Review Open Access

Review of Artificial Intelligence-Based Systems: Evaluation, Standards, and Methods

Shahrzad Oveisi[1*], Faezeh Gholamrezaei[1], Negin Gajari[1], Marjan Goodarzi[1], Mohammad Shahram Moein[1]

1. Center for Innovation and Development of Artificial Intelligence, Information and Communication Technology Research Institute, Tehran, Iran.

Abstract

The rapid expansion of Artificial Intelligence (AI) technologies and algorithms in various industries necessitates a focus on protecting the public interest. Major economies have heavily invested in AI initiatives, underscoring the significance of these advancements. Ensuring the dependability and quality of these systems is crucial to mitigate potential risks stemming from AI failures. In response, there have been efforts to establish monitoring frameworks and evaluation standards for AI products.

This paper presents a comprehensive analysis of more than 200 standards and publications to identify quantitative and qualitative metrics for evaluating AI systems throughout development and operation stages. The study also examines the methodologies, AI evaluation, and standards associated with these assessment criteria. The findings emphasize the importance of implementing robust evaluation frameworks to ensure the safety and effectiveness of AI systems. By reviewing various metrics and standards, this research offers valuable insights for policymakers, regulators, and industry professionals aiming to enhance AI oversight and governance. Moreover, the study highlights the necessity of continuous monitoring and evaluation throughout the AI development process to address potential risks and challenges. By advocating transparency and accountability in AI practices, stakeholders can build trust and confidence in the deployment of these technologies.

Keywords Artificial Intelligence (AI), Standard, Trustworthy AI, AI evaluation, Evaluation Criteria

1. Introduction & background research

The dramatic increase in Artificial Intelligence (AI) capabilities has led to a wide range of innovations that can advance nearly every aspect of our society and economy—from business and healthcare to transportation and cybersecurity. AI technologies are often utilized to exert a beneficial effect by informing, advising or simplifying tasks [1]; [2]; [3]. Since these products and services are among the most sophisticated technologies available, many companies are engaged in research and development in this field. Due to the ever-increasing growth of AI technology, this technology is expected to make great contribution to the transformation of raw data and the improvement of business processes in the near future.

Up to now, there have many debates over AI failures. According to a 2019 IDC survey, "most organizations have reported AI failures in some of their projects, with a quarter having a failure rate of up to 50%. These failures have historically been a strong and compelling motivator for adequate software testing, and industry surveys suggest that AI is one of the most important trends for software testing [4]. Therefore, requirements for the production of AI products have been presented in the form of standards due to the importance of such systems and the high costs

* Correspondence:
shahrzad.oveisi@gmail.com

related to their failure [5].
Standardization of AI products is a catalyst, indicating that a number of prerequisites are defined during the production of AI products from defining specifications to providing and processing services in order to lay the basis of production. Standards and evaluations have a special role in creating a stable framework for AI, promoting the rapid transfer of technologies through research and playing a decisive role in the expansion of international markets [6]. Besides, the implementation of standards promotes profitability. For example, the production of AI-based systems based on standards has resulted in a profit of approximately 17,000,000 euros for Germany [7]. AI is evaluated using checklists based on standards. A comprehensive and detailed examination of AI systems can prevent possible accidents or other potentially undesirable consequences of AI systems as shown in the case of smart cities [8]. AI Along with standards, a number of criteria have been developed for the evaluation of an AI system. The evaluation criteria of AI systems are not limited to the measurement of accuracy and error but include the quality assessment of the AI system. These criteria are divided into functional and non-functional evaluation criteria. The former are specific tasks that a system should be able to perform and are related to the intended purpose of the system, which include things like input/output, processing, and data storage requirements [10] [11] [12]. On the other hand, non-functional criteria are constraints that specify how well the system should perform its functional tasks and how it must behave under different conditions, including such things as performance, scalability, security and usability [9].
In this paper, after examining more than 200 sources, we presented the evaluation methods of AI products for the first time according to both quantitative and qualitative criteria during the life cycle of AI products (data, hardware, software, and machine learning models). Through comprehensive evaluations across these four categories, stakeholders gain valuable insights into the strengths and weaknesses of AI-based systems. This facilitates improvements, advancements, and the development of more reliable and effective AI solutions) by reviewing standards, checklists and methods. Accordingly, we divided these criteria during the production life cycle of AI systems into evaluation methods as well as trustworthiness criteria. Trustworthiness AI evaluation is a process aimed at assessing and ensuring the reliable behavior and performance of AI. This process involves defining appropriate criteria and indicators for evaluating AI, designing and generating test samples, conducting tests, and analyzing the results. Ultimately, to enhance the trustworthiness of AI, necessary adjustments and optimizations can be made to the system. This process is crucial for ensuring confidence and dependability of AI in various applications. According to experts' opinion, the most important and prioritized criteria are as follows: 1) Transparency, explainability and interpretability [325]; [326]; [327]; 2) Safety and reliability [328]; [329]; [330]; 3) Bias [331]; [332]; 4) Robustness [333]; [334]; [335]; [336]; 5) Security [337]; [338]. Besides, specific testing methods have been presented for AI products section, including software, hardware, data, and machine learning methods.
The rest of the paper is organized as follows. After the introduction, in the second section, the basics of research are presented, including importance of standardization and checklists. In the third section, methods are defined with an emphasis on the evaluation of AI-based systems (1- Hardware, 2- Software, 3- Data, and 4- Machine Learning models) and trustworthiness evaluation measurements. In the fourth section, summary and results are presented.

2. Research Fundamentals

In this section, before starting the main sections, necessary research background has been briefly reviewed. In section 2.1, the important role of standards in life cycle of AI-based products has been mentioned, and in section 2.2, the application of checklists in the evaluation of AI-based systems is presented.

2.1. The Importance of AI Standardization

Standards have become important indicators for measuring technology development rates. The culture of countries and regions as well as the formal rules of introducing products into the market are among the basic, supporting and guiding reasons for the standardization of AI. The important and key roles of AI standardization are presented in Table 1 [15].

2.2. The Importance of Artificial Intelligence Checklists

AI-based products are utilized in various applications, so that the trained model is accurate, fair, robust and resistant to attacks [13]. Systematic evaluation of AI-based products is done by checklists. AI-based product evaluation checklists include guidelines and criteria used by manufacturers and laboratory AI-based system assessors. By validating the information presented in the checklist, the evaluation laboratory provides a reliable, explainable and transparent product for all the stakeholders [14].
In Table 2, the three key roles of an AI-based system are presented. Each of these roles must provide the required documentation based on the specified task, and the documentation provided by the manufacturers is examined by the laboratory assessor.

Table 1. The Important Role of AI Standardization

The role of standardization in Artificial Intelligence	Standardization
• Facilitating and accelerating innovations in the field of artificial intelligence • Assisting in commercializing achievements in the field of artificial intelligence Furthermore, standardization can serve as a tool for strengthening technical accomplishments".	**The guiding role of standardization in technology innovation and support for industrial development.**
• Improving the quality of artificial intelligence products and services • Developing an integrated standard accompanied by methods for conducting necessary conformity Evaluation tests and evaluations	**Fast Realization of Innovation Universality**
• Increasing user safety • Protecting user rights • Adhering to human principles • Ensuring information security	
• Establishing fair and open ecosystems Currently, industry giants use methods such as open-source algorithms to create frameworks for deep learning and other ecosystems, which makes it more difficult to transfer user data. This requires unified standards for achieving collaboration and coordination between producers to prevent industry monopolies and user lock-ins, and to create fair and competitive industrial ecosystems.	

Table 2. Key roles of AI-based system

1. Ensuring testing and evaluation of the system according to established guidelines, criteria, and regulations. 2. Developing, deploying, and maintaining an AIsystem. 3. Making informed risk-reward decisions. 4. Responsible for operating and utilizing the algorithm. 5. Ensuring sustainable AIoperations.	Owner of AI
1. The algorithm owner is responsible for ensuring transparency, accountability, and ethicality of the algorithm. 2. Ensuring testing and evaluation of the algorithm's compliance with established guidelines, criteria, and regulations. 3. Documenting the algorithm's performance in documentation. 4. Monitoring, maintaining, and updating algorithms. 5. Making informed risk-reward decisions.	Algorithm Owner

3. Methods

The evaluation of AI-based systems is divided into four categories (Figure 1): 1- Hardware, 2- Software, 3- Data, and 4- Machine Learning models. In section 3.1, we will examine trustworthiness evaluation measurements in AI system life cycle. In sections 3.2 to 3.5, the four categories of AI based systems and evaluation of these methods are presented.

3.1. Trustworthiness Evaluation Measures in AI system Life Cycle

AI system life cycle provides people with a suitable framework that can guide them towards their goal. The main role of AI system life cycle is to distribute the development of AI project into different phases so that the development becomes easier and clearly understandable, and the phases should be more specific to effectively achieve the best possible output. In general, architecture of the production cycle of AI-based products includes seven phases as follows: 1) Conceptualization; 2) Design and development; 3) Validation and verification; 4) Deployment; 5) Operations and monitoring, 6) Re-evaluation; and 7) Completion of operations.

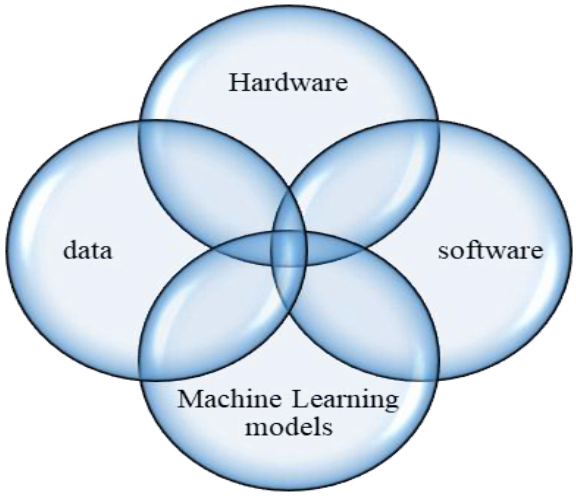

Fig 1. AI system Layers

Quantitative and qualitative assessments are used to test and evaluate the performance of AI software. The former

Table 3. Checklist of AI

Ethical Quality model AI	Robust AI	Legal AI
1. Technical 2. Non-technical	1. Human agency and oversight 2. Technical robustness and safety 3. Privacy preservation and data governance 4. Transparency 5. Diversity 6. Non-discrimination and fairness 7. Environmental and societal well-being 8. Accountability	• Explainability • Fairness • Respect for human autonomy • Prevention of harm

is evaluation based on system code testing and implementation methods and the latter is divided into several categories, which are the subsets of risk evaluation criteria and reliability measurement as follows: 1) Transparency, explainability and interpretability; 2) Safety and reliability; 3) Bias; 4) Sustainability; 5) Security, which should be checked to measure the trustworthiness of AI systems. To achieve these goals, the architecture shown in Figure 2 was taken into account in this paper to review the standards, checklists, and evaluation criteria. This architecture performs development review operations in phases 2-6; transparency and explainability, security and privacy review operations in phases 2-7; risk management and governance operations in phases 1-7.

Also, by reviewing the papers and standards, the checklists of this field have been presented in detail in Table 3.

3.1.1. Transparency and Explainability

The complexity of AI-based systems can lead to problems in understanding for both users and developers. This "understanding" can generally be considered in terms of clarity, interpretability and explainability of a system as follows:

-Transparency- The level of access to the algorithm and data used by the system based on AI ;

-Interpretability- The level of understanding of the way the underlying technology works;

-Explainability- The level of understanding how the AI-based system reached a certain result .

By reviewing the papers and standards, in Table 4 and Figures 3 and 4, the standards and methods of this field have been examined in detail [14], [17], [18], [242], [243], [244], [245], [246], [247], [248], [249], [250], [251], [252].

Evaluation methods related to explainability are important for assessing the transparency and interpretability of machine learning models. Two common subcategories of evaluation methods for explainability are explainable modeling and posthoc explanation .

Explainable modeling involves designing and developing machine learning models that are inherently interpretable. This can involve using simpler algorithms that are more transparent, incorporating human-readable features, or providing a clear understanding of the decision-making process within the model .

Posthoc explanation methods, on the other hand, involve explaining the decisions made by a complex machine learning model after it has already been trained. This can include techniques such as feature importance analysis, local explanation methods (e.g. LIME or SHAP), or generating text or visual explanations to justify model predictions.

3.1.2. Security and Privacy

AI-based systems are expanding and developing, and they are used in many fields as a significant alternative to traditional methods. However, due to the high complexity and extensive potentials that these systems provide, the security and privacy may be at risk, which is why security evaluation is of particular importance .

By reviewing papers and standards, Tables 4 and 5 of standards and checklists [14], methods [19], [20], [21], [22] [23], [24], [25] have been reviewed in detail in Figures 5, 6 and 7 [245], [246], [247], [248], [249], [250], [251], [252].

As seen in Figure 5, various security evaluation methods are described. Security evaluation methods are techniques used to assess the security of a system or application. These methods involve systematically analyzing the security controls, vulnerabilities, and potential risks associated with the target system.

Privacy Evaluation Methods in Figure 6 are described. Privacy evaluation methods are techniques used to assess the level of privacy protection and compliance within an organization or system. These methods can include conducting privacy impact assessments, auditing privacy practices, or using privacy control frameworks to evaluate the effectiveness of privacy protections in place.

In Figure 7, we discussed penetration check methods, which include two groups: fault tolerance and failure tests. Fault tolerance methods are designed to prevent system failures by providing redundancy or backup mechanisms, such as duplication or mirroring of critical components. Failure tests, on the other hand, are used to actively test

Table 4. Transparency& Explainability

Transparency& Explainability
ISO/IEC 24027 [180]; ISO/IEC 24028 [181]; ISO/IEC 5338 [182];ISO/IEC 24368 [183]; ISO/IEC 4213 [184];IEEE P7000- 14 [185]; IEEE P2863 [186];IEEE P3652-1 [187];ISO/IEC 23894.2 [188];ISO/IEC 42001 [189];Transparency of Autonomous Systems (defining levels of transparency for measurement) (IEEE P7001) [190]; Certification for Products and Services In Transparency, Accountability, And Algorithmic Bias In Systems (IEEEECPAIS) [191].

