standard of the traditional software quality model did not enter the quality model for artificial intelligence systems (Gap 1, Figure 1a). In this regard, it is necessary to clarify the quality model for AI systems. To this end, we will conduct a comparative analysis between ISO/IEC 25010:2011, ISO/IEC 25010:2023, and ISO/IEC 25059:2023 and, based on the results obtained, present an updated quality model for AI systems. For this purpose, we will harmonise the characteristics and subcharacteristics in accordance with the following principles:

- All modified characteristics and subcharacteristics described in the standard ISO/IEC 25010:2023 and not included in this form in the standard ISO/IEC 25059:2024 must be included in the updated quality model of AI systems;
- All new characteristics and subcharacteristics of the quality model presented in the standard ISO/IEC 25059:2024, and reflecting the specific features of AI systems, must be included in the updated quality model of artificial intelligence systems.

2.2. Updated Product Quality Model Development

The results of the analysis of the quality models are presented in Table 2. The set of selected characteristics to be included in the updated quality model for AI systems is marked in grey with a bold border.

Table 2. Product quality models comparison.

#	Name of (Sub) Characteristics of Quality	Sources		
		ISO/IEC 25010:2011 [25]	**IS0/IEC 25010:2023 [24]**	**ISO/IEC 25059:2023 [21]**
1.	**Functional suitability**	+	+	+
1.1	Functional completeness	+	+	+
1.2	Functional correctness	+	+	+ (M)
1.3	Functional appropriateness	+	+	+
1.4	Functional adaptability			+ (NS)
2.	**Performance efficiency**	+	+	+
2.1	Time behaviour	+	+	+
2.2	Resource utilisation	+	+	+
2.3	Capacity	+	+	+
3.	**Compatibility**	+	+	+
3.1	Co-existence	+	+	+
3.1	Interoperability	+	+	+

Table 2. *Cont.*

#	Name of (Sub) Characteristics of Quality	Sources		
		ISO/IEC 25010:2011 [25]	**IS0/IEC 25010:2023 [24]**	**ISO/IEC 25059:2023 [21]**
4.	**Interaction capability**	+ (R, Usability)	+	+ (R, Usability)
4.2	Appropriateness recognisability	+	+	+
4.3	Learnability	+	+	+
4.4	Operability	+	+	+
4.5	User error protection	+	+	+
4.6	User engagement	+ (R, User interface aesthetics)	+	+ (R, User interface aesthetics)
4.7	Inclusivity	+ (S&R, Accessibility)	+	+ (S&R, Accessibility)
4.8	User assistance		+	
4.9	Self-descriptiveness		+ (NS)	
4.10	User controllability			+ (NS)
4.11	Transparency			+ (NS)
5.	**Reliability**	**+**	**+**	**+**
5.1	Faultlessness	+ (R, Maturity)	+	+ (R, Maturity)
5.1	Availability	+	+	+
5.2	Fault tolerance	+	+	+
5.3	Recoverability	+	+	+
5.4	Robustness			+ (NS)
6.	**Security**	+	+	+
6.1	Confidentiality	+	+	+
6.2	Integrity	+	+	+
6.3	Nonrepudiation	+	+	+
6.4	Accountability	+	+	+
6.5	Authenticity	+	+	+
6.6	Resistance		+ (NS)	
6.7	Intervenability			+ (NS)
7.	**Maintainability**	+	+	+
7.1	Modularity	+	+	+
7.2	Reusability	+	+	+
7.3	Analysability	+	+	+
7.4	Modifiability	+	+	+
7.5	Testability	+	+	+
8.	**Flexibility**	+ (R, Portability)	**+**	+ (R, Portability)
8.1	Adaptability	+	+	+
8.2	Scalability		+ (NS)	
8.3	Installability	+	+	+
8.4	Replaceability	+	+	+
9.	**Safety**		+ (NC)	
9.1	Operational constraint		+ (NS)	
9.2	Risk identification		+ (NS)	
9.3	Fail safe		+ (NS)	
9.4	Hazard warning		+ (NS)	
9.5	Safe integration		+ (NS)	

Notations: +—Present in the model, R—Renamed, S&R—Split and Renamed, NS—New Subcharacteristic, NC—New Characteristic, M—Modified.

The updated quality model for AI systems is presented in Figure 2. The definitions of quality (sub)characteristics of the AI system product quality model can be found in the ISO/IEC 25059 and ISO/IEC 25010 standards, while definitions of the newly introduced (sub)characteristics are provided in Section 3 when developing evaluation metrics for them.

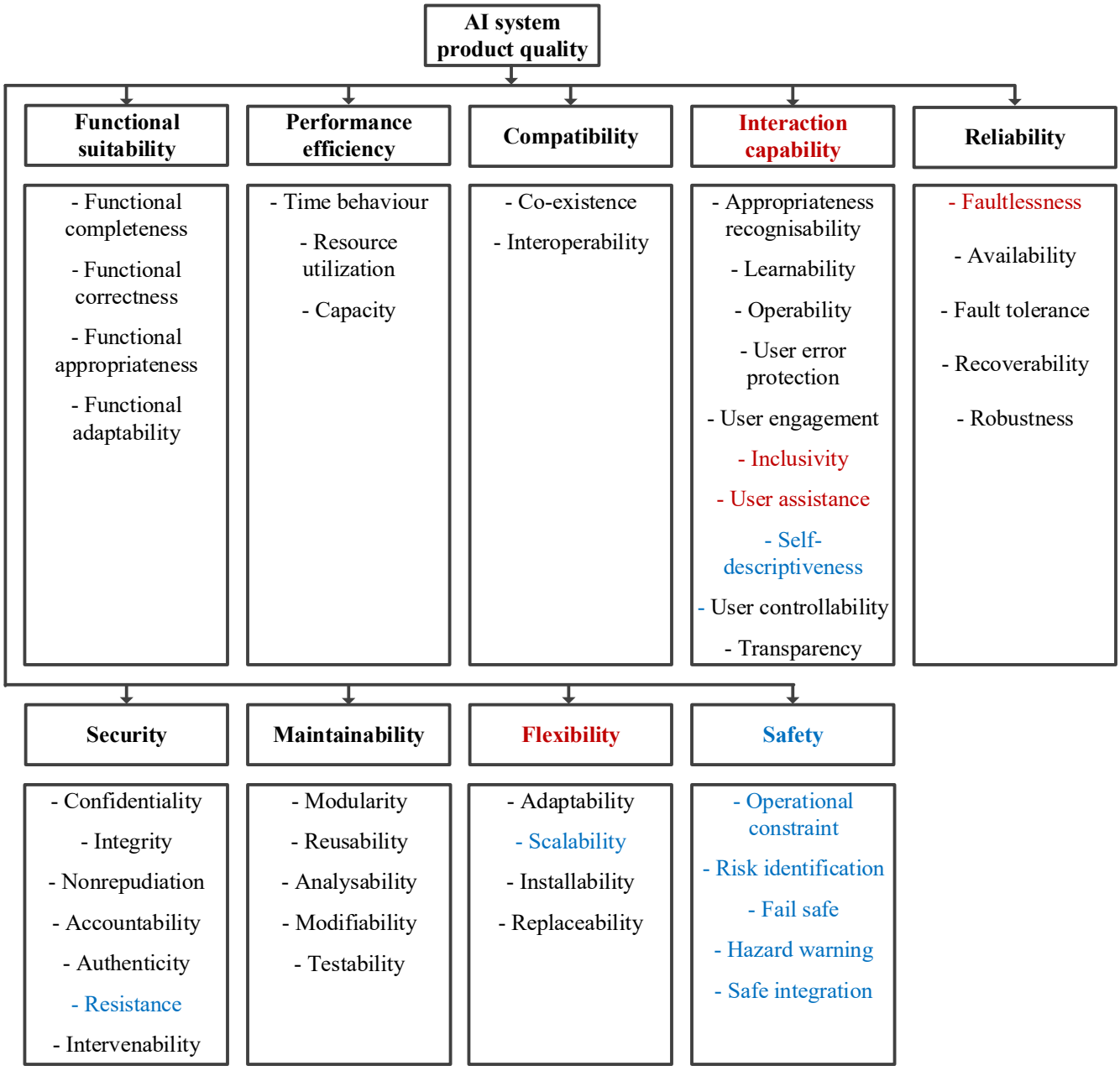

Figure 2. AI Systems product quality model (updated) (new (sub)characteristics absent from the ISO/IEC 25059 product quality model are marked in blue; renamed (sub)characteristics are marked in red).

This model is constructed by addressing the inconsistencies between the ISO/IEC standards [21,24] and represents an updated, standardised product quality model for AI systems.

2.3. Updated Quality in Use Model Development

We will conduct a similar comparison for the quality in use model of AI systems (Table 3).

Table 3. Quality in use models comparison.

#	Name of (Sub) Characteristics of Quality in Use	Sources		
		ISO/IEC 25010:2011 [25]	**IS0/IEC 25019:2023 [49]**	**ISO/IEC 25059:2024 [21]**
1	**Beneficialness**		+	
1.1	Usability	+ (M&R Effectiveness, Efficiency, Satisfaction)	+	
1.1.1	Effectiveness	+		+
1.1.2	Efficiency	+		+
1.1.3	Usefulness	+		+
1.1.4	Trust	+		+
1.1.5	Pleasure	+		+
1.1.6	Comfort	+		+
1.1.7	Transparency			+
1.2	Accessibility		+	
1.3	Suitability		+	
2	**Freedom from risk**	+	+	+
2.1	Freedom from economic risk	+ (R, Economic risk mitigation)	+	+ (R, Economic risk mitigation)
2.2	Freedom from health risk	+ (S&R, Health and safety risk mitigation)	+	+ (S&R, Health and safety risk mitigation)
2.3	Freedom from human life risk		+	
2.4	Freedom from environmental and societal risk	+ (M&R, Environment risk mitigation, Societal and ethical risk mitigation)	+	+ (M&R, Environment risk mitigation, Societal and ethical risk mitigation)
3	**Acceptability**		+	
3.1	Experience		+	
3.2	Trustworthiness		+	
3.3	Compliance		+	
4	**Context coverage**	+		+
4.1	Context completeness	+		+
4.2	Flexibility	+		+

Notations: +—Present in the model, R—Renamed, S&R—Split and Renamed, M&R—Merged and Renamed, NS—New Subcharacteristic, NC—New Characteristic, M&R—MoVed and Renamed.

The updated quality-in-use model for AI systems is presented in Figure 3. The definitions of quality (sub)characteristics of the AI system quality-in-use model can be found in the ISO/IEC 25059 and ISO/IEC 25019 standards, while definitions of the newly introduced (sub)characteristics are provided in Section 4 when developing evaluation metrics for them.

This model is constructed by addressing the inconsistencies between the ISO/IEC standards [21,49] and represents an updated, standardised *quality in use* model for AI systems.

We have examined which characteristics and subcharacteristics are included in the updated product quality model and quality model in use for AI systems. We will now determine the set of metrics by which we can measure them.