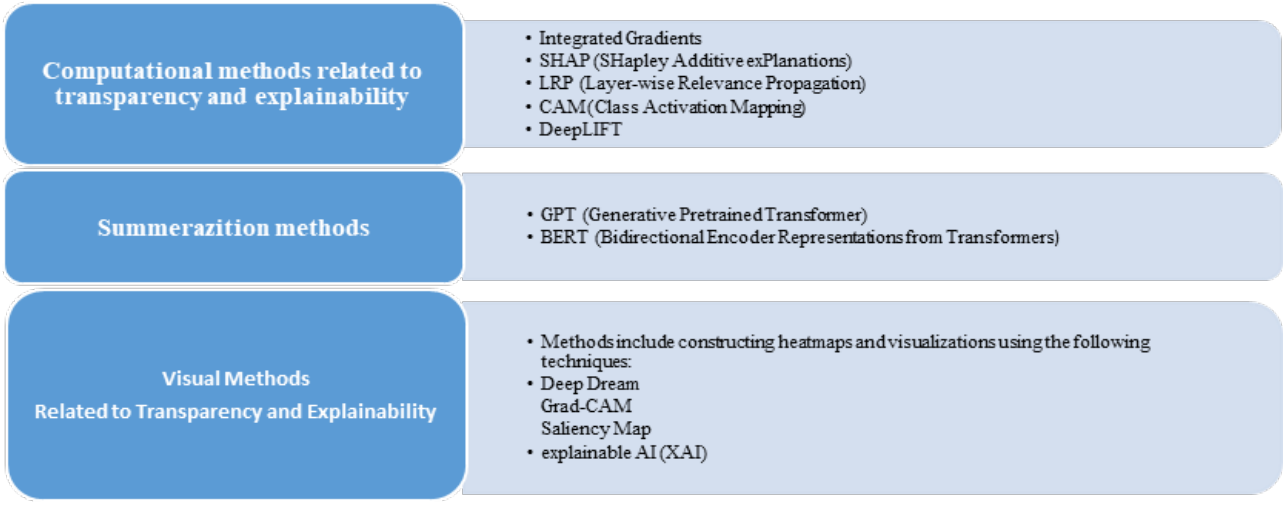

Fig 3. Methods related to transparency and explainability

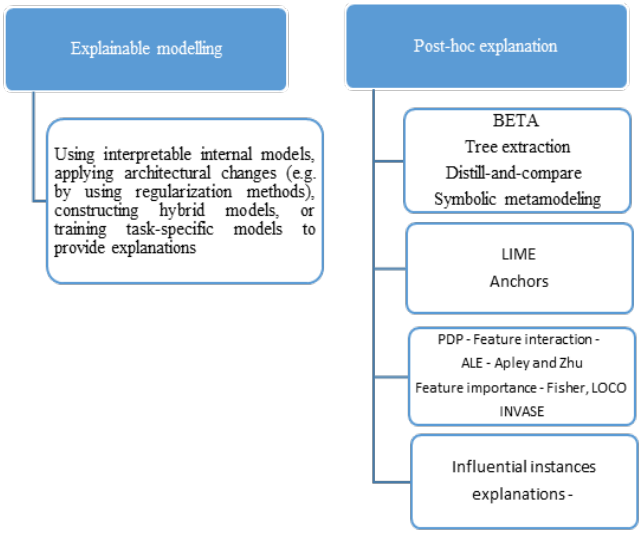

Fig 4. Evaluation Methods Related to Explainability

system vulnerabilities and identify weak points that may lead to system failures. These tests involve intentionally causing faults or failures in the system to assess its resilience and ability to recover from such events. Both fault tolerance and failure tests are important components of penetration testing, as they help to ensure the overall security and reliability of a system.

3.1.3. Bias

Bias is an important factor in AI evaluation, which refers to the concept of distortion in the collection, processing and/or interpretation of data by AI systems. Bias can lead to discrimination, as well as unfair decisions and results obtained by AI.[201][202][203].

Different approaches can be utilized to evaluate AI-based systems in terms of bias. The first step in bias evaluation is to identify and analyze the training data of the system. For example, if the training data contains only samples with certain characteristics, the system may make wrong decisions for data with different characteristics. By reviewing the papers and standards, in Figure 8, methods of this field have been examined in detail [204][205][206][207][208][209][210][211][212].

That can be divided into three categories: Bias mitigation algorithms aim to reduce the impact of biases in data and decision-making processes, while bias reduction techniques focus on eliminating biases entirely from a dataset or model, Data Bias refers to systematic errors introduced during the data collection process or data analysis that result in a misrepresentation of the true information .

3.1.4. Robustness Evaluation Methods

Recognizing the strengths and weaknesses of AI systems against new data is one of the most important challeng-

Table 4. Security and privacy standards

Standardization of security and privacy for AI-based products
ISO/IEC TS 4213 [184]; ISO/IEC 5338 [182]; ISO/IEC 4669 [192]; ISO/IEC 5469 [193]; ISO/IEC 23894.2 [188]; ISO/IEC 24029-1 [194]; ISO/IEC 24668 [195]; ETSI SAI 006 [196]; ISA/IEC 62443 [197], [198], [199], [200].

Table 5. Checklists Based on Security and Privacy from the Perspective of AI Governance

Continuous monitoring	Privacy Impact Evaluation	Third-party security	Risk Evaluation	Compliance	Incident Response Plan	Data retention	Encryption	Access control	Data protection measures
on-site evaluation of continuous monitoring processes	Risks and data protection impact Evaluations (DPIAs).	Evaluation of existing security measures for vendors and third-party partners.	Examination of risk Evaluation and current management processes.	Evaluation of system compliance with relevant regulations and standards such as GDPR and ISO 27001	Evaluation of system incident response plans and procedures.	Evaluation of data retention policies and procedures.	Examination of encryption methods used.	Evaluation of system access control mechanisms	**Evaluation of existing measures for protecting sensitive data and ensuring privacy**

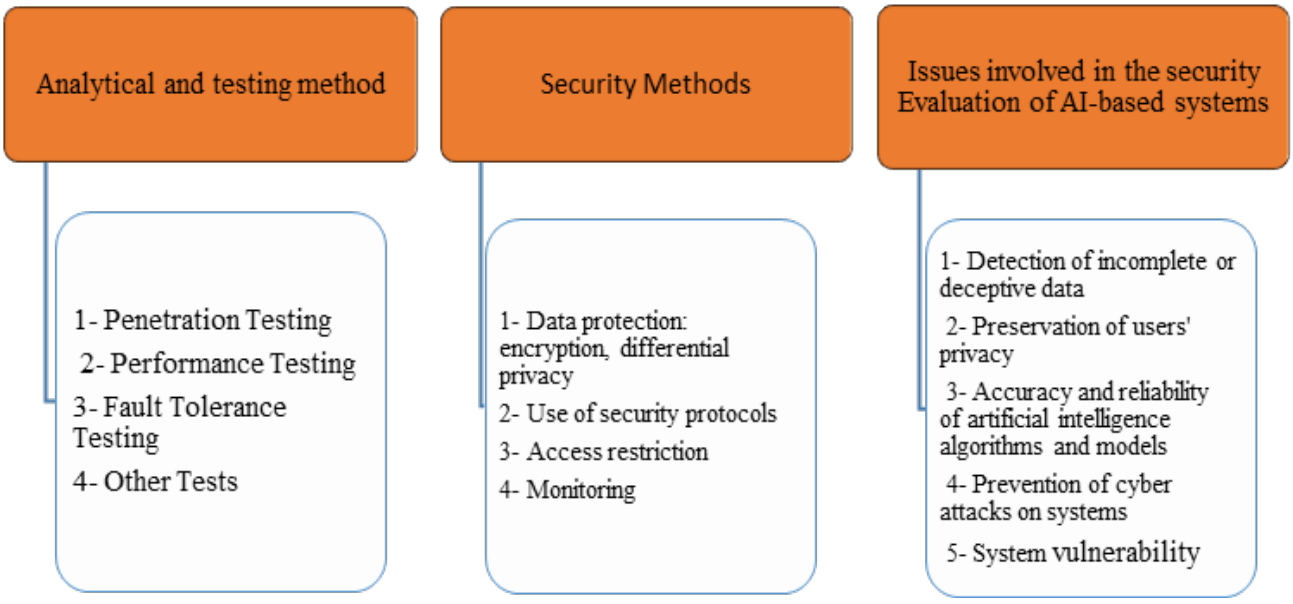

Fig 5. Security Evaluation Methods

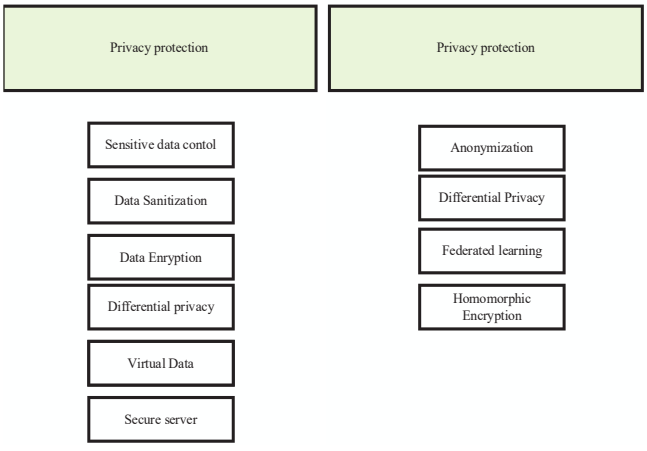

Fig 6. Privacy Evaluation Methods

es for this industry. Therefore, most AI systems have adopted the approach of being highly sensitive to environmental changes and noises and being prone to easy failure, which is why robustness evaluation methods have become of high importance. By reviewing the papers and standards, in Table 9 and Figure 9, the methods and standards of this field have been presented in detail in [257], [258], and [259].

In Figure 9, we have categorized the methods of robustness evaluation into four categories: removing outliers, dealing with data noise, evaluating the performance of several neurons, and adding new data.

Removing outliers: This category includes methods that focus on identifying and removing outlier data points from the dataset. Outliers can significantly affect the per-

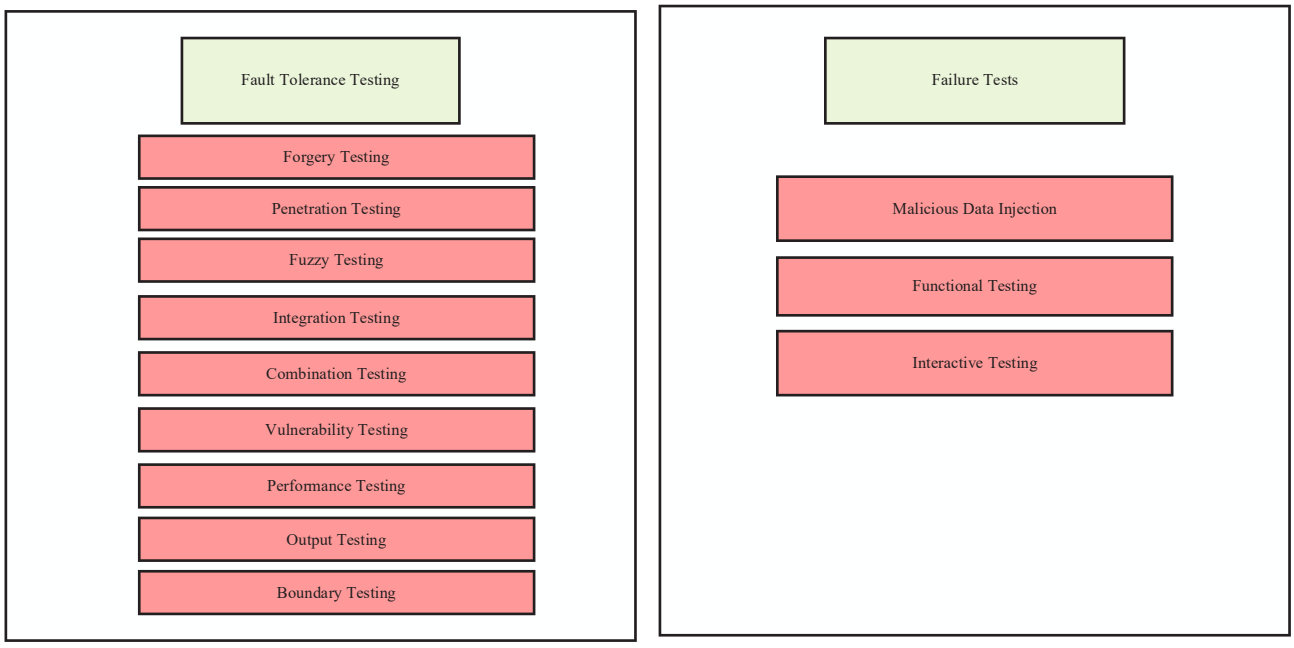

Fig 7. Penetration Check Methods

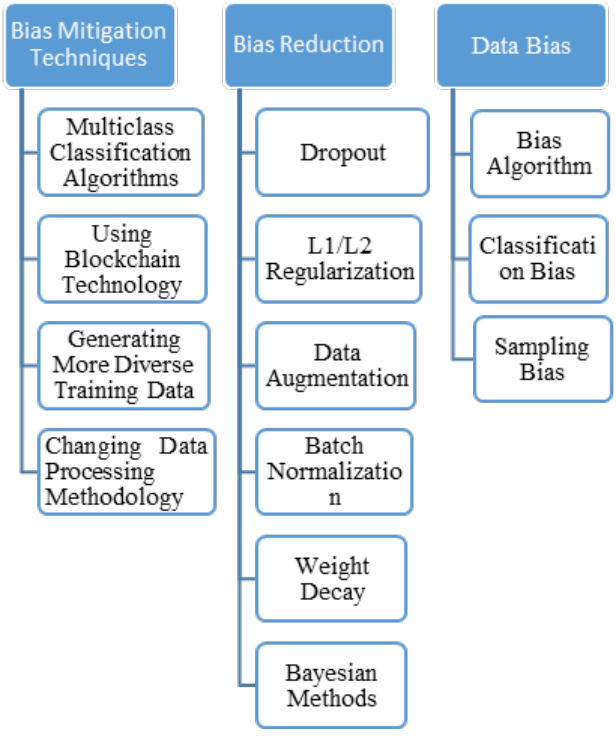

Fig 8. Bias Evaluation methods

Table 9. Robustness Standards

Robustness Standards and Methods
ISO/IEC TR 24029-1

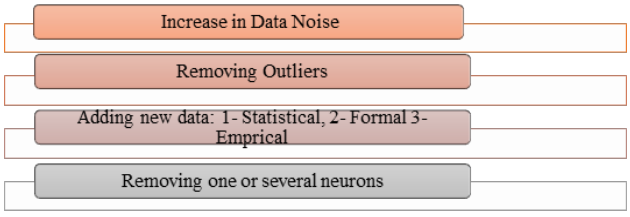

Fig 9. Robustness Evaluation Methods

formance of a model and by removing them, we can improve the robustness of the model.

Dealing with data noise: This category includes methods that aim to address the presence of noise in the dataset. Noisy data can lead to inaccurate predictions and can make the model less robust. Techniques such as smoothing or filtering can be used to reduce the impact of noise on the model. Robustness in the context of removing one or several neurons refers to the ability of a neural network to maintain its performance and functionality even when certain neurons are removed. This is an important aspect of evaluating the effectiveness and reliability of neural networks, as it can provide insights into how well the network can adapt to changes and disruptions.

3.1.5. Safety and Reliability

In computer science, safety means ensuring that a system, software or device poses no risk to the user or the surrounding environment under any circumstances. In

Table 6. Safety Evaluation standards

Safety - Application
Railway- EN 20126 20128- EN20129; Elevator EN 811/PrA2; Autonomus/ ISO 26262; Machinary/ISOM13849 Process/IEC 61511; ISO/PAS 21448; UL/4600; ISO/IEC AW TR 5469; VDE-AR-E-2842-61-1; IEC 61508; ISO/IEC 24027: ISO/IEC 24028; IEEEP2846; ISO/DIS Road Vehicle -Functional Safety; IEC, "Functional safety of electrical/electronic/programmable electronic safety-related systems," IEC 61508:2010; ISO, "Road Vehicles – Functional Safety"; ISO, "Road Vehicles – Functional Safety" ISO 26262:2018; ISO, "Road Vehicles – Safety of the Intended Function"; ISO/PAS 21448:2019; Koopman, P. & Wagner, M., "Toward a framework for highly automated vehicle safety validation," SAE 2018-01-1071, 2018; Koopman, P. & Fratrik, F. "How many operational design domains, objects, and events?" SafeAI 2019; Ministry of Defence, "Safety Management Requirements for Defence Systems," Defence Standard 00-56, 2017; SAE, Guidelines and Methods for Conducting the Safety Evaluation Process on Civil Airborne Systems and Equipment, ARP4761, 2012; US Dept. of Commerce, https://www.commerce.gov/issues/regulatory-reform, 7 June 2019. US DoD, "Standard Practice: System Safety", MIL-STD-882E, 11-May-2012; AC 23.1309-1, System Safety Analysis and Evaluation for Part 23 Airplanes.

Table 7. General safety of artificial intelligence software [240], [241]

Software development phase	Software Safety Tasks
Conceptual Design	• Initial Hazard Analysis • Software Safety Program
Software Requirements Analysis	• Safety requirements analysis • Hazard testing • Review of safety requirements • FTA/FMEA based on user specifications
Software Architecture Design	**Software Architecture Design:** FTA/FMEA of software systems
Coding and Detailed Designing	• **Partial software safety analysis:** FMEA/FTA • Code-level safety analysis Defense Programming
Testing, Integration, and Verification	• **Software safety testing, analysis of software safety tests, and software safety case studies:** • Defense Programming

general, safety indicates preventing accidents and reducing risks associated with the use of complex systems and technologies [261], [262], [263], [264], [265], [266], [267], [268], [269], [270], [271], [272], [273], and [279]. In various industries, including automotive, aerospace, medical, and military industries, safety is one of the most important factors for designing and manufacturing products. In addition, safety is considered an important topic in the field of computer security as well as protection of data and computer systems .

To check the safety of AI systems, we assess these methods from two perspectives :

•Software safety in the production cycle of AI products

•Safety of algorithms of machine learning models in the production cycle of AI products

By reviewing the articles and standards, Tables 6, 7 and 8 [280], [281], [282], [283], [284], [285], [286], [284], [288], [286], [290], [291], [292], [293], [294], [295], [296], [297], [298], [299], [300], [301], [301], [302], [303], [304], [305], [306], [307], [308], [309], [310], [311], [312], [313], [313], [314], [315], [316], [317], [318], [319], [320], [321], [322], [323] present the checklists, methods and standards of this area in detail.

3.2. Hardware Development

We use the RAMI model as one of the most common and famous models in the development of AI products. RAMI 4.0 is the real reference architecture for Industry 4.0 with relevant standards. Accordingly, standardization during the development of AI systems is carried out according to Figure 10.

By reviewing the papers and standards, the standards of this field have been presented in detail in Table 9.

3.3. Software Testing

Most AI-based systems are composed of one or more AI components (e.g., an ML model) surrounded by a substantial array of traditional software that provides a supporting infrastructure, typically consisting of common components such as a user interface and a database. Even "pure" AI components are implemented in software and therefore can be flawed like any other software. There-

Table 8. Safety of Algorithms Of Machine Learning Models

Examples	Method categories	Method categories
scenario coverage [337] - input space ontology [338]	Data representatively requirements	**Requirements Engineering** Use available domain knowledge to formulate use-case and safety requirements
- adversarial robustness metric [339]	Robustness requirements	
- runtime monitoring - model diversity measure for redundancy [342,343,366]	System fault tolerance requirements	
- occlusion sensitivity [352]	Safety performance measures	
- AD behavior rules - sensible intermediate steps [358] - domain specific rules (physical, legal)	Plausibility requirements	
	Experience collection	
experience on model type, training method, initialization values [342]	Design based on experience	**Development Apply** Reasonable Design Choices at All Decision Points
[345,346]	Incorporation of uncertainty	
via loss function NNs [348, 36] – via topology NNs [349] - model repair RL [350] - safe learning [351, 352]	Inclusion of expert knowledge	
- regularization [353] - robustification of training data [353] - counterexample -guided data augmentation [354]	Robustness enhancements	
- solvers NNs [355,356] - boundary approximation NNs [357, 358] - search algorithms [359]	Formal verification of rules, model, KPIs	**Verification** Check Against Test Data and Model Requirements
		Validation **Find Missing.**
	Input space coverage checks	**. . . Test Cases**
	Experience coverage checks	
- concolic testing NNs [360] - counterexample generation [360, 361, 362]	Model coverage checks	
- back-propagation based, e.g. NNs [363] - model agnostic, e.g. [364,365,366]	Attention analysis through heatmapping	**. . . Requirements Via Qualitative Model Analysis**
- method collection [367]	Feature visualization	
- textual explanations NNs [368] - hierarchical information [369]	Explanatory output	
- locally via ILP [370] - global model agnostic, e.g. VIA [371] - global NN specific methods [372]	Rule extraction	**. . . Requirements Via Quantitative Model Analysis**
- Neural Stethoscopes NNs [373] - concept embedding and attribution analysis NNs [375, 374] - ReNN modularized topology [374]	Sub-task/concept analysis	

fore, when testing an AI-based system, conventional software testing methods are still required. However, AI-based systems have a number of special features that necessitate additional testing relative to conventional software systems. Figure 11 shows the types of tests in products based on AI .

3.4. Data governance

In AI(AI), governance is often presented in two parts: data and AI. The former refers to the policies, procedures and practices that organizations have in place to manage the data they collect, store and use. It is critical for ensuring the quality, integrity, and security of data, which are in turn essential to the development and deployment of AI systems.

Data governance also involves privacy, ethics and data transparency. On the other hand, AI governance refers to the policies, regulations, and guidelines established to control the development, deployment, and use of AI systems [26][27]. In data governance, the quality and integrity of data, communication with other areas, compliance with access protocols and data processing capabilities are covered.

In the field of data governance, there are three key roles

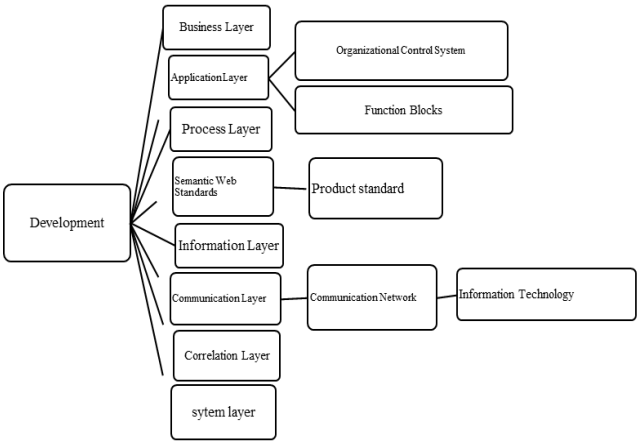

Fig 10. AI-based product development layers

of data management, data supervisor and data technician. All three roles are responsible for overseeing the three areas of data operations, data quality monitoring, and data quality improvement. Figure 12 shows the tasks of each role in each area.

By reviewing the papers and standards, Tables 10, 11, 12, 13 and 14 present the checklists, methods and standards of this field in detail [28]; [16]; [29]; [30]

Data governance frameworks provide a systematic approach to managing and regulating data management in an organization. The most famous frameworks are as follows: Strategic Information Management Model (SAM), Asset Information Model (AIM), DAMA-DMBOK Framework, DMBOK Framework, DAMA Advanced Data Management Framework, Data Governance Institute (DGI)