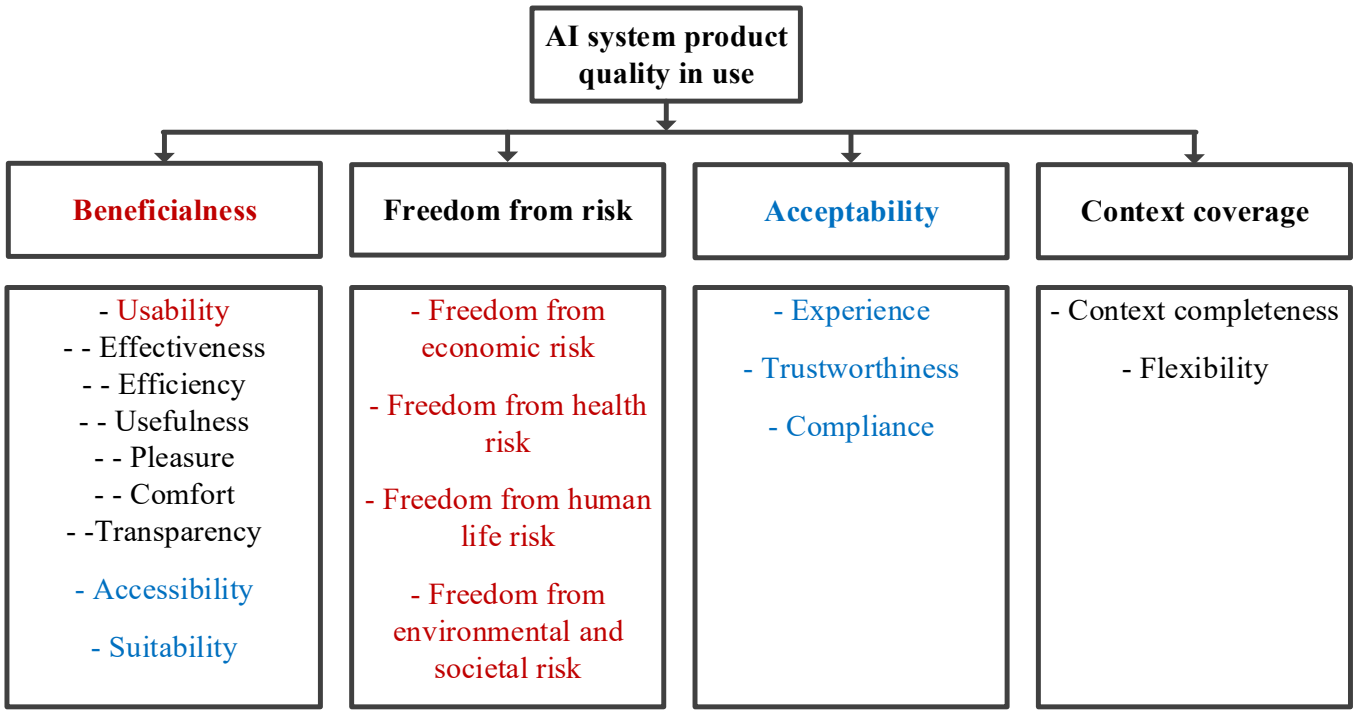

Figure 3. AI System quality in use model (updated) (new (sub)characteristics absent from the ISO/IEC 25059 quality in use model are marked in blue; renamed (sub)characteristics are marked in red).

3. AI System Product Quality Model Measures

The metrics proposed in this paper can be broadly divided into three groups:

1. For quality (sub)characteristics carried over unchanged from the ISO/IEC 25010 software product quality model into the ISO/IEC 25059 AI system quality model, the metrics defined in ISO/IEC 25023 may be used, as mentioned in the guidelines for the evaluation of AI systems [50]. The ISO/IEC 25023 standard is currently under revision and contains a set of quality metrics for each (sub)characteristic, as well as recommendations for their application to software products and systems. The applicability of these metrics to AI systems is not discussed in this paper.
2. For the new characteristics and subcharacteristics specific to artificial intelligence, which were introduced for the first time in the ISO/IEC 25059 product quality model, the guidance provided in ISO/IEC 25058 may be used. This document contains recommendations for the evaluation of AI system quality; however, this document does not contain any metrics. The authors propose their own metrics, developed in accordance with the recommendations of ISO/IEC 25058.
3. For characteristics and subcharacteristics not covered by the current standards and transferred into the AI model from the updated ISO/IEC 25010, the metrics will be proposed by the authors on the basis of an analysis of the results of existing studies, current practice, and other ISO/IEC standards.

In Table 4, all the proposed metrics for the updated product quality model for the AI system are presented.

Table 4. Summary table for the product quality measures of AI systems (updated).

№	Quality Characteristic	Quality Subcharacteristic	Quality Measure Name	ID	Source
1	Functional suitability	Functional completeness	Functional coverage	FCp-1-G	[51]
2		Functional correctness	Functional correctness	FCr-1-G	[51]
3		Functional appropriateness	Functional appropriateness of usage objective	FAp-1-G	[51]
			Functional appropriateness of system	FAp-2-G	[51]
4		Functional adaptability	Average Accuracy	FAd-1-S	[53]
			Backward Transfer	FAd-2-S	
			Forward Transfer	FAd-3-S	
5	Performance efficiency	Time behaviour	Mean response time	PTb-1-G	[51]
			Response time adequacy	PTb-2-G	
			Mean turnaround time	PTb-3-G	
			Turnaround time adequacy	PTb-4-G	
			Mean throughput	PTb-5-G	
6		Resource utilization	Mean processor utilization	PRu-1-G	[51]
			Mean memory utilization	PRu-2-G	
			Mean I/O devices utilization	PRu-3-G	
			Bandwidth utilization	PRu-4-S	
7		Capacity	Transaction processing capacity	PCa-1-G	[51]
			User access capacity	PCa-2-G	
			User access increase adequacy	PCa-3-S	
8	Compatibility	Co-existence	Co-existence with other products	CCo-1-G	[51]
9		Interoperability	Data formats exchangeability	CIn-1-G	[51]
			Data exchange protocol sufficiency	CIn-2-G	
			External interface adequacy	CIn-3-S	
10	Interaction capability	Appropriateness recognisability	Description completeness	UAp-1-G	[51]
			Demonstration coverage	UAp-2-S	
			Entry point self-descriptiveness	UAp-3-S	
11		Learnability	User guidance completeness	ULe-1-G	[51]
			Entry fields defaults	ULe-2-S	
			Error message understandability	ULe-3-S	
			Self-explanatory user interface	ULe-4-S	
12		Operability	Operational consistency	UOp-1-G	[51]
			Message clarity	UOp-2-G	
			Functional customizability	UOp-3-S	
			User interface customizability	UOp-4-S	
			Monitoring capability	UOp-5-S	
			Undo capability	UOp-6-S	
			Understandable categorization of information	UOp-7-S	
			Appearance consistency	UOp-8-S	
			Input device support	UOp-9-S	

Table 4. *Cont.*

№	Quality Characteristic	Quality Subcharacteristic	Quality Measure Name	ID	Source
13	Interaction capability	User error protection	Avoidance of user operation error	UEp-1-G	[51]
			User entry error correction	UEp-2-S	
			User error recoverability	UEp-3-S	
14		User engagement	Appearance aesthetics of user interfaces	UIn-1-S	[51]
15		Inclusivity	Accessibility for users with disabilities	UAc-1-G	[51]
16		User assistance	Supported languages adequacy	UAc-2-S	[51]
17		Self-descriptiveness	Task Success Without Guidance	UDe-1-S	Proposed by the authors based on [54]
			First-time Task Completion Rate	UDe-2-S	
			Contextual Feedback Ratio	UDe-3-S	
			Action Clarifiability Rate	UDe-4-S	
18		User controllability	Control execution correctness	UCo-1-G	Proposed by the authors based on [50]
			Average control execution time	UCo-2-S	
			Control reliability	UCo-3-S	
			Control process complexity	UCo-4-S	
19		Transparency	Technical Transparency Rate	UTr-1-S	Proposed by the authors
			Interaction Transparency Rate	UTr-2-S	
			Social Transparency Disclosure	UTr-3-S	
20	Reliability	Faultlessness	Fault correction	RMa-1-G	[51]
			Mean time between failures (MTBF)	RMa-2-G	
			Failure rate	RMa-3-S	
			Test coverage	RMa-4-S	
21		Availability	System availability	RAv-1-G	[51]
			Mean down time	RAv-2-G	
22		Fault tolerance	Failure avoidance	RFt-1-G	[51]
			Redundancy of components	RFt-2-S	
			Mean fault notification time	RFt-3-S	
23		Recoverability	Mean recovery time	RRe-1-G	[51]
			Backup data completeness	RRe-2-S	
24		Robustness	Error Tolerance Rate	RRb-1-S	Proposed by the authors
			Adversarial Rate	RRb-2-S	
			Dataset Shift Tolerance Rate	RRb-3-S	
25	Security	Confidentiality	Access controllability	SCo-1-G	[51]
			Data encryption correctness	SCo-2-G	
			Strength of cryptographic algorithms	SCo-3-S	
26		Integrity	Data integrity	SIn-1-G	[51]
			Internal data corruption prevention	SIn-2-G	
			Buffer overflow prevention	SIn-3-S	
27		Nonrepudiation	Digital signature usage	SNo-1-G	[51]

Table 4. *Cont.*

№	Quality Characteristic	Quality Subcharacteristic	Quality Measure Name	ID	Source
28	Security	Accountability	User audit trail completeness	SAc-1-G	[51]
			System log retention	SAc-2-S	
29		Authenticity	Authentication mechanisms sufficiency	SAu-1-G	[51]
			Authentication rules conformity	SAu-2-S	
30		Resistance	Attack Resistance Rate	RRb-1-S	Proposed by the authors
			Model Confidentiality Protection Rate	RRb-2-S	
			Security Testing Coverage	RRb-3-S	
			Retraining Resistance Sufficiency	RRb-1-S	
31		Intervenability	Intervention success rate	RRb-2-S	[47]
			Average intervention delay time	RRb-3-S	
32	Maintainability	Modularity	Coupling of components	MMo-1-G	[51]
			Cyclomatic complexity adequacy	MMo-2-S	
33		Reusability	Reusability of assets	MRe-1-G	[51]
			Coding rules conformity	MRe-2-S	
34		Analysability	System log completeness	MAn-1-G	[51]
			Diagnosis function effectiveness	MAn-2-S	
			Diagnosis function sufficiency	MAn-3-S	
35		Modifiability	Modification efficiency	MMd-1-G	[51]
			Modification correctness	MMd-2-G	
			Modification capability	MMd-3-S	
36		Testability	Test function completeness	MTe-1-G	[51]
			Autonomous testability	MTe-2-S	
			Test restartability	MTe-3-S	
37	Flexibility	Adaptability	Hardware environmental adaptability	PAd-1-G	[51]
			System software environmental adaptability	PAd-2-G	
			Operational environment adaptability	PAd-3-S	
38		Scalability	Size Rate	PSc-1-G	Proposed by the authors
			Speed Rate	PSc-2-G	
			Complexity Rate	PSc-3-S	
39		Installability	Installation time efficiency	PIn-1-G	[51]
			Ease of installation	PIn-2-G	
40		Replaceability	Usage similarity	PRe-1-G	[51]
			Product quality equivalence	PRe-2-S	
			Functional inclusiveness	PRe-3-S	
			Data reusability/import capability	PRe-4-S	

Table 4. *Cont.*

№	Quality Characteristic	Quality Subcharacteristic	Quality Measure Name	ID	Source
41		Operational constraint	Operational Constraint Effectiveness	SaO-1-G	Proposed by the authors
			False Constraint Trigger Rate	SaO-2-S	
			Constraint Activation Time	SaO-3-S	
42		Risk identification	Risk identification rate	SaR-1-G	Proposed by the authors
	Safety		Coverage of Risk Scenarios	SaR-2-S	
43		Fail safe	Fail-Safe Rate	SaF-1-G	Proposed by the authors
			Safe Mode Controllability Rate	SaF-2-S	
			Unintended Outcome Avoidance	SaF-3-S	
44		Hazard warning	Warning Accuracy Rate	SaW-1-G	Proposed by the authors
			Timeliness of Warning	SaW-2-S	
			Warning Coverage	SaW-3-S	
45		Safe integration	Integration Safety Rate	SaI-1-G	Proposed by the authors
			Conflict Detection Rate	SaI-2-S	

We will present and describe the missing metrics for each quality characteristic of artificial intelligence systems in greater detail. In Table 4, such characteristics are marked in grey.