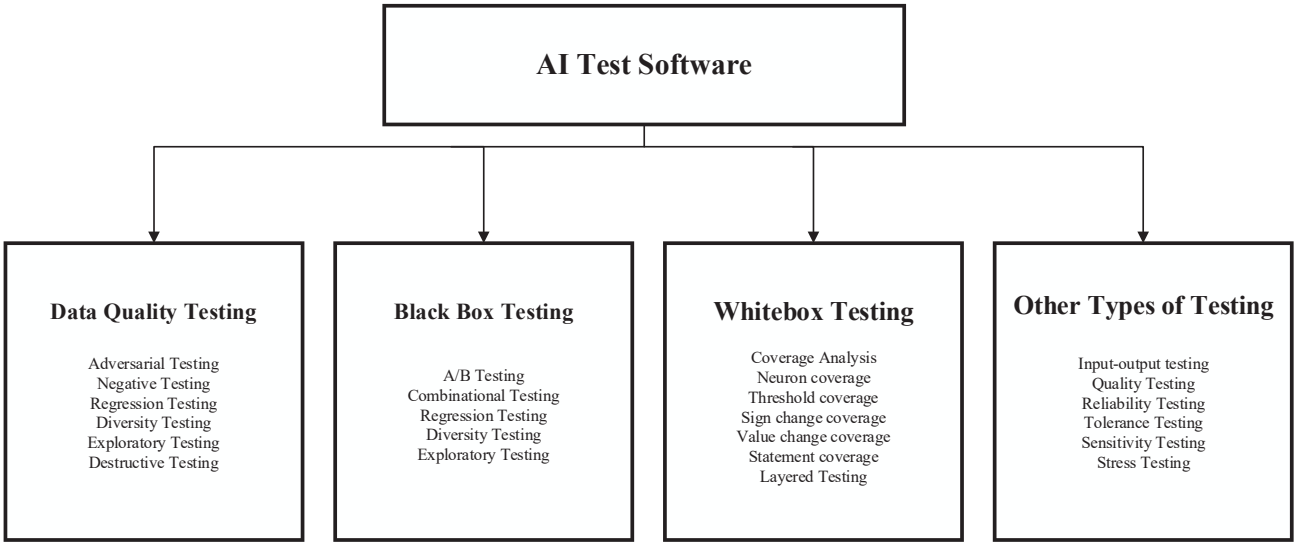

Fig 11. Types of Testing in Products Based on Artificial Intelligence

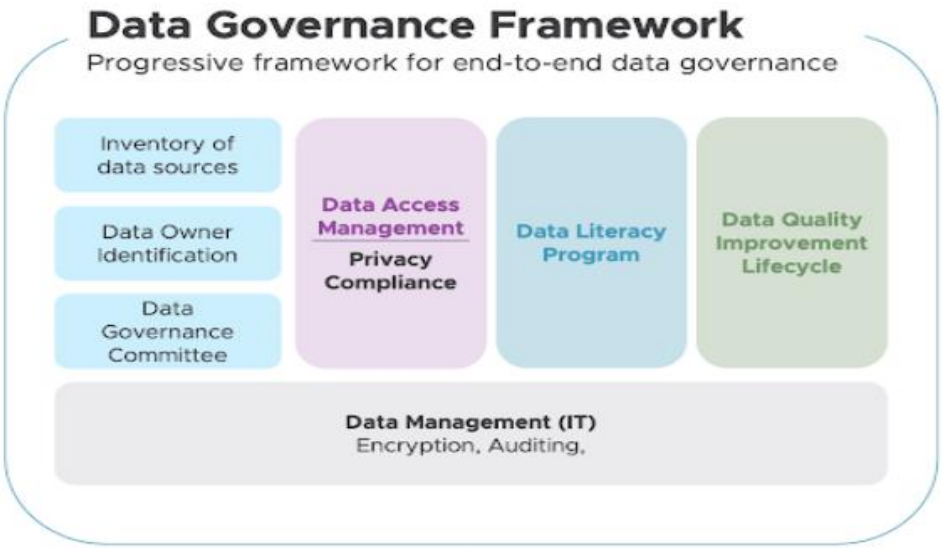

Fig 12. Data Quality

Table 9. Development phase standards

ISO/IEC JTC1 AG08 Standard	Development Layer
IEC/ISO 62264 - Enterprise Control System [35], [36], [37], [38], [39], [40]; ISO/IEC 30141 [41]; IEC 61499 – Function blocks [42], [43], [44]; IEC 61512 – Batch control [45], [46], [47], [48]; ISO/IEC 42010 – Systems and software engineering – Architecture description [49]	**Business Layer**
ISO/IEC JTC 1 SC41 [50]	**Application Layer**
RFC8259 – The JavaScript Object Notation (JSON) Data Interchange Format [51]; IEC 63339 ED1 [52]; ISO 13584-4/IEC 61360 – Classification and product description [53], [54]; IEC 61987 – Industrial-process measurement and control – Data structures and elements in process equipment catalogues [55]; ISO 29002-5 [56]; AML IEC 62714 – Engineering data exchange format for use in industrial automation systems engineering – Automation Markup Language [57], [58], [59], [60], [61]; ISO 13584-4/IEC 61360 – Classification and product description [62], [63]; IEC 61987 – Industrial-process measurement and control – Data structures and elements in process equipment catalogues [64];IEC 62714 – Engineering data exchange format for use in industrial automation systems engineering – Automation Markup Language [65], [66], [67], [68], [69]; ISO/IEC 19788-7:2019 – Information technology – Learning, education and training – Metadata for learning resources [70]; ISO 16684-1:2019 – Graphic technology – Extensible metadata platform (XMP) [71]; ISO/IEC 15944-10:2013 – Information technology – Business operational view – Part 10: IT-enabled coded domains as semantic components in business transactions [72] ISO 1087:2019 – Terminology work and terminology science – Vocabulary [73]; ISO/IEC 1179 – Metadata Registry (MDR) [74, 75]; ISO 21597-1:2020 – Information container for linked document delivery – Exchange specification [76]	**Process Layer**
IEC 61158-1 [77]; IEC 61784-2 [78]; ISO/TC 184 – IEC/TC 65/JWG 21 [79], [80]; IEC 62591:2016 [81]; IEC 62601 (WIAPA), IEC 62734 (ISA100a) [82], [83]; IEC 62948 (WIA-FA) [84]; IEC 62657-2 [85]; IEC 62541-1 presents the concepts and overview of the OPC Unified Architecture and serves as the basis for further specification [86].; IEC 62541-2 – security model [87], IEC 62541-3 Address Space Model [88]; IEC 62541-4 services [89], IEC 62541-5 Information Mode [90]l; IEC 62541-6 Mappings [91]; IEC 62541- 7 Profiles [92]; IEC 62541-8 Data access [93], IEC 62541-9 Alarms and Conditions [94]; IEC 62541-10 Programs [95]; IEC 62541-11 Historical access [96]; IEC 62541-12 Discovery and global [97]; IEC 62541-13 Aggregates [98]; IEC 62541-14 PubSub [99]; IEC 62541-100 Device interface [100]; IEC 61784 – Industrial communication networks [101]; IEC 61158 series, to be used in the design of devices involved in communication in factory manufacturing and process control [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126], [127], [128], [129], [130], [131], [132], [133], [134], [135], [136], [137], [138], [139]; IEC 61784 2 [140]; IEC 61784 3 series [141]; IEC 61784-5 series [142]; IEC 61918. [143]; IEC 61850:2020 – Communication networks and systems for power utility automation [144]; ANSI/MTC1.4 – 2018 – MT Connect [145]; ISO/IEC 19464:2014 – Information Technology – Advanced Message Queuing Protocol (AMQP) v1.0 specification [146] ; The IEEE FIPA (Foundation for Intelligent Physical Agents) specifications represent a collection of standards which are intended to promote the interoperati of heterogeneous agents and the services that they can represent [147];.IEC 61850:2020 – Communication networks and systems for power utility automation [148];ISO/IEC 19464:2014 – Information Technology – Advanced Message Queuing Protocol (AMQP) v1.0 specification [149]; ISO 18000-7 – Information technology – Radio frequency identification for item management – Part 7: Parameters for active air interface communications at 433 MHz [150]; IEEE 802.15.1 – Bluetooth LE [151]; IEEE 802.15.4 – ZigBee [152]; IEEE 802.11 – Wireless Local Networks [153];	**Information Layer**
ISO 11354 - Advanced automation technologies and their applications — Requirements for establishing manufacturing enterprise process interoperability [154]; IEC 63339 [155]; ISO 15745 Industrial automation systems and integration — Open systems application integration framework [156]; IEC 62264 Enterprise-control system integration [157], [158], [159], [160], [161], [162], [163]; IEEE 2660.1-2020 - IEEE Recommended Practice for Industrial Agents: Integration of Software Agents and Low-Level Automation Functions [164]; VDI/VDE 2653 - Multi-agent systems in industrial automation [165].	**Integration Layer**
IEC 62832 CD 2 Part 1 [166]; IEC 61360 [167], [168], [169]; ISO 29002-5 [56]; ISO/IEC 11179-5 [56]; IEC 61131 – Programmable controllers [170]; IEC 62832 [171], [172], [173]; IEC 61508 - functional safety requirements [174]; IEC 62443 - Industrial communication networks - IT security for networks and systems [175], [176], [177], [178]; IEC 61131 – Programmable controllers [179].	**System Layer**

Table 10. Data governance standards

Data Governance Standards
ISO/IEC 5259-6 [213], PWI 8183 [227]; 8183 AI Data life; 5259-X Data quality [228];22989 AI Functional Overview [223];ISO/IEC 38507 Information Technology- Governance of IT- Governance implications of the use of artificial intelligence by organization [229]; ISO/IEC 5259-(1-4) [230]; ISO/IEC 24372 [225]; ISO/IEC 24668 Process management framework for big data analytics [195]; ISO/ IEC 5259-x (1-7) Data quality for analytics and ML [231], [232], [233], [234], [235]; ISO/IEC 20547-1: 2020 Information technology- Bigdata inference architecture – part1: framework and application process [236]; ISO/IEC 20547-2: 2020 Information technology- Bigdata inference architecture – part2: use cases and derived requirements [237]; ISO/IEC 20547-3:2018 Information Technology- Bigdata reference architecture- part 5: Standards Roadmap [238].

Table 11. General and specific evaluation criteria of data sources of AI-based systems

Data Evaluation Criteria	
General data protection regulations (GDPR) Sensitive data quality	**Cloud computing**
Security	**Hidden data**
Integration of various resources	**Internet of Things (IoT)**
Security in distributed processing	**Blockchain**

Table 12. Data governance Evaluation checklist

1.Features of content management systems 2. Organizational content management systems 3. Document management systems 4. Content and Document management standards 5. KPIs for content and document management (KPI stands for Key Performance Indicators)	**Document management**
1. Extract, Transform, Load (ETL) process 2. Data latency or data lag 3. Interaction models 4. Data interoperability and integration architecture 5. Interoperability standards 6.KPIs for data interoperability and integration (KPI stands for Key Performance Indicators)	**Data integration**
1.Secondary data 2. Primary data 3. Difference between primary and secondary data 4. Standards for primary and secondary data 5.KPIs for primary and secondary data (KPI stands for Key Performance Indicators(	**Master data management**
1.. Data encryption 2. Types of data security 3. Classification of data security 4.Data security risks 5.Data security standards 6.Data security KPIs (KPI stands for Key Performance Indicators(	**Data security**
1.Data warehouse structure 2. Data warehouse standards 3. Data warehousing KPIs (KPI stands for Key Performance Indicators)	**Data warehousing and business intelligence**
1.Data characteristics in a database 2. Database components 3. Database architectures 4. Approaches to organizational data storage 5. Data storage standards 6. Data storage KPIs (KPI stands for Key Performance Indicators(	**Cloud data management**
1.Data model components 2. Data model granularity levels" 3. Factors influencing data model improvement 4. Data modeling standards 5. Data model KPIs (KPI stands for Key Performance Indicators)	**Data modeling**
1.Enterprise architecture framework 2. Organizational data architecture 3. Data architecture standards 4. Data architecture KPIs (KPI stands for Key Performance Indicators)	**Data architecture**

Table 13. Data quality Evaluation checklist

Data cascade					**Data management**		
Data quality	Data fitting	Trustworthy performance	Scalability solution	Performance testing	Data cleansing for structured data	Fairness	Data cleaning
					Consistency constraints	Understanding potential data interactions Artificial data analysis	Cleaning poisoned data/Cybersecurity /Self-control/Filtering spam

Table 14. Evaluation criteria

Artificial Intelligence Performance Evaluation Metrics			
Supervised learning		Unsupervised learning	Reinforcement Learning
Regression	Classification	Clustering	–
MAPE RMSE MAE RSquard (R^2)	Accuracy Precision ROC Curve Confusion Matrix	Jaccard Index Silhouette Coefficient Rand Index (RI) Davies-Bouldin Index	–

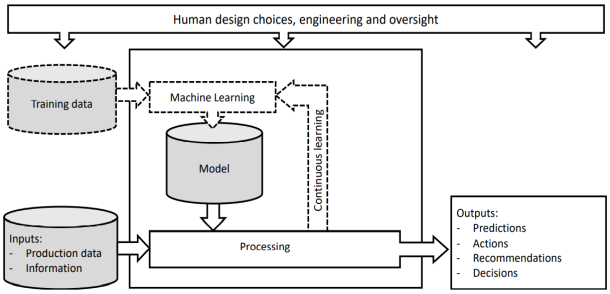

Fig 13. Examining the stages of AI software development

Table 15. AI reliability Evaluation in the life cycle

Trustworthy AI										
Establishment and monitoring	Modeling			Collect and understand data						
Trustworthy deployment Choose Trustworthy monitoring tools Evaluation the Trustworthy of the model	Trustworthy post-processing	Trustworthy training	trustworthy preprocessing	data privacy	crowdsourcing	augmenting data	unbiased data	quality control	Administrative and organizational data	Social network data
Border erosion	Unfair post-processing Adversarial attacks	Domain robustness Fairness-enhancing processing Adversaries Attack Explainability	Domain adaptation Unfair pre-processing Data cleansing Explainability		Unbiased Quality control Impartiality	collapse	Social unbias Representation unbias Data preparation unbias data non-poisoning Temporal unbias	Targeted data collection Obvious cheatin	Historical prejudices Label correction	Quality precision Reliability Folk accents, non-standard dialects, misspellings

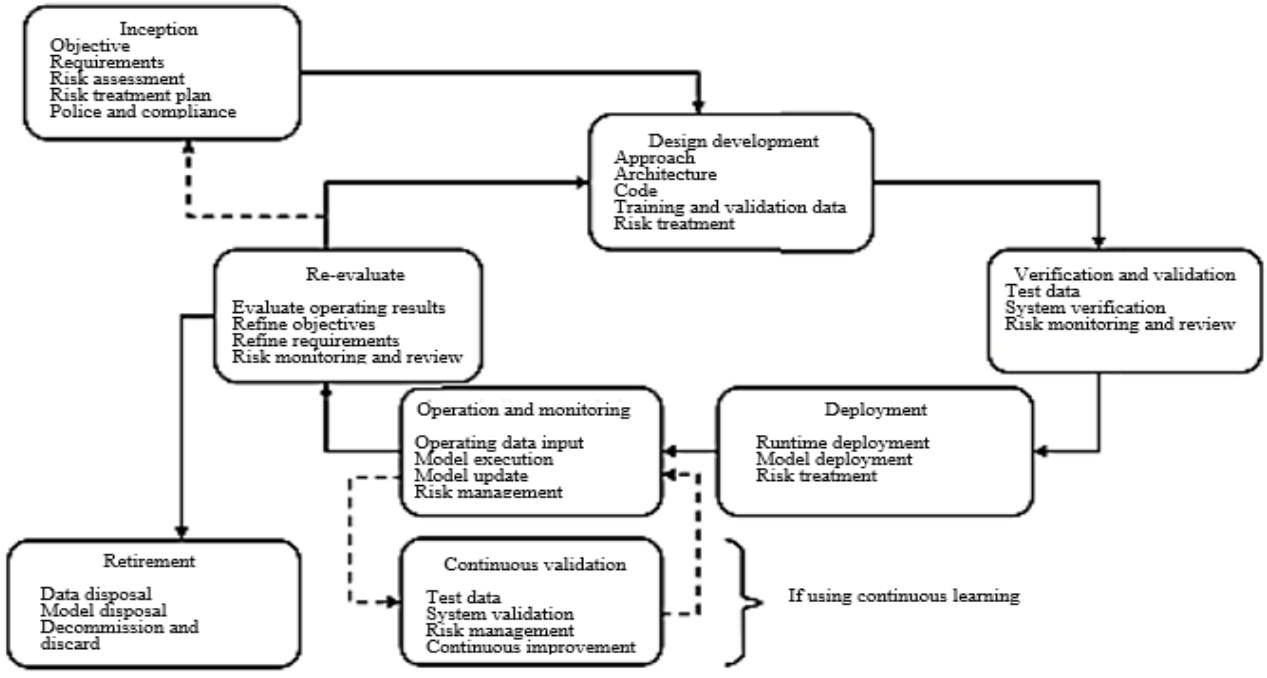

Fig 14. Details of Construction Stages of AI-Based Systems

Table 16. Life cycle-based system Evaluation

Trustworthy AI										
Establishment and monitoring	Modeling			Collect and understand data						
Trustworthy deployment Choose Trustworthy monitoring tools Evaluation the Trustworthy of the model	Trustworthy post-processing	Trustworthy training	trustworthy preprocessing	data privacy	crowdsourcing	augmenting data	unbiased data	quality control	Administrative and organizational data	Social network data
Border erosion	Unfair post-processing Adversarial attacks	Domain robustness Fairness-enhancing processing Adversaries Attack Explainability	Domain adaptation Unfair pre-processing Data cleansing Explainability		Unbiased Quality control Impartiality	collapse	Social unbias Representation unbias Data preparation unbias data non-poisoning Temporal unbias	Targeted data collection Obvious cheatin	Historical prejudices Label correction	Quality precision Reliability Folk accents, non-standard dialects, misspellings

Framework, IBM Data Governance Council Framework, SAS Data Governance Framework. These frameworks provide a comprehensive approach to managing data from creation to provision (throughout the data lifecycle), including data quality, data architecture, data modeling, cloud data management, data warehousing and business intelligence, data storage and operations, data security, master data management, data integration, content and documentation management. Based on this framework, the following checklist should be checked. Based on [13] and [31], the data governance Evaluation checklist and data quality Evaluation are compiled in Tables 12 and 13.

3.5. Evaluation Criteria in Machine Learning Algorithms

The evaluation of AI products involves functional [32]; [33] and non-functional [34] evaluation; in this section, the focus is on performance criteria of AI systems. The criteria that are obtained based on mathematical formulas are meant for measuring the quantitative behavior of a system. Table 14 shows the evaluation criteria in different AI issues.

4. Discussion section & Results

In this section, we present the final conclusion of this paper. Figure 13 shows that a majority of these evaluations is effective in the data section; the designed architecture model of machine learning and the results of these tests and evaluations can be seen in the outputs section.

Based on [16] and [324], Table 15 shows the checklist of reliability evaluation of AI-based system during its life cycle.

Considering the lack of a comprehensive standard covering all aspects of AI product production, we decided to present a roadmap for standardization by studying, collecting and categorizing standards in the field of AI product production. In Figure 16, all stages of life cycle, the steps of evaluation, validation and elimination of risks are presented. These steps are achieved by evaluating the quantitative and qualitative criteria mentioned in this paper.

5.Conclusion

The proliferation of AI products and the rapid adoption of algorithms in business and global society have created the need for regulatory frameworks and evaluation criteria to ensure that the public interest is maintained. Standards, checklists, and Evaluation criteria provide a structured approach to the development, evaluation, and monitoring of AI-based products. The use of standards, tests and checklists has become necessary to create public trust, accountability and operationalization of AI systems. This paper has been presented to guide researchers and practitioners in the field of evaluation and testing of AI systems. For the first time, this survey presented a comprehensive and complete review (by examining more than 200 papers and standards) on how to evaluate AI systems according to the categories of standards, checklists and methods. These categories have been offered according to the review of papers in terms of quantitative and qualitative evaluations during the development of AI products. As a future research direction, we may present specific and clear examples of how to use checklists, test and evaluations in critical industries and use cases.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations
Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: Feb. 2024 Accepted: Mar. 2024
Published online: Mar. 2024

DOI: 10.22034/ASAS.2024.450378.1055

Reference
1. F. Morandín-Ahuerma، ""What is Artificial Intelligence?."،" Journal homepage: www. ijrpr. com، ISSN 2582، p. 7421، 1948.
2. A. Dafoe، ""AI governance: a research agenda."،" Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK، 1442 ا، p. 1443، 2018.
3. M. Nadimpalli، ""Artificial intelligence risks and benefits."،" International Journal of Innovative Research in Science, Engineering and Technology، p. 6.6، 2017.
4. R. K. a. A. K. S. Behara، ""Artificial Intelligence Methodologies in Smart Grid-Integrated Doubly Fed Induction Generator Design Optimization and Reliability Evaluation: A Review. ،" Energies، 7164، 15.19، 2022.
5. J. e. Moor، "The Turing test: the elusive standard of artificial intelligence.،" Springer Science & Business Media، ا Vol. 30.، 2003.
6. A. I. M. a. I. L... Pīlēna، ""Standardization as a catalyst for open and responsible innovation."،" Journal of Open Innovation: Technology, Market, and Complexity، 7.3، p. 187، 2021.
7. R. M. e. a. Al Batayneh، "" "IT governance framework and smart services integration for future development of Dubai infrastructure utilizing AI and big data, its reflection on the citizens standard of living."،" Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021).Cham: Springer International Publishing. 2021..
8. A. Vanolo، ""Smartmentality: The smart city as disciplinary strategy."،" Urban studies، 51.5 ا، pp. 883-898.، (2014).
9. S. e. a. Soltani، ""Automated planning for feature model configuration based on functional and non-functional requirements."،" Proceedings of the 16th International Software Product Line Conference-Volume، 2012.
10. Y. e. a. Zhai، ""Tracing the evolution of AI: conceptualization of artificial intelligence in mass media discourse."،" Information discovery and delivery، 137-149، p. 48.3، 2020.
11. A. Panagariya، "" "Growth and Reforms during 1980s and 1990s."،" Economic and Political Weekl.22581-2594.، 2004.
12. B. Boehm، "" "A view of 20th and 21st century software engineering."،" Proceedings of the 28th international conference on Software engineering.، 2006.
13. K. H. Whang، "Data Cleaning for Accurate, Fair, and Robust Models: A Big Data - AI Integration Approach.،" Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning.، (2019).
14. "AIGA،" (2022). Available: Retrieved from https://ai-governance.eu/ai-governance-framework/the-ai-governance-lifecycle/.
15. K. Schweichhart، "" "Reference architectural model industrie 4.0 (rami 4.0)."،" An Introduction، 40، 2016.
16. K. R. Varshney، "Trustworthy Machine Learning.".
17. T. Hagendorff، "The ethics of AI ethics: An evaluation of guidelines،" Minds and machines، pp. 99-120، 2020.
18. A. A. A. M. BERRADA, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE, vol. 6, pp. 52138-52160, 2018.
19. L. Sweeney، "k-anonymity: A model for protecting privacy،" ، p. 557–570، 2002.
20. Z. L. V. S. a. V. S. T. Li، "Privacy for free: Communication-efficient learning with differential privacy using sketches،" arXiv Prepr. arXiv1911.00972، 2019.
21. D. K. J. G. a. M. V. A. Machanavajjhala، "" l-diversity: Privacy beyond k-anonymity،" 1، pp. 3-es، 2007.
22. C. G. S. H. a. V. V. M. van Dijk، "Fully homomorphic encryption over the integers،" in Annual international conference on the theory and applications of cryptographic techniques، pp. 24-43، 2010.
23. H. A. A. S. U. a. M. C. A. Acar، "A survey on homomorphic encryption schemes: Theory and implementation،" ACM Comput. Surv، vol. 21، p. 1–35، 2018.
24. L. S. a. A. M. P. Martins، "A survey on fully homomorphic encryption: An engineering perspective،" ACM Comput. Surv، vol. 50، pp. 1-33، 2017.
25. C. G. a. V. V. Z. Brakerski، "(Leveled) fully homomorphic encryption without bootstrapping،" ACM Trans. Comput. Theory، vol. 6، pp. 1-36، 2014.
26. M. Cannarsa، "Ethics Guidelines for Trustworthy AI.،" The Cambridge Handbook of Lawyering in the Digital Age.، 2021.
27. I. D. S. a. M. D... Alhassan، ""Data governance activities: an analysis of the literature."،" Journal of Decision Systems، vol. 25.sup1، (2016): 64-75.
28. M. E. Barbierato، "The Dual Role of Artificial Intelligence in Developing Smart Cities.،" Smart Cities، (2022).
29. J. R. Suryn، "On the Use of ISO/IEC Standards to Address Data Quality Aspects in Big Data Analytics Cloud Services.،"(2017).
30. Laware, Gilbert W. "STRATEGIC BUSINESS PLANNING Aligning Business Goals with Technology." INFORMATION SYSTEM MANAGEMENT 8.4 (1991): 44-49.
31. N. S. a. F. I. a. M. v. d. Schaar، "DC-Check: A Data-Centric

AI checklist to guide the development of reliable machine learning systems‹" ArXiv‹ p. abs/2211.05764‹ 2022.
32. a. A. P. J. Diana Sadykova‹ "Quality Evaluation Metrics for Edge Detection and Edge-aware Filtering: A Tutorial Review‹" arXiv:1801.00454v1‹ 2018.
33. P. P. S. 2. U. a. A. H. A. P. M.Z. Naser‹ "Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine 1‹" Architecture, Structures and Construction‹ 2021.
34. P. B. S. J. J. J. ,. B. Z. BOLI‹ "Trustworthy AI :From Principles to Practices‹" arXiv:2110.01167v2‹ 26 May 2022.
35. "IEC 62264-1:2013 Enterprise-control system integration - Part 1: Models and terminology‹" 2013-05-22.
36. "IEC 62264-2:2013 Enterprise-control system integration - Part 2: Object and attributes for enterprise-control system integration‹" 2013-06-27.
37. "IEC 62264-3:2016 Enterprise-control system integration - Part 3: Activity models of manufacturing operations management‹" 2016-12-16.
38. "IEC 62264-4:2015 Enterprise-control system integration - Part 4: Objects models attributes for manufacturing operations management integration‹" 2015-12-16.
39. "IEC 62264-5:2016 Enterprise-control system integration - Part 5: Business to manufacturing transactions‹" 2016-07-19.
40. "IEC 62264-6:2020 Enterprise-control system integration - Part 6: Messaging service model‹" 2020-06-26.
41." ISO/IEC 30141:2018 Internet of Things (IoT) — Reference Architecture‹" 2018-08-30.
42. "IEC 61499-1:2012 Function blocks - Part 1: Architecture‹" 2012-11-07.
43." IEC 61499-2:2012 Function blocks - Part 2: Software tool requirements‹" 2012-11-07.
44." IEC 61499-4:2013 Function blocks - Part 4: Rules for compliance profiles‹" 2013-01-30.
45. "IEC 61512-1:1997 Batch control - Part 1: Models and terminology‹" 1997-08-26.
46." IEC 61512-2:2001 Batch control - Part 2: Data structures and guidelines for languages‹" 2001-11-15.
47." IEC 61512-3:2008 Batch control - Part 3: General and site recipe models and representation‹" 2008-07-08.
48." IEC 61512-4:2009 Batch control - Part 4: Batch production records‹" 2009-10-13.
49." ISO/IEC/IEEE 42010:2022 Software, systems and enterprise - Architecture description‹" 2022-11-07.
50."ISO/IEC JTC 1/SC 41 Internet of things and digital twin‹" 2017.
51. "The JavaScript Object Notation (JSON) Data Interchange Format‹" December 2017.
52. "IEC 63339 ED1 Unified reference model for smart manufacturing‹" Nov 18, 2022.
53. "ISO 13584-42:2010 Industrial automation systems and integration — Parts library — Part 42: Description methodology: Methodology for structuring parts families‹" 2010-12.
54. "IEC 61360-6:2016 Standard data element types with associated classification scheme for electric components - Part 6: IEC Common Data Dictionary (IEC CDD) quality guidelines‹" 2016-10-04.
55. "IEC 61987-31:2022 Industrial-process measurement and control - Data structures and elements in process equipment catalogues - Part 31: List of Properties (LOPs) of infrastructure devices for electronic data exchange – Generic structures‹" 2022-12-14.
56. "ISO/TS 29002-5:2009 Industrial automation systems and integration — Exchange of characteristic data — Part 5: Identification scheme‹" 2
57. "IEC 62714-1:2018 Engineering data exchange format for use in industrial automation systems engineering - Automation Markup Language - Part 1: Architecture and general requirements‹" 2018-04-30.
58. "IEC 62714-2:2022 Engineering data exchange format for use in industrial automation systems engineering - Automation Markup Language - Part 2: Semantics libraries‹" 2022-10-20.
59. "IEC 62714-3:2017 Engineering data exchange format for use in industrial automation systems engineering - Automation markup language - Part 3: Geometry and kinematics‹" 2017-01-25.
60. "IEC 62714-4:2020 Engineering data exchange format for use in industrial automation systems engineering - Automation markup language - Part 4: Logic‹" 2020-06-16.
61. "IEC 62714-5:2022 Engineering data exchange format for use in industrial automation systems engineering - Automation markup language - Part 5: Communication‹" 2022-03-11.
62." ISO 13584-42:2010 Industrial automation systems and integration — Parts library — Part 42: Description methodology: Methodology for structuring parts families‹" 2010-12.
63. "IEC 61360-6:2016 Standard data element types with associated classification scheme for electric components - Part 6: IEC Common Data Dictionary (IEC CDD) quality guidelines‹" 2016-10-04.
64. "IEC 61987-1:2006 Industrial-process measurement and control - Data structures and elements in process equipment catalogues - Part 1: Measuring equipment with analogue and digital output‹" 2006-12-14.
65." IEC 62714-1:2018 RLV Redline version Engineering data exchange format for use in industrial automation systems engineering - Automation Markup Language - Part 1: Architecture and general requirements‹" 2018-04-30.
66. "IEC 62714-2:2022 Engineering data exchange format for use in industrial automation systems engineering - Automation Markup Language - Part 2: Semantics libraries‹" 2022-10-20.
67. "IEC 62714-3:2017 Engineering data exchange format for use in industrial automation systems engineering - Automation markup language - Part 3: Geometry and kinematics‹" 2017-01-25.
68. "IEC 62714-4:2020 Engineering data exchange format for

use in industrial automation systems engineering - Automation markup language - Part 4: Logic‹" 2020-06-16.
69." IEC 62714-5:2022 Engineering data exchange format for use in industrial automation systems engineering - Automation markup language - Part 5: Communication‹" 2022-03-11.
70." ISO/IEC 19788-7:2019 Information technology - Learning, education and training - Metadata for learning resources - Part 7: Bindings‹" 2019-04-10.
71. "ISO 16684-1:2019 Graphic technology — Extensible metadata platform (XMP) — Part 1: Data model, serialization and core properties‹" 2019-04.
72." ISO/IEC 15944-10:2013 Information technology -- Business Operational View -- Part 10: IT-enabled coded domains as semantic components in business transactions‹" 2013-02-12.
73. "ISO 1087:2019 Terminology work and terminology science — Vocabulary‹" 2019-09.
74. "ISO/IEC 11179-5:2015 Information technology — Metadata registries (MDR) — Part 5: Naming principles‹" 2015-04.
75." ISO/IEC 11179-1:2023 Information technology — Metadata registries (MDR) — Part 1: Framework‹" 2023-01.
76." ISO 21597-1:2020 Information container for linked document delivery — Exchange specification — Part 1: Container‹" 2020-04.
77. "IEC 61158-1:2022 PRV Pre release version Industrial communication networks - Fieldbus specifications - Part 1: Overview and guidance for the IEC 61158 and IEC 61784 series‹" 2022-12-30.
78. "IEC 61784-2-19:2022 PRV Pre release version Industrial networks - Profiles - Part 2-19: Additional real-time fieldbus profiles based on ISO/IEC/IEEE 8802-3 – CPF 19‹" 2022-12-30.
79. "ISO/TC 184 Automation systems and integration‹" 2023.
80." ISO/TC 184 - IEC/TC 65/JWG 21 - Smart Manufacturing Reference Model(s‹" (2023-01-27.
81. "IEC 62591:2016 Industrial networks - Wireless communication network and communication profiles - WirelessHARTTM‹" 2016-03-30.
82." IEC 62601:2015 Industrial networks - Wireless communication network and communication profiles - WIA-PA‹" 2015-12-09.
83." IEC 62734:2014+AMD1:2019 CSV Consolidated version Industrial networks - Wireless communication network and communication profiles - ISA 100.11a‹" 2019-07-31.
84." IEC 62948:2017 Industrial networks - Wireless communication network and communication profiles - WIA-FA‹" 2017-07-27.
85." IEC 62657-2:2022 Industrial networks - Coexistence of wireless systems - Part 2: Coexistence management‹" 2022-06-09.
86." IEC TR 62541-1:2020 RLV Redline version OPC unified architecture - Part 1: Overview and concepts‹" 2020-11-18.
87." IEC TR 62541-2:2020 RLV Redline version OPC unified architecture - Part 2: Security Model‹" 2020-11-17.
88." IEC 62541-3:2020 RLV Redline version OPC Unified Architecture - Part 3: Address Space Model‹" 2020-07-08.
89." IEC 62541-4:2020 RLV Redline version OPC Unified Architecture - Part 4: Services‹" 2020-07-13.
90." IEC 62541-5:2020 RLV Redline version OPC Unified Architecture - Part 5: Information Model‹" 2020-07-10.
91." IEC 62541-6:2020 RLV Redline version OPC Unified Architecture - Part 6: Mappings‹" 2020-07-13.
92." IEC 62541-7:2020 RLV Redline version OPC unified architecture - Part 7: Profiles‹" 2020-06-22.
93. "IEC 62541-8:2020 RLV Redline version OPC Unified Architecture - Part 8: Data Access‹" 2020-06-22.
94." IEC 62541-9:2020 RLV Redline version OPC Unified Architecture - Part 9: Alarms and Conditions‹" 2020-06-18.
95." IEC 62541-10:2020 RLV Redline version OPC Unified Architecture - Part 10: Programs‹" 2020-07-07.
96." IEC 62541-11:2020 RLV Redline version OPC Unified Architecture - Part 11: Historical Access‹" 2020-06-23.
97." IEC 62541-12:2020 OPC Unified Architecture - Part 12: Discovery and global services‹" 2020-06-16.
98." IEC 62541-13:2020 RLV Redline version OPC Unified Architecture - Part 13: Aggregates‹" 2020-06-11.
99." IEC 62541-14:2020 OPC Unified Architecture - Part 14: PubSub‹" 2020-07-08.
100." IEC 62541-100:2015 OPC Unified Architecture - Part 100: Device Interface‹" 2015-03-25.
101. "IEC 61784-1-22:2022 PRV Pre release version Industrial networks - Profiles - Part 1-22: Fieldbus profiles - Communication Profile Family 22‹" 2022-12-30.
102. "IEC 61158-1:2022 PRV Pre release version Industrial communication networks - Fieldbus specifications - Part 1: Overview and guidance for the IEC 61158 and IEC 61784 series‹" 2022-12-30.
103. "IEC 61158-2:2022 PRV Pre release version Industrial communication networks - Fieldbus specifications - Part 2: Physical layer specification and service definition‹" 2022-12-30.
104. "IEC 61158-3-1:2014 Industrial communication networks - Fieldbus specifications - Part 3-1: Data-link layer service definition - Type 1 elements‹" 2014-08-13.
105. "IEC 61158-3-2:2014+AMD1:2019 CSV Consolidated version Industrial communication networks - Fieldbus specifications - Part 3-2: Data-link layer service definition - Type 2 elements‹" 2019-04-18.
106." IEC 61158-3-3:2014 Industrial communication networks - Fieldbus specifications - Part 3-3: Data-link layer service definition - Type 3 elements‹" 2014-08-13.
107. "IEC 61158-3-4:2019 Industrial communication networks - Fieldbus specifications - Part 3-4: Data-link layer service definition - Type 4 elements‹" 2019-04-24.
108. "IEC 61158-3-7:2007 Industrial communication networks - Fieldbus specifications - Part 3-7: Data-link layer service definition - Type 7 elements‹" 2007-12-14.
109." IEC 61158-3-8:2007 Industrial communication networks - Fieldbus specifications - Part 3-8: Data-link layer ser-