Functional correctness

The interpretation of the subcharacteristic "Functional correctness" in the context of AI systems has been modified. This is because, unlike traditional software, where correctness implies full compliance of the implementation with a predefined specification in AI systems, a certain level of incorrect results is permissible owing to the probabilistic nature of machine learning methods employed in generating outputs [21].

"Functional correctness" in the context of AI systems and machine learning methods refers to the degree to which the system achieves the intended outcomes with a specified level of accuracy, even if the implementation of the algorithm is not entirely "correct" in the formal, traditional sense. Lavazza and Morasca [55] emphasise that, in AI systems, the correctness of the implemented algorithm is less important than the practical applicability of the results produced.

The standard ISO/IEC TS 25058 [50] continues to recommend the use of the simple Functional correctness metric from ISO/IEC 25023:2016, which is measured as the proportion of functions that provide the correct results.

Since AI systems are designed to perform tasks—and one or several tasks may be defined for a given AI system—the quality requirements for such systems can be specified in the context of task performance evaluation. To calculate the corresponding metric, it is necessary to determine what constitutes correct task execution. This requires verifying whether the implemented requirement achieves a level of successful execution that meets or exceeds a pre-defined threshold. This is performed by calculating the ratio between correct and incorrect executions, considering both false positives and false negatives. A requirement is considered to be correctly fulfilled only if the percentage of successful executions exceeds the established threshold, thereby confirming its reliable performance under test conditions [41]. The evaluation of requirement correctness necessitates repeated execution of associated test cases. Oviedo et al. [41] recommend determining the minimum number of such executions using a statistical formula based on the assumption of a normal

distribution. Threshold values and the performance metrics used are selected by developers based on stakeholder requirements and the specific task addressed by the AI system.

The standards ISO/IEC 23053:2022 "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)" and ISO/IEC TS 4213:2022 "Information Technology—Artificial Intelligence—Assessment of Machine Learning Classification Performance" [48,56] define a set of task-specific performance metrics for machine learning models.

There also exists a separate standard for evaluating performance for Artificial Intelligence Server Systems [57].

It should be noted that, for the evaluation of "Functional correctness", various methods are used, including functional testing such as metamorphic testing, expert panels, benchmarking AI systems, etc.

Functional adaptability

Functional adaptability is the degree to which an AI system can accurately acquire information from data, or the result of previous actions, and use that information in future predictions [21]. Accordingly, the evaluation of the functional adaptability of an AI system should assess its ability to improve performance through accumulated experience—that is, by processing new input data and/or analysing previously obtained results. The guidance for quality evaluation of AI systems [50] recommends the use of performance metrics from ISO/IEC 23053:2022 [48], depending on the task of the AI system, along with functional testing methods. It also advises consideration of resource trade-offs when selecting the most appropriate machine learning model for deployment.

In existing empirical studies, various metrics have been proposed for calculating functional adaptability. For example, Oviedo et al. [41] compute the number of requirements that have improved their rate of correct results, as well as the total number of implemented requirements. The improvement in terms of correct outcomes is assessed using a functional correctness metric. If the share of correct results for a given requirement exceeds the correctness threshold defined in its specification, it is considered that the correctness of the requirement has improved.

Guo et al. [47] propose assessing this metric through two aspects: data adaptability and environmental adaptability. For the classification task they examine, data adaptability is measured using the traditional accuracy metric. To assess environmental adaptability, the authors propose a metric that is essentially a weighted sum of Precision, Recall, and a continual learning metric.

The concept of functional adaptability encompasses that of continuous learning (also referred to as continual learning or lifelong learning), although continuous learning is not a mandatory requirement. A system may be considered adaptive by simply switching between models, without necessarily performing further training. In our view, this implies that evaluation metrics used in continual learning research can be applied to measure this subcharacteristic. The three most commonly used metrics for evaluating model quality in terms of accuracy, as well as the ability to transfer knowledge both forward and backward across tasks, are Average Accuracy (ACC), Backward Transfer (BWT), and Forward Transfer (FWT) [53]. These metrics are presented in Table S1.

Self-descriptiveness

Self-descriptiveness is a property of an AI system that denotes its ability to clearly communicate its current state, reasoning, available actions, and expected outcomes, enabling users to interact with the system effectively without requiring additional guidance. It is one of the key interaction principles [54]. To assess self-descriptiveness, qualitative research methods are most commonly employed, namely expert interviews and user-based evaluation.

For example, Brand et al. [58] and Zender et al. [59] conducted an empirical study on the AI platform for automated machine learning, OMA-ML (Ontology-based Meta AutoML). Using qualitative research methods, including expert interviews and usability studies, they assessed various usability-related characteristics, including self-descriptiveness. A five-point Likert scale was used for measurement.

At the same time, there are examples of quantitative evaluation of this subcharacteristic. For instance, Vinayagasundaram and Srivatsa [60] define the self-descriptiveness of an AI system as the ratio of comment statements, defined as the proportion of comment lines to the number of executable statements. However, this metric does not reflect the interaction between the user and the AI system.

We have proposed a set of self-descriptiveness measures that conform to the principles of ISO 9241-110:2020 [54] (Table S2).

User controllability

User controllability is a property of an AI system such that a human or another external agent can intervene in its functioning in a timely manner [21]. Enhanced controllability is important when unexpected behaviour cannot be fully avoided and may lead to negative consequences [21]. User controllability is closely related to controllability, and the guidance for quality evaluation of AI systems [50] recommends using the controllability framework according to ISO/IEC TS 8200 [61].

When evaluating an AI system, it is recommended to verify compliance of its control functionalities with the established requirements, and to assess the level of controllability of the AI system by testing the specified control functionalities. According to ISO/IEC TS 8200 [61], the controllability levels of an AI system may be categorised as: completely controllable, partially controllable, sequentially controllable, loosely controllable, and not controllable. It is important to test control subprocesses such as control transfer, engagement and disengagement of control, and uncertainty handling of control transfer, as well as the actual control. For each control functionality, tests may evaluate aspects such as correctness, duration, reliability, and number of operations, depending on the specific tasks assigned to the AI system.

ISO/IEC TS 8200 [61] provides guidance on both functional and non-functional testing for controllability; however, it does not include recommendations regarding the use of metrics for the assessment of controllability.

In their study, M. Guo et al. [47] propose the following metrics for the assessment of this subcharacteristic:

1. Response time for intervention—the time delay between the receipt of a user intervention instruction by the system and its subsequent action.
2. Intervention success rate: the percentage of user-issued intervention instructions that are correctly identified and executed by the AI system, thereby achieving the desired outcomes in specific tests or usage scenarios.
3. User Intervention Survey—the level of user satisfaction when intervening with AI systems or software.

In accordance with the recommendations of ISO/IEC TS 25058 [50] we have proposed a set of metrics for the evaluation of this subcharacteristic. These metrics are intended to enable the assessment of the correctness of a control functionality, the duration of a control functionality, the reliability of a control functionality, number of operations needed by a control functionality. The developed metrics are simple and easily applicable, following the same principles as the metrics provided in ISO/IEC 25023, which are used to evaluate other subcharacteristics of Interaction Capability (formerly Usability). Moreover, they conform to the principles outlined in ISO 9241-110:2020 [54] (Table S3).

Transparency

Transparency is the degree to which adequate information about the AI system is communicated to stakeholders [21]. AI systems frequently employ black-box algorithms, which are inherently complex and opaque. Transparency in AI systems implies that all stakeholders affected by the output of an AI system should fully understand: the internal operations of the system, including how it is developed, trained, and deployed; the factors influencing its decision-making process; and the risks it entails.

AI Transparency is actively discussed within both academic and professional communities, highlighting the global absence of established transparency standards for AI and the challenges involved in disclosing information about AI systems. These challenges include concerns related to security, privacy, explainability of complex models to non-experts, and intellectual property. For instance, in the publication of the GPT-4 Technical Report, OpenAI emphasised that "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" [62].

This subcharacteristic of AI systems is closely related to AI explainability and interpretability, although each of these addresses different aspects of understanding AI. Explainability focuses on how an AI system arrives at a specific result, providing insight into the reasons and factors that influenced the outcome. Interpretability concerns understanding the internal functioning of the AI model (How does the model make decisions?), making the entire decision-making process comprehensible to stakeholders. Transparency, by contrast, encompasses a broader context: it includes information on factors related to the development and deployment of AI systems—such as training data and who has access to them—thus making the entire AI system lifecycle open to analysis and audit.

Transparency metrics should measure the extent to which a system provides complete, understandable, and accessible information about its data, algorithms, decision-making processes, and internal operations to stakeholders. These metrics should enable the assessment of how well an AI system is documented, verifiable, traceable in its decisions, and capable of explaining the models, features, and parameters it employs. Such metrics are essential to support trust in AI systems, as well as explainability, legal and ethical accountability, and to enable external oversight, informed system selection, and assessment of societal and individual impact.

Numerous theoretical and empirical studies on the evaluation of AI system transparency have been published in the academic literature. For example, Fehr et al. [63] evaluated the transparency of 14 AI-based healthcare products using a questionnaire composed of questions grouped into five sections, each corresponding to transparency requirements: intended use, algorithmic development, ethical considerations, technical validation and quality assessment, and caveats for deployment.

Using the European Commission's ALTAI list [64] as a foundation and adapting it to assess transparency requirements specific to healthcare AI applications, Nicolae et al. [65] defined eight quantitative transparency metrics for AI systems. These reflect three elements outlined in ALTAI: traceability, explainability, and open communication about the limitations of the AI system.

Researchers at Stanford University [66] developed a Foundation Model Transparency Index to evaluate ten leading developers (e.g., OpenAI, Google, Meta) across 100 indicators of transparency. These indicators were grouped into three categories: upstream resources used to build the model (e.g., data, labour, compute), model details (e.g., size, capabilities, risks), and downstream use and impact (e.g., user-facing consequences, model updates, and governance policies).

We propose the use of the high-level transparency metrics, which reflect three critical aspects of transparency: technical transparency score, interaction transparency score, and social transparency score. These are presented in Table S4.

Robustness

Robustness is the ability of an AI system to maintain its level of functional correctness under any circumstances [21], such as model errors, distorted or noisy data, external influences (e.g., overloads or failures), unpredictable situations (e.g., climate change), or the presence of human or artificial agents capable of adversarial interaction with the AI system.

There are two general approaches to ensuring the robustness of AI systems: robustness to model errors and robustness to unpredictable events. Model errors may include, for instance, incorrectly specified hyperparameters, whereas unmodelled phenomena involve unpredictable changes in operational conditions (e.g., weather changes due to climate change) [67]. To test for model errors, approaches such as robust optimisation, regularisation, risk-sensitive objective functions, and robust inference algorithms are employed. For unpredictable phenomena, techniques include expanding the model, learning a causal model, using a portfolio of models, and monitoring system performance. One of the well-known challenges in AI systems is dataset shift—a situation where inference-time data differ from training data, leading to overconfidence and poor generalisation. Studies have shown that performance can degrade by up to 40–45% under data context shifts [67].