vice definition - Type 8 elements‹" 2007-12-14.
110." IEC 61158-3-11:2007 Industrial communication networks - Fieldbus specifications - Part 3-11: Data-link layer service definition - Type 11‹" 2007-12-14.
111. "IEC 61158-3-12:2019 Industrial communication networks - Fieldbus specifications - Part 3-12: Data-link layer service definition - Type 12‹" 2019-04-24.
112." IEC 61158-3-13:2014 Industrial communication networks - Fieldbus specifications - Part 3-13: Data link layer service definition - Type 13 elements‹" 2014-08-13.
113. "IEC 61158-3-14:2014 Industrial communication networks - Fieldbus specifications - Part 3-14: Data-link layer service definition - Type 14 elements‹" 2014-08-13.
114." IEC 61158-3-16:2007 Industrial communication networks - Fieldbus specifications - Part 3-16: Data-link layer service definition - Type 16 elements‹" 2007-12-14.
115." IEC 61158-3-17:2007 Industrial communication networks - Fieldbus specifications - Part 3-17: Data-link layer service definition - Type 17 elements‹" 2022-12-30.
116. "IEC 61158-3-18:2007 Industrial communication networks - Fieldbus specifications - Part 3-18: Data-link layer service definition - Type 18 elements‹" 2007-12-14.
117. "IEC 61158-3-19:2019 Industrial communication networks - Fieldbus specifications - Part 3-19: Data-link layer service definition - Type 19 elements‹" 2019-04-24.
118." IEC 61158-3-20:2014 Industrial communication networks - Fieldbus specifications - Part 3-20: Data-link layer service definition - Type 20 elements‹" 2014-08-13.
119." IEC 61158-3-21:2019 Industrial communication networks - Fieldbus specifications - Part 3-21: Data-link layer service definition - Type 21 elements‹" 2019-04-24.
120. "IEC 61158-3-22:2014 Industrial communication networks - Fieldbus specifications - Part 3-22: Data-link layer service definition - Type 22 elements‹" 2014-08-13.
121." IEC 61158-3-24:2014 Industrial communication networks - Fieldbus specifications - Part 3-24: Data-link layer service definition - Type 24 elements‹" 2014-08-13.
122. "IEC 61158-3-25:2019 Industrial communication networks - Fieldbus specifications - Part 3-25: Data-link layer service definition - Type X elements‹" 2019-04-10.
123." IEC 61158-4-1:2014 Industrial communication networks - Fieldbus specifications - Part 4-1: Data-link layer protocol specification - Type 1 elements‹" 2014-08-15.
124." IEC 61158-4-2:2019 Industrial communication networks - Fieldbus specifications - Part 4-2: Data-link layer protocol specification - Type 2 elements‹" 2019-04-18.
125. "IEC 61158-4-3:2019 Industrial communication networks - Fieldbus specifications - Part 4-3: Data-link layer protocol specification - Type 3 elements‹" 2019-04-18.
126." IEC 61158-4-4:2019 Industrial communication networks - Fieldbus specifications - Part 4-4: Data-link layer protocol specification - Type 4 elements‹" 2019-04-18.
127." IEC 61158-4-7:2007 Industrial communication networks - Fieldbus specifications - Part 4-7: Data-link layer protocol specification - Type 7 elements‹" 2007-12-14.
128. "IEC 61158-4-8:2007 Industrial communication networks - Fieldbus specifications - Part 4-8: Data-link layer protocol specification - Type 8 elements‹" 2007-12-14.
129." IEC 61158-4-11:2014 Industrial communication networks - Fieldbus specifications - Part 4-11: Data-link layer protocol specificationn - Type 11 elements‹" 2014-08-15.
130. "IEC 61158-4-12:2019 Industrial communication networks - Fieldbus specifications - Part 4-12: Data-link layer protocol specification - Type 12 elements‹" 2019-04-18.
131. "IEC 61158-4-13:2014 Industrial communication networks - Fieldbus specifications - Part 4-13: Data-link layer protocol specificationn - Type 13 elements‹" 2014-08-15.
132." IEC 61158-4-14:2014 Industrial communication networks - Fieldbus specifications - Part 4-14: Data-link layer protocol specificationn - Type 14 elements‹" 2014-08-15.
133." IEC 61158-4-16:2007 Industrial communication networks - Fieldbus specifications - Part 4-16: Data-link layer protocol specification - Type 16 elements‹" 2007-12-14.
134." IEC 61158-4-17:2007 Industrial communication networks - Fieldbus specifications - Part 4-17: Data-link layer protocol specification - Type 17 elements‹" 2007-12-14.
135. "IEC 61158-4-18:2010 Industrial communication networks - Fieldbus specifications - Part 4-18: Data-link layer protocol specification - Type 18 elements‹" 2010-08-05.
136." IEC 61158-4-19:2019 Industrial communication networks - Fieldbus specifications - Part 4-19: Data-link layer protocol specification - Type 19 elements‹" 2019-04-18.
137. "IEC 61158-4-20:2014 Industrial communication networks - Fieldbus specifications - Part 4-20: Data-link layer protocol specification - Type 20 elements‹" 2014-08-15.
138." IEC 61158-4-21:2019 Industrial communication networks - Fieldbus specifications - Part 4-21: Data-link layer protocol specification - Type 21 elements‹" 2019-04-10.
139." IEC 61158-4-22:2014 Industrial communication networks - Fieldbus specifications - Part 4-22: Data-link layer protocol specificationn - Type 22 elements‹" 2014-08-15.
140." IEC 61784-2-19:2022 PRV Pre release version Industrial networks - Profiles - Part 2-19: Additional real-time fieldbus profiles based on ISO/IEC/IEEE 8802-3 – CPF 19‹" 2022-12-30.
141." IEC 61784-3:2021 Industrial communication networks - Profiles - Part 3: Functional safety fieldbuses - General rules and profile definitions‹" 2021-02-16.
142." IEC 61784-5-1:2013 Industrial communication networks - Profiles - Part 5-1: Installation of fieldbuses - Installation profiles for CPF 1‹" 2013-09-11.
143." IEC 61918:2018+AMD1:2022 CSV Consolidated version Industrial communication networks - Installation of communication networks in industrial premises‹" 2022-03-09.
144." IEC TS 61850-1-2:2020+AMD1:2022 CSV Consolidated version Communication networks and systems for power utility automation - Part 1-2: Guideline on extending IEC

61850‹" 2022-07-11.
145." ANSI/MTC1.4-2018 - MTConnect‹" 12-7-2018.
146." ISO/IEC 19464:2014 Information technology — Advanced Message Queuing Protocol (AMQP) v1.0 specification‹" 2014-05.
147." IEEE FOUNDATION FOR INTELLIGENT PHYSICAL AGENTS (FIPA"(2012-¬-01-¬-09.
148." IEC 61850-8-1:2011+AMD1:2020 CSV Consolidated version Communication networks and systems for power utility automation - Part 8-1: Specific communication service mapping (SCSM) - Mappings to MMS (ISO 9506-1 and ISO 9506-2) and to ISO/IEC 8802-3‹" 2020-02-21.
149." ISO/IEC 19464:2014 Information technology -- Advanced Message Queuing Protocol (AMQP) v1.0 specification‹" 2014-04-29.
150." ISO/IEC 18000-7:2014 Information technology — Radio frequency identification for item management — Part 7: Parameters for active air interface communications at 433 MHz‹" 2014-09.
151." IEEE Standard for Information technology-- Local and metropolitan area networks-- Specific requirements-- Part 15.1a: Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications for Wireless Personal Area Networks (WPAN"(2003-09-11.
152." TM IEEE Standard for Information technology— Telecommunications and information exchange between systems— Local and metropolitan area networks— Specific requirements Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specific‹" 12 May 2003.
153." IEEE 802.11 – Wireless Local Networks‹" september 1997.
154." ISO 11354-1:2011 Advanced automation technologies and their applications — Requirements for establishing manufacturing enterprise process interoperability — Part 1: Framework for enterprise interoperability‹" 2011-09.
155." IEC/DIS 63339 Unified reference model for smart manufacturing".
156. "ISO 15745-1:2003 Industrial automation systems and integration — Open systems application integration framework — Part 1: Generic reference description‹" 2003-03.
157." IEC 62264-1:2013 Enterprise-control system integration - Part 1: Models and terminology‹" 2013-05-22.
158." IEC 62264-2:2013 Enterprise-control system integration - Part 2: Object and attributes for enterprise-control system integration‹" 2013-06-27.
159." IEC 62264-3:2016 Enterprise-control system integration - Part 3: Activity models of manufacturing operations management‹" 2016-12-16.
160." IEC 62264-4:2015 Enterprise-control system integration - Part 4: Objects models attributes for manufacturing operations management integration‹" 2015-12-16.
161." IEC 62264-5:2016 Enterprise-control system integration - Part 5: Business to manufacturing transactions‹" 2016-07-19.
162." IEC 62264-6:2020 Enterprise-control system integration - Part 6: Messaging service model‹" 2020-06-26.
163." ISO 20140-5:2017 Automation systems and integration - Evaluating energy efficiency and other factors of manufacturing systems that influence the environment - Part 5: Environmental performance evaluation data‹" 2017-04-19.
164. " IEEE Recommended Practice for Industrial Agents: Integration of Software Agents and Low-Level Automation Functions‹" IEEE Std 2660.1-2020 doi: 10.1109/IEEESTD.2021.9340089.‹ vol. no.‹ pp..1-43‹ 29 January 2021.
165." Agentensysteme in der Automatisierungstechnik - Entwicklung‹" 2018-02.
166." IEC 62832-1:2020 Industrial-process measurement, control and automation - Digital factory framework - Part 1: General principles‹" 2020-10-26.
167." IEC 61360-1:2017 Standard data element types with associated classification scheme - Part 1: Definitions - Principles and methods‹" 2017-07-27.
168." IEC 61360-2:2012 Standard data element types with associated classification scheme for electric components - Part 2: EXPRESS dictionary schema‹" 2012-10-02.
169." IEC 61360-6:2016 Standard data element types with associated classification scheme for electric components - Part 6: IEC Common Data Dictionary (IEC CDD) quality guidelines‹" 2016-10-04.
170." IEC 61131:2023 SER Series Programmable controllers - ALL PARTS‹" 2023-01-03.
171. "IEC 62832-1:2020 Industrial-process measurement, control and automation - Digital factory framework - Part 1: General principles‹" 2020-10-26.
172." IEC 62832-2:2020 Industrial-process measurement, control and automation - Digital factory framework - Part 2: Model elements‹" 2020-10-26.
173. "IEC 62832-3:2020 Industrial-process measurement, control and automation - Digital factory framework - Part 3: Application of Digital Factory for life cycle management of production systems‹" 2020-10-27.
174. "IEC 61508:2010 CMV Commented version Functional safety of electrical/electronic/programmable electronic safety-related systems - Parts 1 to 7‹" 2010-04-30.
175." IEC TS 62443-1-1:2009 Industrial communication networks - Network and system security - Part 1-1: Terminology, concepts and models‹" 2009-07-30.
176." IEC 62443-2-1:2010 Industrial communication networks - Network and system security - Part 2-1: Establishing an industrial automation and control system security program‹" 2010-11-10.
177." IEC TR 62443-3-1:2009 Industrial communication networks - Network and system security - Part 3-1: Security technologies for industrial automation and control systems‹" 2009-07-30.
178." IEC 62443-4-1:2018 Security for industrial automation

and control systems - Part 4-1: Secure product development lifecycle requirements‹" 2018-01-15.
179." IEC 61131:2023 SER Series Programmable controllers - ALL PARTS‹" 2023-01-03.
180." ISO/IEC TR 24027:2021 Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making‹" 2021-11-05.
181. "ISO/IEC TR 24028:2020 Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence‹" 2020-05-28.
182." ISO/IEC DIS 5338 Information technology — Artificial intelligence — AI system life cycle processes".
183. " ISO/IEC TR 24368:2022 Information technology - Artificial intelligence - Overview of ethical and societal concerns‹" 2022-08-19.
184. " ISO/IEC TS 4213:2022 Information technology - Artificial intelligence - Evaluation of machine learning classification performance‹" 2022-10-13.
185. " IEEE P7000 Working Group Meeting Draft Minutes‹" 14 December 2018.
186. "IEEE P2863 Organizational Governance of Artificial Intelligence Working Group‹" January 6, 2022.
187." IEEE 3652.1-2020 IEEE Guide for Architectural Framework and Application of Federated Machine Learning‹" 2021-03-19.
188." ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management‹" 2023-02.
189. "ISO/IEC DIS 42001 Information technology — Artificial intelligence — Management system".
190." IEEE P7001: A Proposed Standard on Transparency‹" 26 July 2021.
191." IEEE Invites Stakeholders Globally to Expand on Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) Work‹" March 6, 2020.
192. "ISO/PRF 4669-1 Document management — Information classification, marking and handling — Part 1: Requirements".
193." ISO/IEC CD TR 5469 Artificial intelligence — Functional safety and AI systems".
194. " ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) - Evaluation of the robustness of neural networks - Part 1: Overview‹" 2021-03-10.
195. "ISO/IEC 24668:2022 Information technology - Artificial intelligence - Process management framework for big data analytics‹" 2022-11-17.
196. "ETSI GR SAI 006 Securing Artificial Intelligence (SAI‹"(2022-03.
197." IEC TS 62443-1-1:2009 Industrial communication networks - Network and system security - Part 1-1: Terminology, concepts and models‹" 2009-07-30.
198." IEC 62443-2-1:2010 Industrial communication networks - Network and system security - Part 2-1: Establishing an industrial automation and control system security program‹" 2010-11-10.
199. "IEC TR 62443-3-1:2009 Industrial communication networks - Network and system security - Part 3-1: Security technologies for industrial automation and control systems‹" 2009-07-30.
200." IEC 62443-4-1:2018 Security for industrial automation and control systems - Part 4-1: Secure product development lifecycle requirements‹" 2018-01-15.
201. "ISO/IEC DIS 25059 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems".
202. "ISO/IEC AWI 24029-2 Artificial intelligence (AI) - Evaluation of the robustness of neural networks - Part 2: Methodology for the use of formal methods."
203."IEEE - P7009 - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems‹" 2017-06-15.
204."IEEE P2807 Knowledge Graph Working Group‹" 2022-09-21.
205. "IIEEE P2846: Assumptions for Models in Safety-Related Automated Vehicle Behavior‹" June 2020.
206." ISO 26262-1:2018 Road vehicles — Functional safety — Part 1: Vocabulary‹" 2018-12.
207. "ISO/PAS 21448:2019 Road vehicles — Safety of the intended functionality‹" 2019-01.
208. P. a. M. W. Koopman‹ "Toward a framework for highly automated vehicle safety validation.‹" SAE Technical Paper, Tech. Rep‹ 2018.
209. P. a. F. F. Koopman‹ "How many operational design domains, objects, and events?‹" Safeai@ aaai 4‹ 2019.
210. "J. Inge,Defence Standard 00-56 Issue 4: Safety Management Requirements for Defence Systems‹" 2007.
211." GUIDELINES AND METHODS FOR CONDUCTING THE SAFETY EVALUATION PROCESS ON CIVIL AIRBORNE SYSTEMS AND EQUIPMENT ARP4761‹" 1996-12-01.
212. "US Dept. of Commerce,‹" 7 June 2019.
213."MIL-STD-882E, DEPARTMENT OF DEFENSE STANDARD PRACTICE: SYSTEM SAFETY‹" 11-MAY-2012.
214. "AC 23.1309-1E - System Safety Analysis and Evaluation for Part 23 Airplanes‹" November 17, 2011.
215."ISO 20126:2012 Dentistry — Manual toothbrushes — General requirements and test methods‹" 2012-01.
216." ISO 13849-1:2015 Safety of machinery — Safety-related parts of control systems — Part 1: General principles for design‹" 2015-12.
217. "IEC 61511:2023 SER Series Functional safety - Safety instrumented systems for the process industry sector - ALL PARTS‹" 2023-01-03
218. " UL 4600, Standard for Safety for the Evaluation of Autonomous Products)‹" March 15, 2022)
219. "VDE-AR-E 2842-61-1 Anwendungsregel:2021-07 Development and trustworthiness of autonomous/cognitive systems‹" 2021-07.
220." ISO/IEC 20547-3:2020 Information technology — Big

data reference architecture — Part 3: Reference architecture‹" 2020-03.
221." ISO/IEC AWI TS 6254 Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems".
222. "ISO/IEC AWI TS 25058 Software and systems engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of AI systems".
223." ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology‹" 2022-07.
224. " ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML‹"(2020-03.
225. "ISO/IEC TR 24372:2021 Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems‹" 2021-12.
226." ISO/IEC DIS 5392 Information technology — Artificial intelligence — Reference architecture of knowledge engineering".
227." ISO/IEC DIS 8183 Information technology — Artificial intelligence — Data life cycle framework".
228."ISO/IEC JTC 1/SC 42(AI)/WG 2(Data) Data Quality for Analytics and Machine Learning (ML))‹" May 24, 2022.(
229. "ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations‹" 2022-04.
230. "ISO/IEC CD 5259-4 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework".
231. "ISO/IEC CD 5259-1 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples".
232. "ISO/IEC AWI 5259-2 Data quality for analytics and ML - Part 2: Part 2: Data quality measures".
233."ISO/IEC CD 5259-3 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelin es".
234. "ISO/IEC AWI 5259-4 Data quality for analytics and ML - Part 4: Data quality process framework".
235. "ISO/IEC AWI 5259-5 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance".
236. "ISO/IEC TR 20547-1:2020 Information technology — Big data reference architecture — Part 1: Framework and application process‹" 2020-08-20.
237. "ISO/IEC TR 20547-2:2018 Information technology - Big data reference architecture - Part 2: Use cases and derived requirements‹" 2018-01-10.
238. "ISO/IEC 20547-3:2020 Information technology - Big data reference architecture - Part 3: Reference architecture‹" 2020-03-04.
239. C. M. A. S. L. Olah‹ "Feature visualization.‹" Distill 2(11), e7‹ 2017.
240. Oveisi, S., & Ravanmehr, R. (2017). Analysis of software safety and reliability methods in cyber physical systems. International journal of critical infrastructures, 13(1), 1-15.
241. Oveisi, S., Moeini, A., Mirzaei, S., & Farsi, M. A. (2023). Software reliability prediction: A survey. Quality and Reliability Engineering International, 39(1), 412-453.
242. Singh, A., & Kaur, N. (2020). Towards Transparent and Explainable Artificial Intelligence (AI). International Journal of Advanced Computer Science and Applications, 11(6), 83-89.
243. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
244. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36-43.
245. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
246. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 93
247. Ignatiev, A., & Hoffmann, J. (2020). Transparency in Artificial Intelligence: A Survey. arXiv preprint arXiv:2010.09831.
248. Weller, A., & Barria-Pineda, J. (2019). Transparent and explainable artificial intelligence: The role of accuracy. Journal of Business Research, 98, 365-376.
249. Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., ... & Muller, U. (2018). Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911.
250. Kim, M., & Park, C. (2020). Transparency in the Age of Artificial Intelligence and Big Data: Challenges and Opportunities. Journal of Open Innovation: Technology, Market, and Complexity, 6(4), 134.
251. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2019). The ethics of algorithms: Mapping the debate. Big Data & Society, 6(2), 2053951719848483.
252. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144)
253. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. Proceedings of the IEEE Symposium on Security and Privacy, 39-57.
254. Li, F., & Li, Y. (2020). Enhancing adversarial robustness via feature disentanglement with orthogonality regularization. Pattern Recognition, 97, 107030.
255. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A.

(2018). Towards deep learning models resistant to adversarial attacks. Proceedings of the International Conference on Learning Representations.
256. Pang, T., Du, X., Li, Y., & Sun, X. (2020). Enhancing the robustness of deep neural networks via adversarial training with Triplet loss. Neurocomputing, 383, 191-200.
257. Ramkumar, S. D., Ganesh, K. V., & Arjunan, M. R. (2019). A Survey of Artificial Intelligence in Privacy and Security. International Journal of Computer Applications, 182(29), 9-13. doi: 10.5120/ijca2019918993
258. Alawadhi, I., Baskaran, B., Vaidya, J., & Adam, N. (2021). Privacy and Security Issues in Artificial Intelligence: A Systematic Review. Journal of Big Data, 8(1), 1-39. doi: 10.1186/s40537-021-00441-5
259. Rahnavard, N., & Kangavari, M. R. (2020). Privacy and Security Issues in Machine Learning: A Survey. Journal of Information Security and Applications, 50, 102423. doi: 10.1016/j.jisa.2019.102423
260. European Committee for Electrotechnical Standardization. (2004). Railway applications - Insulation coordination, Part 1: Basic principles, requirements and tests (EN 20126). Retrieved from https://www.en-standard.eu/
261. European Committee for Electrotechnical Standardization. (2021). Safety rules for the construction and installation of lifts - Examination and tests - Part 3: Passenger and goods passenger lifts (EN 81-20/PrA2). Retrieved from https://www.en-standard.eu/
262. International Organization for Standardization. (2018). Road vehicles -- Functional safety (ISO 26262). Retrieved from https://www.iso.org/standard/68383.html
263. International Organization for Standardization. (2015). Safety of machinery -- Safety-related parts of control systems -- Part 1: General principles for design (ISO 13849-1). Retrieved from https://www.iso.org/standard/61427.html
264. International Electrotechnical Commission. (2016). Functional safety -- Safety instrumented systems for the process industry sector (IEC 61511). Retrieved from https://www.iec.ch/iec-61511/
265. International Organization for Standardization. (2019). Road vehicles -- Safety of the intended functionality (SOTIF) (ISO/PAS 21448). Retrieved from https://www.iso.org/standard/72458.html
266. Underwriters Laboratories Inc. (2020). Outline of investigation for autonomous vehicle safety (UL 4600). Retrieved from https://standardscatalog.ul.com/ProductDetail.aspx?productId=UL4600
267. International Organization for Standardization, & International Electrotechnical Commission. (2019). Information technology -- Artificial intelligence -- Overview of trustworthiness (ISO/IEC AWI TR 5469). Retrieved from https://www.iso.org/standard/69056.html
268. Association for Electrical, Electronic & Information Technologies (VDE). (2018). Safety of machinery - Part 61: Functional safety - Development of software for safety related machine control systems (VDE-AR-E 2842-61-1). Retrieved from https://www.vde-verlag.de/iec/dokumente/1916675/safety-of-machinery-part-61-functional-safety-development-of-software-for-safety-related-machine-control-systems-vde-ar-e-2842-61-1-2018-01.html
269. International Electrotechnical Commission. (2010). Functional safety of electrical/electronic/programmable electronic safety-related systems (IEC 61508). Retrieved from https://www.iec.ch/iec-61508/
270. International Organization for Standardization, & International Electrotechnical Commission. (2019). Information technology -- Security techniques -- Governance of information security (ISO/IEC 27000). Retrieved from https://www.iso.org/standard/54534.html
271. IEEE Standards Association. (2019). Standard for safety and test requirements for autonomous vehicles (IEEE P2846). Retrieved from https://standards.ieee.org/project/2846.html
272. International Organization for Standardization. (n.d.). ISO/DIS 26262-3: Road vehicles -- Functional safety -- Part 3: Concept phase (ISO/DIS 26262-3). Retrieved from https://www.iso.org/standard/71294.html
273. US Department of Defense. (2012). Standard practice: System safety (MIL-STD-882E). Retrieved from https://assist.dla.mil/quicksearch/basic_profile.cfm?ident_number=36008
274. US Federal Aviation Administration. (2000). Advisory Circular 23.1309-1: System safety analysis and Evaluation for part 23 airplanes (AC 23.1309-1). Retrieved from https://www.faa.gov/documentLibrary/media/Advisory_Circular/150_5300_13A_Chg1_AC_23_1309_1_System_Safety_Analysis_Evaluation_for_Part_23_Airplanes.pdf
275. Koopman, P., & Wagner, M. (2018). Toward a framework for highly automated vehicle safety validation. In Proceedings of the SAE 2018 World Congress & Exhibition (pp. 1-12). SAE International.
276. Koopman, P., & Fratrik, F. (2019). How many operational design domains, objects, and events? In Proceedings of the SafeAI Conference (pp. 1-20). IEEE.
277. Ministry of Defence. (2017). Safety management requirements for defence systems (Def Stan 00-56). Retrieved from https://www.gov.uk/government/publications/safety-management-requirements-for-defence-systems-def-stan-00-56
278. SAE International. (2012). Guidelines and methods for conducting the safety Evaluation process on civil airborne systems and equipment (ARP4761). Retrieved from https://www.sae.org/standards/content/arp4761/
279. US Department of Commerce. (2019, June 7). Regulatory reform. Retrieved from https://www.commerce.gov/issues/regulatory-reform
280. I.S.: Road Vehicles — Functional Safety — Part 1: Vocab-