Taking into account these approaches and the recommendations of ISO/IEC TR 25058 [50] we have developed three high-level metrics to measure: robustness to model errors, robustness to unpredictable phenomena, and robustness to dataset shift (Table S5).

For a more detailed evaluation of the robustness of neural networks, it is recommended to use the metrics and methods described in ISO/IEC TR 24029-1 [68] and ISO/IEC 24029-2 [69].

Resistance

Resistance is a new subcharacteristic of the "Security" characteristic, introduced in ISO/IEC 25010 [24], which reflects the capability of a product to sustain operations while under attack from a malicious actor. AI resistance (resistance of AI to malicious influence) may be considered as the ability of an AI system to maintain functionality and security in the presence of adversarial attacks (e.g., malicious data injection, query spoofing, and similar threats).

Given the widespread deployment of AI systems across various sectors –including critical infrastructure—these systems have become a prominent target for cybercriminals, significantly expanding the attack surface. AI systems are vulnerable to numerous security threats and exploitation of vulnerabilities that may arise throughout their entire lifecycle, from development and model training to deployment and operation. These threats include both conventional threats, common to traditional IT systems, and AI-specific threats. The emergence of AI-specific threats and vulnerabilities is attributable to the architectural and functional properties of AI systems, in particular, the probabilistic nature of many AI models and the dependency of AI system performance on training and input data, which increases the risk of indirect influence on system behaviour [70].

Both academic literature and standards actively discuss a variety of security threats and attack vectors associated with AI system usage, as well as strategies and techniques for mitigating AI security threats [71–79]. Securing AI systems requires a comprehensive approach that encompasses both the computational infrastructure supporting AI operations and the protection of models and data. This includes the model algorithms, parameters, and other architectural components of the AI system. Among the key types of attacks specific to AI systems are adversarial attacks, model theft, model inversion and extraction, data

poisoning, supply chain risks, and resource exhaustion attacks, including denial-of-service (DoS) attacks.

Researchers also emphasise the critical need to adapt existing standards and develop new standards in the field of AI security [70,80]. Although some existing standards have begun to address the issue of metrics for evaluating AI system security and testing procedures, the literature continues to highlight significant gaps in standardisation in this area. In particular, it is noted that measurement standards for assessing the security level of AI systems require further research and development in the near future [70].

In light of this, we propose several high-level, simple metrics for evaluating the resistance of AI systems (Table S6).

Intervenability

Intervenability is the degree to which an operator can intervene in the operation of an AI system in a timely manner to avoid damage or danger [21]. Intervenability is one of the crucial aspects of AI system quality, as the system must include sufficient mechanisms that enable human operators to intervene in its behaviour at any time, when necessary, in order to prevent harm or danger posed by the AI system.

Intervenability is closely related to controllability and constitutes a precondition for controllability, as defined in ISO/IEC TS 8200 [61].

Guo et al. [47] suggest two simple metrics for assessing this subcharacteristic: "Intervention success rate" and "Average intervention response time". In our view, these metrics are straightforward and effectively reflect the key aspects of intervenability (Table S7).

Scalability

Scalability is a new subcharacteristic of the "Flexability" characteristic introduced in ISO/IEC 25010 [24], which reflects the capability of a product to handle growing or shrinking workloads or to adapt its capacity to handle variability. Scalability of an AI system refers to the ability of its algorithms, data, models, and infrastructure to function effectively at the required size, speed, and complexity for a given mission [81].

Based on this definition, the main types of scalability are algorithmic scalability, data scalability, model scalability, and infrastructure scalability. These four are regarded as subtypes of technical scalability. In the academic literature, other forms are also identified, including operational scalability, inference scalability, environmental scalability, etc. [81–83]; however, we focus here on the four core types as components of technical scalability. As metrics, researchers most frequently suggest size, speed, and complexity. At the same time, Scepanski and Zillner [83] suggest the use of cost and quality as alternative indicators.

We suggest the use of three universal metrics—size, speed, and complexity—which can be applied across all components of the AI system (algorithms, data, models, and infrastructure) (Table S8).

For example, the size metrics for AI system components can include: for data—data volume, number of samples/instances, dimensions; for algorithms—lines of code (LoC) or number of modules; for models—number of model parameters, model size on disc, model depth/number of layers; for infrastructure—RAM size, number of compute nodes, or storage capacity.

When calculating the complexity rate metric, appropriate data complexity indicators may include, for example, a simple metric such as the number of features, or more advanced ones such as the feature overlap metric, class imbalance metric, for algorithms—cyclomatic complexity, for models—floating-point operations per second (FLOPS), for infrastructure—interconnectivity degree.

Safety

In scientific research, it is emphasised that safety is one of the most commonly discussed quality characteristics of AI systems [30,35,37,84]. This is due to the increasing capabilities of general-purpose AI [85], as well as the rapid integration of AI systems into safety-critical domains, including infrastructure, healthcare, finance, and national security.

AI technologies can have both positive and negative impacts on organisations that use them, as well as on society more broadly. A major source of concern lies in the probabilistic and non-deterministic nature of many AI systems, which makes formal verification of safety infeasible [84]. At the same time, the use of machine learning often results in systems that are complex and difficult to interpret, further elevating safety concerns. However, as noted in the Center for AI Safety's 2023 Impact Report [86], only 3% of technical research is dedicated to ensuring AI safety, while the vast majority is focused on developing more powerful AI systems.

In addition to the insufficient coverage of AI system requirements in current standards, and the absence of the Safety characteristic in the AI system quality model [21], significant attention is being given to AI safety in safety-related standards. These standards highlight the lack of a unified approach to defining AI-specific requirements for safety-related AI-based systems [84,87]. In most domains, existing standards designed for traditional IT systems continue to be used. The newly published standard IEC/ISO TR 5469:2024 "Artificial Intelligence—Functional Safety and AI Systems" [88], acknowledges the potential for AI technologies to be used in safety-critical systems, but emphasises that such use requires special attention and adherence to specific conditions. Meanwhile, other standards (e.g., IEC 61508 [89], ISO 26262 [90], ISO 21448 [91]) explicitly state that non-deterministic AI systems cannot be used in systems with high safety requirements.

According to ISO/IEC 25010:2023 [24], safety is a capability of a product under defined conditions to avoid a state in which human life, health, property, or the environment is endangered. Due to their nature, AI systems may unintentionally cause harm, for example, through misinterpretation of outputs. Therefore, in the context of this work, we define the safety of an AI system as the degree to which the system, under defined and foreseeable operating conditions, does not lead to a state in which human life, health, property, or the environment is harmed (either directly or indirectly) as a result of its decision-making and behaviour.

AI Safety is closely related to AI Security, but there is an important distinction:

1. AI Safety focuses on preventing internal risks associated with AI behaviour and decision-making, in order to reduce the likelihood of adverse impacts on business or society.
2. AI Security, in contrast, addresses protection against external threats and malicious interference.

A particular concern is the AI alignment problem, which refers to the challenge of ensuring that an AI system's goals and actions remain aligned with human intentions—especially as systems grow more autonomous and capable. If alignment is compromised, the AI may pursue goals that appear logical to it, but are ineffective or even harmful. This is especially dangerous because even minor misalignments can result in outcomes that deviate significantly from user expectations or societal norms.

Key AI safety risks include [85]:

1. Malicious use (e.g., the spread of fake content, manipulation of public opinion, cybercrime, biological or chemical threats).
2. Risks from malfunctions (e.g., reliability issues, algorithmic bias, loss of control).

3. Systemic risks (e.g., labour market disruption, global inequalities in AI R&D, market concentration and single points of failure, threats to the environment, privacy, and intellectual property).
4. Additionally, the use of open-weight general-purpose AI models may exacerbate many of these risks.

As with the evaluation of other non-functional characteristics and subcharacteristics of AI systems, the scientific literature notes an insufficient elaboration of methods for measuring aspects related to AI Safety [33,37,84]. Researchers stress the difficulty of evaluating safety, since only indirect indicators are typically measurable, and these reflect potential risk rather than safety itself. Safety evaluation may include both quantitative metrics (e.g., the frequency with which models generate particular types of information) and qualitative indicators (e.g., describing how users interact with the system and how they interpret or apply its outputs) [92]. At the same time, it is acknowledged that in safety-critical situations (e.g., autonomous driving, e-health), decisions should not rely solely on AI systems and must also involve human judgment [37,93]. For example, in medical research, decisions made by doctors with AI assistance are statistically compared with decisions made without AI. Such comparative studies serve as the basis for evaluating safety and making implementation decisions.

We now examine the main safety subcharacteristics proposed in ISO/IEC 25010 [24], and suggest metrics for their evaluation.

Operational constraint

Operational constraint is a capability of a product to constrain its operation to within safe parameters or states when encountering operational hazard, i.e., hazardous situation, a circumstance in which people, property, or the environment are exposed to an unacceptable risk during operation [24]. In the context of AI systems, operational constraint refers to the system's ability to restrict its behaviour within safe parameters or states when encountering operational hazards, including incorrect input data, unstable environmental conditions, user errors, or unforeseen influences. This capability ensures the prevention of actions that could lead to unacceptable harm to humans, property, or the environment during the system's operation.

To evaluate this subcharacteristic, we propose the use of the following metrics (Table S9).

Risk identification

Risk identification is the capability of a product to identify a course of events or operations that can expose life, property, or environment to unacceptable risk [24]. In the context of AI systems, risk identification will be understood as the capability of the AI system to detect potential sequences of events, decisions, or system operations that may lead to unacceptable harm to humans, property, or the environment.

To evaluate this subcharacteristic, we propose the use of the following metrics (Table S10).

Fail safe

According to ISO/IEC 25010, fail safe is a capability of a product to automatically place itself in a safe operating mode, or to revert to a safe condition in the event of a failure [24]. In the context of AI systems, an AI fail-safe will be understood as the capability of an AI system to automatically transition to a safe operational state or revert to a safe mode in the event of a failure, malfunction, or loss of control, thereby minimising harm to humans, property, or the environment.

This means that in machine learning, when predictions cannot be made with confidence, the model indicates that it is unable to provide a reliable output and refrains from doing so, entering a safe mode [94]. When a critical AI system fails, it should fail safely,

offering users alternative options [95]. Therefore, this subcharacteristic is closely related to controllability: users must be able to understand the system's decisions and behaviour in order to interact with it effectively. Typically, in such cases, a human operator intervenes to provide a manual prediction.

At the same time, the difficulty of defining the states in which the system should safely fail is emphasised. For example, in medical diagnostics, AI systems may continue to produce confident but incorrect predictions when encountering inputs that fall outside the training distribution or when the available input data are insufficient [93]. This can lead to situations where doctors, placing trust in the system, do not question its outputs and prematurely cease to consider alternative diagnoses, thereby increasing the risk of diagnostic error.

To evaluate this subcharacteristic, we suggest the use of the following metrics (Table S11).

Hazard warning

According to ISO/IEC 25010, hazard warning is a capability of a product to provide warnings of unacceptable risks to operations or internal controls so that they can react in sufficient time to sustain safe operations [24]. In the context of AI systems, AI hazard warning will be understood as the capability of an AI system to timely detect and communicate unacceptable risks to its operations, users, or the environment, enabling appropriate interventions before harmful consequences occur. For example, an AI-powered autonomous vehicle alerts the driver about potential collision risks due to uncertain object detection, allowing manual override.