ulary, vol. 1, 2 edn. (2018)
281. I.S.: Road Vehicles — Functional Safety — Part 6: Product Development at the Software Level, vol. 6, 2 edn. (2018)
282. Akametalu, A.K., Fisac, J.F., Gillula, J.H., Kaynama, S., Zeilinger, M.N., Tomlin, C.J.: Reachability-based safe learning with gaussian processes. In: Proc. 53rd IEEE Conf. Decision and Control, pp. 1424–1431 (2014)
283. Amodei, D., Olah, C., Steinhardt, J., Christiano, P.F., Schulman, J., Mané, D.: Concrete problems in AI safety. CoRR abs/1606.06565 (2016)
284. Bagschik, G., Menzel, T., Maurer, M.: Ontology based scene creation for the development of automated vehicles. In: Proc. 2018 IEEE Intelligent Vehicles Symp., pp. 1813–1820 (2018)
285. Bunel, R.R., Turkaslan, I., Torr, P., Kohli, P., Mudigonda, P.K.: A unified view of piecewise linear neural network verification. In: Advances in Neural Information Processing Systems 31, pp. 4790–4799 (2018)
286. Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to explain: An information-theoretic perspective on model interpretation. In: Proc. 35th Int. Conf. Machine Learning, vol. 80, pp. 883–892 (2018)
287. Cheng, C.H., Nührenberg, G., Huang, C.H., Ruess, H., Yasuoka, H.: Towards dependability metrics for neural networks. In: 16th ACM/IEEE Int. Conf. Formal Methods and Models for System Design, pp. 43–46 (2018)
288. Dreossi, T., Ghosh, S., Yue, X., Keutzer, K., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: Counterexample-guided data augmentation. In: Proc. 27th Int. Joint Conf. Artificial Intelligence, pp. 2071–2078 (2018)
289. Dutta, S., Jha, S., Sankaranarayanan, S., Tiwari, A.: Output range analysis for deep feedforward neural networks. In: Proc. 10th Int. Symp. NASA Formal Methods, vol. 10811, pp. 121–138 (2018)
290. Fong, R., Vedaldi, A.: Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition, pp. 8730–8738 (2018)
291. Fridovich-Keil, D., Herbert, S.L., Fisac, J.F., Deglurkar, S., Tomlin, C.J.: Planning, fast and slow: A framework for adaptive real-time safe trajectory planning. In: Proc. 2018 IEEE Int. Conf. Robotics and Automation, pp. 387–394 (2018)
292. Fuchs, F.B., Groth, O., Kosiorek, A.R., Bewley, A., Wulfmeier, M., Vedaldi, A., Posner, I.: Neural Stethoscopes: Unifying analytic, auxiliary and adversarial network probing. CoRR abs/1806.05502 (2018)
293. Gast, J., Roth, S.: Lightweight probabilistic deep networks. In: Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition, pp. 3369–3378 (2018)
294. Ghosh, S., Lincoln, P., Tiwari, A., Zhu, X.: Trusted machine learning: Model repair and data repair for probabilistic models. In: Workshops 31st AAAI Conf. Artificial Intelligence, vol. WS-17 (2017) Methods for Safety Assurance of Machine Learning 9
295. Guo, J., Jiang, Y., Zhao, Y., Chen, Q., Sun, J.: DLFuzz: Differential fuzzing testing of deep learning systems. In: Proc. ACM Joint Meeting on European Software Engineering Conf. and Symp. Foundations of Software Engineering, pp. 739–743 (2018)
296. Hailesilassie, T.: Rule extraction algorithm for deep neural networks: A review. CoRR abs/1610.05267, 555,555 (2016)
297. Hu, Z., Ma, X., Liu, Z., Hovy, E.H., Xing, E.P.: Harnessing deep neural networks with logic rules. In: Proc. 54th Annu. Meeting of the Association for Computational Linguistics, vol. 1: Long Papers (2016)
298. Huang, X., Kroening, D., Kwiatkowska, M., Ruan, W., Sun, Y., Thamo, E., Wu, M., Yi, X.: Safety and trustworthiness of deep neural networks: A survey. CoRR abs/1812.08342 (2018)
299. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial Examples Are Not Bugs, They Are Features. CoRR abs/1905.02175 (2019)
300. Johnson, C.W.: The increasing risks of risk Evaluation: On the rise of artificial intelligence and non-determinism in safety-critical systems. In: Evolution of System Safety: Proc. Safety-Critical Systems Symp. (2017)
301. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: Proc. 29th Int. Conf. Comput. Aided Verification, pp. 97–117 (2017)
302. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems 30, pp. 5580–5590 (2017)
303. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. 35th Int. Conf. Machine Learning, vol. 80, pp. 2668–2677 (2018)
304. Kim, J., Canny, J.F.: Interpretable learning for self-driving cars by visualizing causal attention. In: Proc. 2017 IEEE Int. Conf. Comput. Vision, pp. 2961–2969 (2017)
305. Kim, J., Rohrbach, A., Darrell, T., Canny, J.F., Akata, Z.: Textual explanations for self-driving vehicles. In: Proc. 15th European Conf. Comput. Vision, Part II, vol. 11206, pp. 577–593 (2018)
306. Kindermans, P.J., Schütt, K.T., Alber, M., Müller, K.R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: PatternNet and PatternAttribution. In: Proc. 6th Int. Conf. on Learning Representations (2018)
307. Koopman, P., Fratrik, F.: How many operational design domains, objects, and events? In: Workshops of the 32nd AAAI Conf. Artificial Intelligence (2019)
308. McAllister, R., Gal, Y., Kendall, A., van der Wilk, M., Shah, A., Cipolla, R., Weller, A.: Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In: Proc. 26th Int. Joint Conf. Artificial Intelligence, pp. 4745–4753 (2017) 10 G. Schwalbe and M. Schels
309. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill 2(11), e7 (2017) [31] Pei, K., Cao, Y., Yang, J., Jana, S.: Towards practical verification of machine learning: The case

of computer vision systems. CoRR abs/1712.01785 (2017)
310. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: Proc. British Machine Vision Conf., p. 151 (2018)
311. Rabold, J., Siebers, M., Schmid, U.: Explaining black-box classifiers with ILP – empowering LIME with Aleph to approximate non-linear decisions with relational rules. In: Proc. Int. Conf. Inductive Logic Programming, pp. 105–117 (2018)
312. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
313. Roychowdhury, S., Diligenti, M., Gori, M.: Image classification using deep learning and prior knowledge. In: Workshops of the 32nd AAAI Conf. Artificial Intelligence, vol. WS-18, pp. 336–343 (2018)
314. Ruan, W., Huang, X., Kwiatkowska, M.: Reachability analysis of deep neural networks with provable guarantees. In: Proc. 27th Int. Joint Conf. Artificial Intelligence, pp. 2651–2659 (2018)
315. Salay, R., Queiroz, R., Czarnecki, K.: An analysis of ISO 26262: Using machine learning safely in automotive software. CoRR abs/1709.02435 (2017)
316. Sensoy, M., Kaplan, L.M., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems 31, pp. 3183–3193 (2018)
317. Shalev-Shwartz, S., Shammah, S., Shashua, A.: On a formal model of safe and scalable self-driving cars. CoRR abs/1708.06374 (2017)
318. Sun, Y., Wu, M., Ruan, W., Huang, X., Kwiatkowska, M., Kroening, D.: Concolic testing for deep neural networks. In: Proc. 33rd ACM/IEEE Int. Conf. Automated Software Engineering, pp. 109–119 (2018)
319. Thrun, S.: Extracting rules from artificial neural networks with distributed representations. In: Advances in Neural Information Processing Systems 7, pp. 505–512 (1995)
320. Wang, H.: ReNN: Rule-embedded neural networks. In: Proc. 24th Int. Conf. Pattern Recognition, pp. 824–829 (2018)
321. Wang, S., Pei, K., Whitehouse, J., Yang, J., Jana, S.: Formal security analysis of neural networks using symbolic intervals. In: Proc. 27th USENIX Security Symp., pp. 1599–1614 (2018)
322. Xiang, W., Musau, P., Wild, A.A., Lopez, D.M., Hamilton, N., Yang, X., Rosenfeld, J.A., Johnson, T.T.: Verification for machine learning, autonomy, and neural networks survey. CoRR abs/1810.01989 (2018)
323. Zhang, Q., Zhu, S.C.: Visual interpretability for deep learning: A survey. Front. IT EE 19(1), 27–39 (2018)
324. Sculley, D. G.-F. ((2015)). Hidden technical debt in machine learning systems. Advances in neural information processing systems 28.
325. Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. npj Digital Medicine, 6(1), 6.
326. Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in biology and medicine, 140, 105111.
327. Fernandez-Quilez, A. (2023). Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. AI and Ethics, 3(1), 257-265.
328. Athavale, J., Baldovin, A., Graefe, R., Paulitsch, M., & Rosales, R. (2020, June). AI and reliability trends in safety-critical autonomous systems on ground and air. In 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W) (pp. 74-77). IEEE.
329. Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-31.
330. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495-504.
331. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data‐driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.
332. Roselli, D., Matthews, J., & Talagala, N. (2019, May). Managing bias in AI. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 539-544).
333. Khan, S., Tsutsumi, S., Yairi, T., & Nakasuka, S. (2021). Robustness of AI-based prognostic and systems health management. Annual Reviews in Control, 51, 130-152.
334. Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and explainability of artificial intelligence. Publications Office of the European Union, 207.
335. Wu, T., Dong, Y., Dong, Z., Singa, A., Chen, X., & Zhang, Y. (2020). Testing Artificial Intelligence System Towards Safety and Robustness: State of the Art. IAENG International Journal of Computer Science, 47(3).
336. Namiot, D., & Ilyushin, E. (2022). On the robustness and security of Artificial Intelligence systems. International Journal of Open Information Technologies, 10(9), 126-134.
337. Cheng, C.H., Nührenberg, G., Huang, C.H., Ruess, H., Yasuoka, H.: Towards dependability metrics for neural networks. In: 16th ACM/IEEE Int. Conf. Formal Methods and Models for System Design, pp. 43–46 (2018)
338. Bagschik, G., Menzel, T., Maurer, M.: Ontology based scene creation for the development of automated vehicles. In: Proc. 2018 IEEE Intelligent Vehicles Symp., pp. 1813–1820 (2018)
339. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: Proc. 29th Int. Conf. Comput. Aided Ver-

ification, pp. 97–117 (2017)
340. Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to explain: An information-theoretic perspective on model interpretation. In: Proc. 35th Int. Conf. Machine Learning, vol. 80, pp. 883–892 (2018)
341. Amodei, D., Olah, C., Steinhardt, J., Christiano, P.F., Schulman, J., Mané, D.: Concrete problems in AI safety. CoRR abs/1606.06565 (2016)
342. Fridovich-Keil, D., Herbert, S.L., Fisac, J.F., Deglurkar, S., Tomlin, C.J.: Planning, fast and slow: A framework for adaptive real-time safe trajectory planning. In: Proc. 2018 IEEE Int. Conf. Robotics and Automation, pp. 387–394 (2018)
343. I.S.: Road Vehicles — Functional Safety — Part 6: Product Development at the Software Level, vol. 6, 2 edn. (2018)
344. Koopman, P., Fratrik, F.: How many operational design domains, objects, and events? In: Workshops of the 32nd AAAI Conf. Artificial Intelligence (2019)
345. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems 30, pp. 5580–5590 (2017)
346. Gast, J., Roth, S.: Lightweight probabilistic deep networks. In: Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition, pp. 3369–3378 (2018)
347. Roychowdhury, S., Diligenti, M., Gori, M.: Image classification using deep learning and prior knowledge. In: Workshops of the 32nd AAAI Conf. Artificial Intelligence, vol. WS-18, pp. 336–343 (2018)
348. Hu, Z., Ma, X., Liu, Z., Hovy, E.H., Xing, E.P.: Harnessing deep neural networks with logic rules. In: Proc. 54th Annu. Meeting of the Association for Computational Linguistics, vol. 1: Long Papers (2016)
349. Wang, H.: ReNN: Rule-embedded neural networks. In: Proc. 24th Int. Conf. Pattern Recognition, pp. 824–829 (2018)
350. Ghosh, S., Lincoln, P., Tiwari, A., Zhu, X.: Trusted machine learning: Model repair and data repair for probabilistic models. In: Workshops 31st AAAI Conf. Artificial Intelligence, vol. WS-17 (2017) Methods for Safety Assurance of Machine Learning 9
351. Fridovich-Keil, D., Herbert, S.L., Fisac, J.F., Deglurkar, S., Tomlin, C.J.: Planning, fast and slow: A framework for adaptive real-time safe trajectory planning. In: Proc. 2018 IEEE Int. Conf. Robotics and Automation, pp. 387–394 (2018)
352. Akametalu, A.K., Fisac, J.F., Gillula, J.H., Kaynama, S., Zeilinger, M.N., Tomlin, C.J.: Reachability-based safe learning with gaussian processes. In: Proc. 53rd IEEE Conf. Decision and Control, pp. 1424–1431 (2014)
353. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial Examples Are Not Bugs, They Are Features. CoRR abs/1905.02175 (2019)
354. Dreossi, T., Ghosh, S., Yue, X., Keutzer, K., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: Counterexample-guided data augmentation. In: Proc. 27th Int. Joint Conf. Artificial Intelligence, pp. 2071–2078 (2018)
355. Dutta, S., Jha, S., Sankaranarayanan, S., Tiwari, A.: Output range analysis for deep feedforward neural networks. In: Proc. 10th Int. Symp. NASA Formal Methods, vol. 10811, pp. 121–138 (2018)
356. Bunel, R.R., Turkaslan, I., Torr, P., Kohli, P., Mudigonda, P.K.: A unified view of piecewise linear neural network verification. In: Advances in Neural Information Processing Systems 31, pp. 4790–4799 (2018)
357. Wang, S., Pei, K., Whitehouse, J., Yang, J., Jana, S.: Formal security analysis of neural networks using symbolic intervals. In: Proc. 27th USENIX Security Symp., pp. 1599–1614 (2018)
358. Ruan, W., Huang, X., Kwiatkowska, M.: Reachability analysis of deep neural networks with provable guarantees. In: Proc. 27th Int. Joint Conf. Artificial Intelligence, pp. 2651–2659 (2018)
359. Pei, K., Cao, Y., Yang, J., Jana, S.: Towards practical verification of machine learning: The case of computer vision systems. CoRR abs/1712.01785 (2017)
360. Sun, Y., Wu, M., Ruan, W., Huang, X., Kwiatkowska, M., Kroening, D.: Concolic testing for deep neural networks. In: Proc. 33rd ACM/IEEE Int. Conf. Automated Software Engineering, pp. 109–119 (2018)
361. Guo, J., Jiang, Y., Zhao, Y., Chen, Q., Sun, J.: DLFuzz: Differential fuzzing testing of deep learning systems. In: Proc. ACM Joint Meeting on European Software Engineering Conf. and Symp. Foundations of Software Engineering, pp. 739–743 (2018)
362. Dreossi, T., Ghosh, S., Yue, X., Keutzer, K., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: Counterexample-guided data augmentation. In: Proc. 27th Int. Joint Conf. Artificial Intelligence, pp. 2071–2078 (2018)
363. Kindermans, P.J., Schütt, K.T., Alber, M., Müller, K.R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: PatternNet and PatternAttribution. In: Proc. 6th Int. Conf. on Learning Representations (2018)
364. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
365. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: Proc. British Machine Vision Conf., p. 151 (2018)
366. Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to explain: An information-theoretic perspective on model interpretation. In: Proc. 35th Int. Conf. Machine Learning, vol. 80, pp. 883–892 (2018)
367. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill 2(11), e7 (2017) [31] Pei, K., Cao, Y., Yang, J., Jana, S.: Towards practical verification of machine learning: The case of computer vision systems. CoRR abs/1712.01785 (2017)

368. Kim, J., Rohrbach, A., Darrell, T., Canny, J.F., Akata, Z.: Textual explanations for self-driving vehicles. In: Proc. 15th European Conf. Comput. Vision, Part II, vol. 11206, pp. 577–593 (2018)
369. Roychowdhury, S., Diligenti, M., Gori, M.: Image classification using deep learning and prior knowledge. In: Workshops of the 32nd AAAI Conf. Artificial Intelligence, vol. WS-18, pp. 336–343 (2018)
370. Rabold, J., Siebers, M., Schmid, U.: Explaining black-box classifiers with ILP – empowering LIME with Aleph to approximate non-linear decisions with relational rules. In: Proc. Int. Conf. Inductive Logic Programming, pp. 105–117 (2018)
371. Thrun, S.: Extracting rules from artificial neural networks with distributed representations. In: Advances in Neural Information Processing Systems 7, pp. 505–512 (1995)
372. Hailesilassie, T.: Rule extraction algorithm for deep neural networks: A review. CoRR abs/1610.05267, 555,555 (2016)
373. Fuchs, F.B., Groth, O., Kosiorek, A.R., Bewley, A., Wulfmeier, M., Vedaldi, A., Posner, I.: Neural Stethoscopes: Unifying analytic, auxiliary and adversarial network probing. CoRR abs/1806.05502 (2018)
374. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. 35th Int. Conf. Machine Learning, vol. 80, pp. 2668–2677 (2018)
375. Fong, R., Vedaldi, A.: Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition, pp. 8730–8738 (2018)
376. Wang, H.: ReNN: Rule-embedded neural networks. In: Proc. 24th Int. Conf. Pattern Recognition, pp. 824–829 (2018)