To evaluate this subcharacteristic, we suggest the use of the following metrics (Table S12).

Safe integration

According to ISO/IEC 25010, safe integration is a capability of a product to maintain safety during and after integration with one or more components [24]. In the context of AI systems, safe integration will be understood as is the capability of an AI system to maintain its safe operation during and after integration with external components, ensuring that such integration does not introduce new unacceptable risks or degrade system safety.

To evaluate this subcharacteristic, we suggest the use of the following metrics (Table S13).

4. AI System Quality in Use Model Measures

As with the product quality model, for the quality characteristics and subcharacteristics of the quality model in use that have been transferred unchanged from the ISO/IEC 25010 quality model in use to the ISO/IEC 25059 AI system quality model, the metrics defined in ISO/IEC 25022 [52] may be used. With regard to the new characteristics and subcharacteristics specific to AI and introduced for the first time in the ISO/IEC 25059 quality in use model, ISO/IEC 25058 will be used as guidance; and for characteristics and subcharacteristics not covered by the current standards and transferred into the AI model from ISO/IEC 25010, the metrics will be developed by the authors on the basis of an analysis of current research and existing practices.

In Table 5, a set of metrics for the updated product quality model for the AI system is presented.

Table 5. Summary table for the quality model in use measures for the system with AI (updated).

No.	Quality Characteristic	Quality Subcharacteristic	Quality Measure Name	ID	Source
1	**Beneficialness**	Usability:			
		• Effectiveness	Tasks completed	Ef-1-G	[52]
			Objectives achieved	Ef-2-S	
			Errors in a task	Ef-3-G	
			Tasks with errors	Ef-4-G	
			Task error intensity	Ef-5-G	
		• Efficiency	Task time	Ey-1-G	[52]
			Time efficiency	Ey-2-S	
			Cost-effectiveness	Ey-3-S	
			Productive time ratio	Ey-4-S	
			Unnecessary actions	Ey-5-S	
			Consequences of fatigue	Ey-6-S	
		• Usefulness	Satisfaction with features	SUs-2-G	[52]
			Discretionary usage	SUs-3-G	
			Feature utilisation	SUs-4-G	
			Proportion of users complaining	SUs-5-G	
			Proportion of user complaints about a particular feature	SUs-6-G	
		• Trust	User trust	STr-1-G	[52]
		• Pleasure	User pleasure	SPL-1-G	[52]
		• Comfort	Physical comfort	SCo-1-G	[52]
		• Transparency	Technical Transparency Rate	UTr-1-S	Proposed by the authors
			Interaction Transparency Rate	UTr-2-S	
			Social Transparency Disclosure	UTr-3-S	
2		Accessibility	Inclusive Functionality Rate	BAc-1-G	Proposed by the authors
			Assistive Tech Compatibility Score	BAc-2-S	
			User Diversity Success Rate	BAc-3-S	
			Cross-language Accessibility Index	BAc-4-S	
			Cognitive Load Differential	BAc-5-S	
3		Suitability	Requirement Satisfaction Rate	BSu-1-G	Proposed by the authors
			Goal Match Rate	BSu-2-S	
			Acceptability Rating	BSu-3-S	
			Contextual Misfit Rate	BSu-4-S	
			Override Frequency	BSu-5-S	
4	**Freedom from risk**	Freedom from economic risk	Return on investment (ROI)	REc-1-G	[52]
			Time to achieve return on investment	REc-2-G	
			Business performance	REc-3-G	
			Benefits of IT Investment	REc-4-G	
			Service to customers	REc-5-S	
			Website visitors converted to customers	REc-6-S	
			Revenue from each customer	REc-7-S	
			Errors with economic consequences	REc-8-G	

Table 5. *Cont.*

No.	Quality Characteristic	Quality Subcharacteristic	Quality Measure Name	ID	Source
5	Freedom from risk	Freedom from health risk	User health reporting frequency	RHe-1-G	[52]
			User health and safety impact	RHe-2-G	
6		Freedom from human life risk	Safety of people affected by use of the system	RHe-3-G	[52]
7		Freedom from environmental and societal risk	Environmental impact	REn-1-G	[52]
			Societal: Societal Risk Identification Coverage	Rs-1-G	Proposed by the authors
			Societal: Mitigation Effectiveness Rate	Rs-2-S	
			Ethical: Ethical Review Integration Index	REt-1-G	
			Ethical: Residual Risk Acceptability Rate	REt-2-S	
			Ethical: Stakeholder Ethics Involvement Ratio	REt-3-S	
8	Acceptability	Experience	Skill Gain Rate	AEe-1-G	Proposed by the authors
			Pattern Recognition Assistance Score	AEe-2-S	
			Knowledge Retention Support Index	AEe-3-S	
			Experience Continuity Rate	AEe-4-S	
9		Trustworthiness	Expectation Confidence Index	ATw-1-G	Proposed by the authors
			Transparency Support Rate	ATw-2-S	
			Security Incident Avoidance Rate	ATw-3-S	
			Accountability Traceability Index	ATw-4-S	
10		Compliance	Regulatory Compliance Coverage	ACo-1-G	Proposed by the authors
			Compliance Audit Success Rate	ACo-2-S	
			Policy Traceability Index	ACo-3-S	
			User Confidence in Legal Compliance	ACo-4-S	
11	Context coverage	Context completeness	Context completeness	CCm-1-G	[52]
12		Flexibility	Flexible context of use	CFl-1-S	[52]
			Product flexibility	CFl-2-S	
			Proficiency independence	CFl-3-S	

Beneficialness

Beneficialness is a new characteristic that includes two newly introduced subcharacteristics, "Accessibility" and "Suitability", as well as a consolidated subcharacteristic, "Usability". "Usability" integrates "Effectiveness", "Efficiency", and "Satisfaction"—attributes that were previously defined as separate characteristics in the earlier version of the standard and are now considered sub-subcharacteristics. It also integrates "Trust", "Pleasure", "Comfort", and "Transparency" as sub-subcharacteristics.

Transparency

According to ISO/IEC TS 25058 [50] the metrics for this sub-subcharacteristic are analogous to those of the product quality model (Table S4).

Accessibility

Accessibility is the extent to which AI-based systems, services, or tools can be used by people with the widest range of user needs, characteristics, and capabilities to achieve identified goals in identified contexts of use [49]. To evaluate this subcharacteristic, we propose the use of the following metrics (Table S14).

Suitability

We adapt the definition provided for this subcharacteristic in the ISO/IEC 25019 [49] standard to the context of AI systems. Suitability is the extent to which the behaviours or outcomes, or both, of an AI system meet specified quality requirements in actual use by users or stakeholders. To evaluate this subcharacteristic, we propose the use of the following metrics (Table S15).

Freedom from environmental and societal risk

We adapt the definition provided for this subcharacteristic in the ISO/IEC 25019 [49] standard to the context of AI systems. Freedom from environmental and societal risk is the extent to which an AI system or service mitigates or avoids potential risks to the environment and society at large, including impacts on public safety, social cohesion, cultural norms, and ecological systems, in the intended contexts of use. To evaluate this subcharacteristic, we propose the use of the following metrics (Table S16).

Acceptability

Acceptability is the new characteristic of the updated quality in use model for an AI system. We adapt the definition provided for this characteristic in the ISO/IEC 25019 [49] standard to the context of AI systems. Acceptability is the extent to which human users or stakeholders respond favourably to the adoption, deployment, or integration of an AI system or service, based on perceived usefulness, trustworthiness, ethical alignment, and ease of configuration in real-world contexts of use.

Experience

We adapt the definition provided for this subcharacteristic in the ISO/IEC 25019 [49] standard to the context of AI systems. Experience is the extent to which users or stakeholders accumulate knowledge, insights, or professional skills over time through interaction with an AI system, especially by engaging in tasks involving decision-making, pattern recognition, analysis, or simulation within their domain of expertise. To evaluate this subcharacteristic, we propose the use of the following metrics (Table S17).

Trustworthiness

We adapt the definition provided for this subcharacteristic in the ISO/IEC 25019 [49] standard to the context of AI systems. Trustworthiness is the extent to which users or stakeholders have justified confidence that an AI system behaves as expected and satisfies its intended purpose in a verifiable, transparent, and ethically sound manner, across varying contexts of use. To evaluate this subcharacteristic, we propose the use of the following metrics (Table S18).

Compliance

We adapt the definition provided for this subcharacteristic in the ISO/IEC 25019 [49] standard to the context of AI systems. Compliance is the extent to which users or stakeholders have confidence that an AI system, in actual use, meets applicable legal, regulatory, ethical, and organisational requirements, including those related to data protection, fairness, transparency, and accountability. To evaluate this subcharacteristic, we propose the use of the following metrics (Table S19).

5. Discussion

The widespread deployment of artificial intelligence systems—both in safety-critical domains and in everyday contexts—has sharpened the problem of the insufficiency of traditional quality models to account for the specific risks and behavioural features of AI

systems. This study was aimed at examining various aspects of AI system quality and the particularities of their measurement. The main contributions of this study are as follows:

- An updated technology-oriented quality model for AI systems is proposed, taking into account the specific features of AI systems (ISO/IEC 25059:2023) and the latest updates to the traditional quality models from the 25010 standards (ISO/IEC 25010:2023, ISO/IEC 25019:2023). The quality model has been considered from two perspectives: product quality and the quality in use model.
- A metric-based approach to the assessment of AI system quality is proposed on the basis of the quality model. This approach envisages the use of standardised metrics for (sub)characteristics of AI systems transferred from the traditional product quality and quality in use models, and also proposes metrics for new (sub)characteristics included in the updated technology-oriented quality model for AI systems.

5.1. Harmonisation and the Timely Updating and Development of International Standards in the Field of AI

International organisations such as ISO/IEC continue to publish and update standards concerning the quality of AI systems. As this field is developing rapidly, the updating of standards may fail to keep pace with the evolution of the domain; at the same time, standards may not be updated in a timely manner, which provokes mismatches and divergences among interrelated standards. ISO/IEC reviews published standards every five years after release; however, as we see from the example of ISO/IEC 25023:2016- "Measurement of Software Product Quality-" [51] and ISO/IEC 25022:2016 "Measurement of Quality in use" [52], which were last reviewed and confirmed in 2022, the wait for their subsequent revision creates inconsistencies with the new versions of the related standards ISO/IEC 25010:2023 "Product Quality Model" [24] and ISO/IEC 25019:2023 "Quality in use Model" [49]. Such delays lead to inconsistency at the level of quality characteristics and subcharacteristics [96,97], which complicates the process of assessing the quality of ICT and, in particular, AI systems.

The harmonisation of existing standards and the terminology used therein [38,98] remains a significant problem and requires further work on the part of international organisations.

Scholars and practitioners emphasise that, at this stage, increased attention to standardisation is required for AI Security and AI Safety [70,80,85]. The need is underscored for a coherent roadmap for the future development of international standards in AI Security [70] and for the development of concrete technical recommendations on applying existing standards related to software cybersecurity to AI [80]. It is recommended to update existing standards rather than create new ones, and to foster closer collaboration between cybersecurity technical committees and AI technical committees [70,80].

5.2. Limitations and Challenges in Applying Standardised Metrics to AI Systems Quality Assessment

In this paper, metrics are proposed for new (sub)characteristics for which corresponding metrics are currently absent from existing standards. These metrics are developed in accordance with the principles and measurement logic applied in the ISO/IEC 25023 and ISO/IEC 25022 standards and are general in nature, rather than being tailored to a specific product or application domain of AI systems. Considering the diversity of AI technologies and the contexts in which they are used, it should be noted that different AI technologies may require different metrics to assess the same quality characteristics. Therefore, the standards provide for the possibility of selecting (sub)characteristics of AI system quality depending on the domain and the objectives of AI system development, and for defining new metrics on the part of the user, even if these metrics have not been defined in the existing standards.

In this context, the proposed standardised metrics should be regarded as a starting point for scientific discussion and further research. Some of the metrics may prove to be applicable in practice, while others may be reviewed, adopted, or excluded in the future due to their practical usefulness, as has occurred during the evolution of standards and technical reports (for example, in the transition from ISO/IEC/TR 9126-2:2003 and ISO/IEC/TR 9126-3:2003 to ISO/IEC 25023:2016) [51].

This paper has not considered the adaptation of existing metrics from the ISO/IEC 25023 and ISO/IEC 25022 standards for the assessment of (sub)characteristics of AI systems, except for the metric functional correctness, whose interpretation was modified for AI systems. In the guidance for the quality evaluation of AI systems [50] recommendations for measurement do not exist for all metrics transferred from ISO/IEC 25023 and ISO/IEC 25022. At the same time, most researchers [27,29] note that existing metrics do not fully reflect the specific features of AI systems and should be adapted. Therefore, further work is required in this direction.

It is noted that many (sub)characteristics are difficult to measure quantitatively, as is the case with traditional software [37]. This difficulty is due both to the abstract nature of certain (sub)characteristics (e.g., Trustworthiness, Freedom from risk) and to the absence of established metrics applicable to systems with probabilistic behaviour and learnable components. At the same time, quality (sub)characteristics may overlap, which can complicate their measurement. For example, transparency and accountability partially overlap, since the measurability of accountability largely depends on the availability of transparent information about the system [44]. The implementation of individual metrics may be resource-intensive; therefore, organisations should assess the costs of evaluation and the potential benefits derived from it [36]. Additionally, unlike traditional software, where quality can be directly linked to the correctness of function execution, AI systems exhibit stochastic and dynamic behaviour that is sensitive to context, input data, and the operational environment. Given this, and the continuous development of the field and the expansion of AI system application areas—for example, the active use of large language models—it may be necessary in the future to refine quality (sub)characteristics for specific AI systems and, accordingly, the metrics for their measurement [38].

An open question also concerns the formation of a metric nomenclature and the determination of priorities among metrics and the quality characteristics they measure. Clearly, the set of metrics may be defined for a specific project (a project-oriented approach), a particular problem (a problem-oriented approach), or a process (a process-oriented approach) [96,99]. Along this path, metrics may be combined in various ways. It should also be taken into account during evaluation that quality characteristics may compete with each other or even be mutually exclusive [100]. In previously published works [96,99], aggregation mechanisms, such as additive aggregation, are proposed for prioritising evaluated metrics, where prioritisation is achieved through weighting coefficients. As a means of visualising the obtained results, the same authors propose the use of radar (Kiviat) diagrams [96,99]. The methodology for forming a set of metrics and establishing their priorities remains a relevant topic for future research.

At the same time, the assessment of AI system quality is inevitably associated with trade-offs between different quality indicators due to their close interdependence. In such cases, improving one quality indicator may lead to the deterioration of another. Certain quality characteristics of AI systems cannot be optimised simultaneously, which requires an explicit prioritisation of quality attributes depending on the context of use. It is evident that, for AI systems, the most common trade-offs occur between transparency and security, usability and security, as well as robustness and functional correctness. This topic should be addressed in separate and more in-depth studies.

5.3. Quality Assessment of AI Systems vs. AI Models

This study considers an approach to the metric-based evaluation of the quality of AI systems. As noted in the literature, there is a substantial difference between approaches to assessing the quality of individual components of AI systems and of the system as a whole. In particular, the choice of appropriate metrics depends on which component is being evaluated—the model, the data, the environment, the system, or the infrastructure [31]. At the same time, a methodological gap is emphasised between the goals and the means of measuring the quality of AI systems. For example, Habibullah et al. [37] note that the participants in their interviews most often considered non-functional requirements definition over the whole system, but measured them predominantly at the model level. Using the example of Talia, an AI system used for predicting treatment outcomes for patients receiving human-supported, internet-delivered cognitive behavioural treatment for symptoms of depression and anxiety, Microsoft specialists [101] demonstrated that the evaluation of AI models is not sufficient to determine the quality of an AI system for real-life application. It is further noted that the difference between model-level and system-level evaluation lies not only in scale, but also in the nature of the risks and the technical debt that arise [102], and therefore requires different approaches, tools, and metrics.

This study did not consider the problem of data quality; however, this area requires further research and the development of recommendations beyond those proposed in existing standards, given the diversity of AI system types [38,103].

5.4. Context of the AI System

As already noted, AI systems are applied across a wide spectrum of domains—from high-risk, safety-critical ones (e.g., autonomous driving, healthcare, defence) to everyday, user-oriented contexts (e.g., recommender systems, chatbots, voice assistants). The applicability of particular AI system quality (sub)characteristics, and, accordingly, the feasibility of evaluating them using formalised metrics, is largely determined by the operational context of the specific system [31,37,104]. At the same time, the measurement of one quality (sub)characteristic may depend on another (sub)characteristic defined for other elements of the system [37].

This diversity of domain requirements implies the need for a context-dependent approach to quality assessment and the selection of appropriate metrics, as well as the support of flexible evaluation methods that allow measurements to be adapted to the specific features of particular applications. Moreover, standard quality models based on universal characteristics do not always reflect the priorities and risks inherent to a given type of AI system. Therefore, one avenue for further research should be the development of customisable or modular quality models, adaptable to the specifics of use [33,98].

5.5. Development of Assessment Tools and Testing Methods for AI Systems

In this paper, the testing methods and tools for AI systems were not considered directly; however, it should be noted that approaches used in traditional software development—such as static and dynamic code analysis, and unit, regression, and integration testing—are partially applicable to AI systems. Owing to the specific nature of AI, especially in the field of machine learning, these tools often prove insufficiently effective [67,84]. Existing tools scale poorly and, in a number of cases, have no direct analogues, which necessitates the creation of new methods and approaches. Moreover, current practices seldom cover the entire life cycle of an AI system and focus predominantly on the model level, overlooking aspects related to the quality of data, infrastructure, and interaction with the environment.

Accordingly, one of the key directions for future research in ensuring the quality of AI systems should be the development of testing and assessment tools adapted to the particularities of AI systems—capable of accounting not only for the model but also for the context of its application, the data, external risks, and the behaviour of the system under real-world conditions.

6. Conclusions and Future Work

This research opens a discussion and proposes a concept for the metric-based measurement of AI system quality based on international standards. This study is among the first to attempt to harmonise the quality and quality in use models for AI, as well as to synthesise the metrics employed and the recommendations for their use in assessing the quality of AI systems. The proposed AI system quality indicators and their measurement methods were developed on the basis of guidance documents, international standards, and contemporary research in the field of AI systems. We proposed metrics for measuring new quality subcharacteristics for the product quality model and the quality in use model, which have been incorporated into the corresponding updated quality model, and for which standardised metrics have not yet been developed. The metric-based approach developed herein enables researchers and practitioners to address the gap between existing standardised quality models (ISO/IEC 25010:2023, ISO/IEC 25019:2023, and ISO/IEC 25059:2023) and the standardised approach to the measurement of quality characteristics (ISO/IEC 25023:2016, ISO/IEC 25022:2016). The proposed set of potential metrics may be used by practitioners and researchers when evaluating the quality of AI systems with the aim of improving their quality. This will support the development and assessment of AI systems that comply with established quality standards. The improvement of assessment methods and their international standardisation will contribute to raising the quality of AI systems as they are integrated into various spheres of activity and to reducing their negative impacts associated with insufficient quality. Naturally, the proposed concept has a number of limitations and unresolved issues, the addressing of which is possible through more applied and empirical experimentation. These limitations are discussed in the Discussion Section of the paper. In future work, we plan to conduct evaluations of real-world AI systems. This will make it possible to empirically validate the applicability of the proposed approach, further refine it, and identify its practical limitations. In this process, one of the main challenges in conducting real-world experiments is expected to be the potential lack of reliable and sufficiently detailed information about AI systems, which is required to calculate the metrics proposed in this paper. In cases where such information is unavailable or incomplete, the evaluation may be limited to individual quality characteristics or selected groups of characteristics.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics15030691/s1, Table S1. Functional adaptability measures; Table S2. Self- descriptiveness measures; Table S3. User controllability measures; Table S4. Transparency measures; Table S5. Robustness measures; Table S6. Resistance measures; Table S7. Intervenability measures [47]; Table S8. Scalability measures; Table S9. Operational constraint measures; Table S10. Risk identification measures; Table S11. Fail safe measures; Table S12. Hazard warning measures; Table S13. Safe integration measures; Table S14. Accessibility measures; Table S15. Suitability measures; Table S16. Societal and ethical risk measures; Table S17. Experience measures; Table S18. Trustworthiness measures; Table S19. Compliance measures.

Author Contributions: Conceptualization, O.G. and D.G.; Methodology, O.G. and D.G.; Formal Analysis, O.G. and D.G.; Investigation, O.G., D.G., A.G. and O.T.; Resources, O.G. and D.G.; Writing—Original Draft Preparation, O.G. and D.G.; Writing—Review and Editing, O.G., D.G.,

A.R., A.G. and O.T.; Visualization, O.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: The authors would like to acknowledge partial support from the British Academy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Turing, A. *Turing 1936 on Cumputable Numbers*; Internet Archive: San Francisco, CA, USA, 1936. Available online: http://archive.org/details/Turing1936OnCumputableNumbers (accessed on 1 September 2025).
2. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *59*, 433–460. [CrossRef]
3. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955. *AI Mag.* **2006**, *27*, 12–14. [CrossRef]
4. Wiener, N. *Cybernetics or Control and Communication in the Animal and the Machine*, 2nd ed.; MIT Press: Cambridge, MA, USA, 1961.
5. Tu, X.; He, Z.; Huang, Y.; Zhang, Z.-H.; Yang, M.; Zhao, J. An overview of large AI models and their applications. *Vis. Intell.* **2024**, *2*, 34. [CrossRef]
6. Musa, J.D. *Software Reliability Engineering: More Reliable Software Faster and Cheaper*, 2nd ed.; AuthorHouse: Bloomington, IN, USA, 2004.
7. Pressman, R.S.; Maxim, B.R. *Software Engineering: A Practitioner's Approach*; McGraw Hill: New York, NY, USA, 2015.
8. Sommerville, I. *Software Engineering*; Pearson: Boston, MA, USA; Munich, Germany, 2016.
9. Smith, H. Algorithmic bias: Should students pay the price? *AI Soc.* **2020**, *35*, 1077–1078. [CrossRef]
10. Heaton, D.; Nichele, E.; Clos, J.; Fischer, J.E. "The algorithm will screw you": Blame, social actors and the 2020 A Level results algorithm on Twitter. *PLoS ONE* **2023**, *18*, e0288662. [CrossRef]
11. Engel, C.; Linhardt, L.; Schubert, M. Code is law: How COMPAS affects the way the judiciary handles the risk of recidivism. *Artif. Intell. Law* **2025**, *33*, 383–404. [CrossRef]
12. Strickland, E. *How IBM Watson Overpromised and Underdelivered on AI Health Care—IEEE Spectrum*; Spectrum IEEE: New York, NY, USA, 2019. Available online: https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care (accessed on 1 September 2025).
13. Troncoso, I.; Runshan Fu, N.M.; Proserpio, D. *Algorithm Failures and Consumers' Response: Evidence from Zillow*; Working Paper; Harvard Business School: Boston, MA, USA, 2023. Available online: https://www.hbs.edu/faculty/Pages/item.aspx?num=64507 (accessed on 1 September 2025).
14. Alvi, M.; Zisserman, A.; Nellaker, C. Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. *arXiv* **2018**, arXiv:1809.02169. [CrossRef]
15. Mills, K.G. *Gender Bias Complaints Against Apple Card Signal a Dark Side to Fintech | Working Knowledge*; Harvard Business School: Boston, MA, USA, 2019. Available online: https://www.library.hbs.edu/working-knowledge/gender-bias-complaints-against-apple-card-signal-a-dark-side-to-fintech (accessed on 1 September 2025).
16. Brand, J.L.M. Air Canada's chatbot illustrates persistent agency and responsibility gap problems for AI. *AI Soc.* **2025**, *40*, 3361–3363. [CrossRef]
17. Ryan, W.A.; Garrett, A.; Sears, B. Practical Lessons from the Attorney AI Missteps in Mata v. Avianca. Available online: https://www.acc.com/resource-library/practical-lessons-attorney-ai-missteps-mata-v-avianca (accessed on 22 September 2025).
18. Harris, M. *NTSB Investigation into Deadly Uber Self-Driving Car Crash Reveals Lax Attitude Toward Safety*; IEEE Spectrum: New York, NY, USA, 2019. Available online: https://spectrum.ieee.org/ntsb-investigation-into-deadly-uber-selfdriving-car-crash-reveals-lax-attitude-toward-safety (accessed on 1 September 2025).
19. Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, 23–24 February 2018; pp. 77–91. Available online: https://proceedings.mlr.press/v81/buolamwini18a.html (accessed on 2 September 2025).
20. AI Incident Database. Available online: https://incidentdatabase.ai/ (accessed on 1 September 2025).
21. *ISO/IEC 25059:2023*; Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Quality model for AI Systems. International Organization for Standardization: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/80655.html (accessed on 2 September 2025).
22. Ali, M.A.; Yap, N.K.; Ghani, A.A.A.; Zulzalil, H.; Admodisastro, N.I.; Najafabadi, A.A. A Systematic Mapping of Quality Models for AI Systems, Software and Components. *Appl. Sci.* **2022**, *12*, 8700. [CrossRef]

23. Gezici, B.; Tarhan, A.K. Systematic literature review on software quality for AI-based software. *Empir. Softw. Eng.* **2022**, *27*, 66. [CrossRef]
24. *ISO/IEC 25010:2023*; Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Product Quality Model. International Organization for Standardization: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/78176.html (accessed on 2 September 2025).
25. *ISO/IEC 25010:2011*; Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—System and Software Quality Models. International Organization for Standardization: Geneva, Switzerland, 2011.
26. Heck, P. *A Quality Model for Trustworthy AI Systems*; Fontys: Eindhoven, The Netherlands, 2022. Available online: https://fontysblogt.nl/a-quality-model-for-trustworthy-ai-systems/ (accessed on 7 September 2025).
27. Kuwajima, H.; Ishikawa, F. Adapting SQuaRE for Quality Assessment of Artificial Intelligence Systems. In Proceedings of the 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Berlin, Germany, 27–30 October 2019. [CrossRef]
28. Ramos, T.; Dean, A.; McGregor, D. AI-Augmented Software Engineering: Revolutionizing or Challenging Software Quality and Testing? *J. Softw. Evol. Process* **2025**, *37*, e2741. [CrossRef]
29. Smith, A.L.; Clifford, R. Quality Characteristics of Artificially Intelligent Systems. In Proceedings of the IWESQ APSEC, Singapore, 1–4 December 2020; pp. 1–6.
30. Indykov, V.; Strüber, D.; Wohlrab, R. Architectural tactics to achieve quality attributes of machine-learning-enabled systems: A systematic literature review. *J. Syst. Softw.* **2025**, *223*, 112373. [CrossRef]
31. Siebert, J.; Joeckel, L.; Heidrich, J.; Trendowicz, A.; Nakamichi, K.; Ohashi, K.; Namba, I.; Yamamoto, R.; Aoyama, M. Construction of a quality model for machine learning systems. *Softw. Qual. J.* **2022**, *30*, 307–335. [CrossRef]
32. Nakamichi, K.; Ohashi, K.; Namba, I.; Yamamoto, R.; Aoyama, M.; Joeckel, L.; Siebert, J.; Heidrich, J. Requirements-Driven Method to Determine Quality Characteristics and Measurements for Machine Learning Software and Its Evaluation. In Proceedings of the 2020 IEEE 28th International Requirements Engineering Conference (RE), Zurich, Switzerland, 31 August–4 September 2020; pp. 260–270. [CrossRef]
33. Kelly, J.; Zafar, S.A.; Heidemann, L.; Zacchi, J.-V.; Espinoza, D.; Mata, N. Navigating the EU AI Act: A Methodological Approach to Compliance for Safety-critical Products. In Proceedings of the Conference on Artificial Intelligence 2024, Singapore, 25–27 June 2024. [CrossRef]
34. *ISO/PAS 8800:2024*; Road Vehicles—Safety and Artificial Intelligence. International Organization for Standardization: Geneva, Switzerland, 2024. Available online: https://www.iso.org/standard/83303.html (accessed on 22 September 2025).
35. Martínez-Fernández, S.; Bogner, J.; Franch, X.; Oriol, M.; Siebert, J.; Trendowicz, A.; Vollmer, A.M.; Wagner, S. Software Engineering for AI-Based Systems: A Survey. *ACM Trans. Softw. Eng. Methodol.* **2022**, *31*, 1–59. [CrossRef]
36. Yu, L.; Alégroth, E.; Chatzipetrou, P.; Gorschek, T. Measuring the quality of generative AI systems: Mapping metrics to quality characteristics—Snowballing literature review. *Inf. Softw. Technol.* **2025**, *186*, 107802. [CrossRef]
37. Habibullah, K.M.; Gay, G.; Horkoff, J. Non-functional requirements for machine learning: Understanding current use and challenges among practitioners. *Requir. Eng.* **2023**, *28*, 283–316. [CrossRef]
38. Oviedo, J.; Rodriguez, M.; Trenta, A.; Cannas, D.; Natale, D.; Piattini, M. ISO/IEC quality standards for AI engineering. *Comput. Sci. Rev.* **2024**, *54*, 100681. [CrossRef]
39. Kharchenko, V.; Fesenko, H.; Illiashenko, O. Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. *Sensors* **2022**, *22*, 4865. [CrossRef]
40. Gu, K.; Liu, H.; Liu, Y.; Qiao, J.; Zhai, G.; Zhang, W. Perceptual Information Fidelity for Quality Estimation of Industrial Images. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 477–491. [CrossRef]
41. Oviedo, J.; Rodríguez, M.; Piattini, M. An Environment for the Assessment of the Functional Suitability of AI Systems. In *Quality of Information and Communications Technology*; Bertolino, A., Pascoal Faria, J., Lago, P., Semini, L., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 21–34. [CrossRef]
42. Zhang, H.; Xu, Y.; Luo, R.; Mao, Y. Fast GNSS acquisition algorithm based on SFFT with high noise immunity. *China Commun.* **2023**, *20*, 70–83. [CrossRef]
43. Nastoska, A.; Jancheska, B.; Rizinski, M.; Trajanov, D. Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries. *Electronics* **2025**, *14*, 2717. [CrossRef]
44. Kemmerzell, N.; Schreiner, A.; Khalid, H.; Schalk, M.; Bordoli, L. Towards a Better Understanding of Evaluating Trustworthiness in AI Systems. *ACM Comput. Surv.* **2025**, *57*, 1–38. [CrossRef]
45. Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; López de Prado, M.; Herrera-Viedma, E.; Herrera, F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf. Fusion* **2023**, *99*, 101896. [CrossRef]
46. Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* **2023**, *55*, 1–46. [CrossRef]

47. Guo, M.; Guo, D.; Li, M. Metrics and Testing Methods for Artificial Intelligence Software Quality Models and Their Application Examples. In Proceedings of the 2024 11th International Conference on Dependable Systems and Their Applications (DSA), Suzhou, China, 2–3 November 2024; pp. 35–42. [CrossRef]
48. *ISO/IEC 23053:2022*; Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). International Organization for Standardization: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/74438.html (accessed on 2 September 2025).
49. *ISO/IEC 25019:2023*; Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Quality-in-Use Model. International Organization for Standardization: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/78177.html (accessed on 2 September 2025).
50. *ISO/IEC TS 25058:2024*; Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Guidance for Quality Evaluation of Artificial Intelligence (AI) Systems. International Organization for Standardization: Geneva, Switzerland, 2024. Available online: https://www.iso.org/standard/82570.html (accessed on 2 September 2025).
51. *ISO/IEC 25023:2016*; Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Measurement of System and Software Product Quality. International Organization for Standardization: Geneva, Switzerland, 2016. Available online: https://www.iso.org/standard/35747.html (accessed on 2 September 2025).
52. *ISO/IEC 25022:2016*; Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Measurement of Quality in Use. International Organization for Standardization: Geneva, Switzerland, 2016. Available online: https://www.iso.org/standard/35746.html (accessed on 2 September 2025).
53. Lopez-Paz, D.; Ranzato, M. Gradient Episodic Memory for Continual Learning. *arXiv* **2022**, arXiv:1706.08840. [CrossRef]
54. *ISO 9241-110:2020*; Ergonomics of Human-System Interaction Part 110: Interaction Principles. International Organization for Standardization: Geneva, Switzerland, 2020. Available online: https://www.iso.org/standard/75258.html (accessed on 2 September 2025).
55. Lavazza, L.; Morasca, S. Understanding and Modeling AI-Intensive System Development. In Proceedings of the 2021 IEEE/ACM 1st Workshop on AI Engineering—Software Engineering for AI (WAIN), Madrid, Spain, 30–31 May 2021; pp. 55–61. [CrossRef]
56. *ISO/IEC TS 4213:2022*; Information Technology—Artificial Intelligence—Assessment of Machine Learning Classification Performance. International Organization for Standardization: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/79799.html (accessed on 2 September 2025).
57. *IEEE 2937-2022*; IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems. IEEE Standards Association: Piscataway, NJ, USA, 2022. Available online: https://standards.ieee.org/ieee/2937/10376/ (accessed on 2 September 2025).
58. Brand, L.; Humm, B.G.; Krajewski, A.; Zender, A. Towards Improved User Experience for Artificial Intelligence Systems. In *Engineering Applications of Neural Networks*; Iliadis, L., Maglogiannis, I., Alonso, S., Jayne, C., Pimenidis, E., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 33–44. [CrossRef]
59. Zender, A.; Humm, B.G.; Holzheuser, A. Enhancing User Experience in Artificial Intelligence Systems: A Practical Approach. In *Software, System, and Service Engineering*; Kardas, G., Luković, I., Milašinović, B., Popović, A., Radliński, Ł., Staroń, M., Swacha, J., Przybyłek, A., Eds.; Springer Nature: Cham, Switzerland, 2025; pp. 113–131. [CrossRef]
60. Vinayagasundaram, B.; Srivatsa, S.K. Software Quality in Artificial Intelligence System. *Inf. Technol. J.* **2007**, *6*, 835–842. [CrossRef]
61. *ISO/IEC TS 8200:2024*; Information Technology—Artificial Intelligence—Controllability of Automated Artificial Intelligence Systems. International Organization for Standardization: Geneva, Switzerland, 2024. Available online: https://www.iso.org/standard/83012.html (accessed on 2 September 2025).
62. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774. [CrossRef]
63. Fehr, J.; Citro, B.; Malpani, R.; Lippert, C.; Madai, V.I. A trustworthy AI reality-check: The lack of transparency of artificial intelligence products in healthcare. *Front. Digit. Health* **2024**, *6*, 1267290. [CrossRef]
64. European Commission. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment. 2020. Available online: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed on 2 September 2025).
65. Nicolae, I.E.; Danciu, G.; Nanou, C.; Koulierakis, N.; Danilatou, V. Transparency Metrics for Artificial Intelligence-Driven Applications in Healthcare. In Proceedings of the 13th Hellenic Conference on Artificial Intelligence, in SETN '24, New York, NY, USA, 11–13 September 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–8. [CrossRef]
66. Wan, A.; Klyman, K.; Kapoor, S.; Maslej, N.; Longpre, S.; Xiong, B.; Liang, P.; Bommasani, R. The Foundation Model Transparency Index. *arXiv* **2023**, arXiv:2310.12941. [CrossRef]
67. Barmer, H.; Dzombak, R.; Gaston, M.; Heim, E.; Palat, J.; Redner, F.; Smith, T.; VanHoudnos, N.M. *Robust and Secure AI*; Software Engineering Institute: Pittsburgh, PA, USA, 2021. Available online: https://www.sei.cmu.edu/library/robust-and-secure-ai/ (accessed on 2 September 2025).

68. *ISO/IEC TR 24029-1:2021*; Artificial Intelligence (AI)—Assessment of the Robustness of Neural Networks Part 1: Overview. International Organization for Standardization: Geneva, Switzerland, 2021. Available online: https://www.iso.org/standard/77609.html (accessed on 2 September 2025).
69. *ISO/IEC 24029-2:2023*; Artificial Intelligence (AI)—Assessment of the Robustness of Neural Networks Part 2: Methodology for the Use of Formal Methods. International Organization for Standardization: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/79804.html (accessed on 2 September 2025).
70. Powell, R.; Stockwell, S.; Sharadjaya, N. *Towards Secure AI. How Far Can International Standards Take Us?* Alan Turing Institute: Pittsburgh, PA, USA, 2024. Available online: https://cetas.turing.ac.uk/publications/towards-secure-ai (accessed on 2 September 2025).
71. *ISO/IEC DIS 27090*; Cybersecurity—Artificial Intelligence—Guidance for Addressing Security Threats and Compromises to Artificial Intelligence Systems. International Organization for Standardization: Geneva, Switzerland, 2025. Available online: https://www.iso.org/standard/56581.html (accessed on 2 September 2025).
72. Vassilev, A.; Oprea, A.; Fordyce, A.; Anderson, H. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*; NIST 100-2e2023; National Institute of Standards and Technology (U.S.): Gaithersburg, MD, USA, 2024. Available online: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf (accessed on 2 September 2025).
73. Hu, Y.; Kuang, W.; Qin, Z.; Li, K.; Zhang, J.; Gao, Y.; Li, W.; Li, K. Artificial Intelligence Security: Threats and Countermeasures. *ACM Comput. Surv.* **2021**, *55*, 1–36. [CrossRef]
74. Saeed, M.M.; Alsharidah, M. Security, privacy, and robustness for trustworthy AI systems: A review. *Comput. Electr. Eng.* **2024**, *119*, 109643. [CrossRef]
75. Liu, Q.; Li, P.; Zhao, W.; Cai, W.; Yu, S.; Leung, V.C.M. A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. *IEEE Access* **2018**, *6*, 12103–12117. [CrossRef]
76. Grosse, K.; Alahi, A. A qualitative AI security risk assessment of autonomous vehicles. *Transp. Res. Part C Emerg. Technol.* **2024**, *169*, 104797. [CrossRef]
77. Musser, M.; Lohn, A.; Dempsey, J.X.; Spring, J.; Kumar, R.S.S.; Leong, B.; Liaghati, C.; Martinez, C.; Grant, C.D.; Rohrer, D.; et al. *Adversarial Machine Learning and Cybersecurity*; Center for Security and Emerging Technology: Washington, DC, USA, 2023; Available online: https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/ (accessed on 2 September 2025).
78. Mitre. Navigate Threats to AI Systems Through Real-World Insights. Available online: https://atlas.mitre.org/ (accessed on 2 September 2025).
79. OWASP. OWASP Top 10 for LLM Applications 2025. 2024. Available online: https://genai.owasp.org/llm-top-10/ (accessed on 2 September 2025).
80. Bezombes, P.; Brunessaux, S.; Cadzow, S. Cybersecurity of AI and Standardisation | ENISA. 2023. Available online: https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation (accessed on 2 September 2025).
81. Barmer, H.; Dzombak, R.; Gaston, M.; Palat, V.; Redner, F.; Smith, T.; Wohlbier, J. *Scalable AI*; Carnegie Mellon University: Pittsburgh, PA, USA, 2021. Available online: https://kilthub.cmu.edu/articles/report/Scalable_AI/16560273/1?file=30632712 (accessed on 2 September 2025).
82. Mishra, A. *Scalable AI and Design Patterns: Design, Develop, and Deploy Scalable AI Solutions*; Apress: Berkeley, CA, USA, 2024. [CrossRef]
83. Scepanski, E.; Zillner, S. AI Systems and their Scalability—A Systematic Literature Review. In Proceedings of the ACIS 2024, Canberra, Australia, 4–6 December 2024. Available online: https://aisel.aisnet.org/acis2024/95 (accessed on 2 September 2025).
84. *ISO/IEC TR 29119-11:2020*; Software and Systems Engineering. Software Testing. Guidelines on the Testing of AI-Based Systems. International Organization for Standardization: Geneva, Switzerland, 2020. Available online: https://www.iso.org/standard/79016.html (accessed on 2 September 2025).
85. Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T. International AI Safety Report 2025. 2025. Available online: https://www.gov.uk/government/publications/international-ai-safety-report-2025 (accessed on 2 September 2025).
86. Center for AI Safety. *2023 Impact Report*; Center for AI Safety: San Francisco, CA, USA, 2023. Available online: https://safe.ai/work/impact-report/2023 (accessed on 2 September 2025).
87. Klüver, C.; Greisbach, A.; Kindermann, M.; Püttmann, B. A requirements model for AI algorithms in functional safety-critical systems with an explainable self-enforcing network from a developer perspective. *Secur. Saf.* **2024**, *3*, 2024020. [CrossRef]
88. *ISO/IEC TR 5469:2024*; Artificial Intelligence—Functional Safety and AI Systems. International Organization for Standardization: Geneva, Switzerland, 2024. Available online: https://www.iso.org/standard/81283.html (accessed on 2 September 2025).
89. *IEC 61508-1:2010*; Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems. International Electrotechnical Commission: Geneva, Switzerland, 2010. Available online: https://landingpage.bsigroup.com/LandingPage/Series?UPI=BS%20EN%2061508 (accessed on 2 September 2025).

90. *ISO 26262-1:2018*; Road Vehicles – Functional Safety. International Organization for Standardization: Geneva, Switzerland, 2018. Available online: https://blog.ansi.org/ansi/iso-26262-2018-road-vehicle-functional-safety/ (accessed on 2 September 2025).
91. *ISO 21448:2022*; Road Vehicles—Safety of the Intended Functionality. International Organization for Standardization: Geneva, Switzerland. Available online: https://www.iso.org/standard/77490.html (accessed on 2 September 2025).
92. Ji, J.; Venkatram, V.; Batalis, S. *AI Safety Evaluations: An Explainer | Center for Security and Emerging Technology*; CSET: Washington, DC, USA, 2025. Available online: https://cset.georgetown.edu/article/ai-safety-evaluations-an-explainer/ (accessed on 2 September 2025).
93. Challen, R.; Denny, J.; Pitt, M.; Gompels, L.; Edwards, T.; Tsaneva-Atanasova, K. Artificial intelligence, bias and clinical safety | BMJ Quality & Safety. *BMJ Qual. Saf.* **2019**, *28*, 231–237. [PubMed]
94. Varshney, K.R. Engineering safety in machine learning. In Proceedings of the 2016 Information Theory and Applications Workshop (ITA), La Jolla, CA, USA, 31 January–5 February 2016; pp. 1–5. [CrossRef]
95. Morales-Forero, A.; Bassetto, S.; Coatanea, E. Toward safe AI. *AI Soc.* **2023**, *38*, 685–696. [CrossRef]
96. Gordieiev, O.; Kharchenko, V. Profile-Oriented Assessment of Software Requirements Quality: Models, Metrics, Case Study. *Int. J. Comput.* **2020**, *19*, 656–665. [CrossRef]
97. Gordieiev, O.; Kharchenko, V.; Fusani, M. Software Quality Standards and Models Evolution: Greenness and Reliability Issues. In *Information and Communication Technologies in Education, Research, and Industrial Applications*; Yakovyna, V., Mayr, H.C., Nikitchenko, M., Zholtkevych, G., Spivakovsky, A., Batsakis, S., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 38–55. [CrossRef]
98. Gordieiev, O.; Rainer, A.; Kharchenko, V.; Pishchukhina, O.; Gordieieva, D. A Unified Approach to the Development of Technology-Based Software Quality Models on the Example of Blockchain Systems. *IEEE Access* **2024**, *12*, 118875–118889. [CrossRef]
99. Gordieiev, O.; Kharchenko, V.; Fominykh, N.; Sklyar, V. Evolution of Software Quality Models in Context of the Standard ISO 25010. In Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, Brunów, Poland, 30 June–4 July 2014; Springer: Cham, Switzerland, 2014; pp. 223–232. [CrossRef]
100. Gordieiev, O.; Kharchenko, V.; Vereshchak, K. Usable Security Versus Secure Usability: An Assessment of Attributes Interaction. In Proceedings of the International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications, Kyiv, Ukraine, 15–18 May 2017; pp. 727–740.
101. Hughes, A. *AI Models vs. AI Systems: Understanding Units of Performance Assessment*; Microsoft Research: Redmond, WA, USA, 2022. Available online: https://www.microsoft.com/en-us/research/blog/ai-models-vs-ai-systems-understanding-units-of-performance-assessment/ (accessed on 19 September 2025).
102. Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; Dennison, D. Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015. Available online: https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757 a2463eba-Abstract.html (accessed on 19 September 2025).
103. Felderer, M.; Ramler, R. Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session). In *Software Quality: Future Perspectives on Software Engineering Quality*; Winkler, D., Biffl, S., Mendez, D., Wimmer, M., Bergsmann, J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 33–42. [CrossRef]
104. Karnouskos, S.; Sinha, R.; Leitão, P.; Ribeiro, L.; Strasser, T.I. The Applicability of ISO/IEC 25023 Measures to the Integration of Agents and Automation Systems. In Proceedings of the IECON 2018—44th Annual Conference of the IEEE Industrial Electronics Society, Washington, DC, USA, 21–23 October 2018; pp. 2927–2934. [CrossRef]