AI Systems Trustworthiness Assessment: State of the Art

Afef Awadid, Kahina Amokrane-Ferka, Henri Sohier, Juliette Mattioli, Faouzi Adjed, Martin Gonzalez, Souhaiel Khalfaoui

AI Systems Trustworthiness Assessment: State of the Art

Afef AWADID[1], Kahina AMOKRANE-FERKA[1], Henri SOHIER[1], Juliette MATTIOLI[2], Faouzi ADJED[1], Martin GONZALEZ[1], Souhaiel KHALFAOUI[1,3]

[1]*IRT SystemX, France*

[2]*Thales, France*

[3]*Valeo, France*

{*afef.awadid, kahina.amokrane-ferka, henri.sohier, faouzi.adjed, martin.gonzalez*}*@irt-systemx.fr,*
juliette.mattioli@thalesgroup.com, souhaiel.khalfaoui@valeo.com

Keywords: AI-Based Systems, Trustworthiness Assessment, Trustworthiness Attributes, Metrics, State of the Art Review.

Abstract: Model-based System Engineering (MBSE) has been advocated as a promising approach to reduce the complexity of AI-based systems development. However, given the uncertainties and risks associated with Artificial Intelligence (AI), the successful application of MBSE requires the assessment of AI trustworthiness. To deal with this issue, this paper provides a state of the art review of AI trustworthiness assessment in terms of trustworthiness attributes/ characteristics and their corresponding evaluation metrics. Examples of such attributes include data quality, robustness, and explainability. The proposed review is based on academic and industrial literature conducted within the Confiance.ai research program.

1 INTRODUCTION

Central to Model-based Systems Engineering (MBSE) is "the formalized application of modeling to support system requirements, design, analysis, verification, and validation activities beginning in the conceptual design phase and continuing throughout development and later life cycle phases" (INCOSE, 2007). MBSE, therefore, advocates "the use of models to perform systems engineering activities that are traditionally performed using documents" (Mann, 2009).

This promotes the understanding of complex systems engineering processes including Artificial Intelligence (AI) systems engineering as a multi-engineering process (Mattioli et al., 2023d). However, the successful application of MBSE requires the assessment of AI trustworthiness defined by the ISO/IEC DIS 30145-2 standard as the "ability to meet stakeholders' expectations in a verifiable way". Indeed, without an accompanying assessment of trustworthiness from the early stages of development, the deployment of an AI component within a safety critical systems such as in avionics, mobility, healthcare and defense becomes risky (Mattioli et al., 2023b).

In view of this, it is not surprising that the quantification of AI-based system trustworthiness has become a hot topic (Braunschweig et al., 2022). AI system trustworthiness is defined in terms of characteristics/ attributes such as reliability, safety, and resiliency (AI, 2019). In this context, the paper at hand provides a state of the art review of AI trustworthiness assessment. Such review focuses on the main trustworthiness attributes as well as their evaluation metrics, and is based on academic and industrial literature conducted within the Confiance.ai research program.

The rest of the paper is organized as follows. Section 2 introduces the context and motivation of this work. Section 3 presents the state of the art review of AI systems trustworthiness assessment with respect to trustworthiness attributes and their evaluation metrics. Finally, Section 4 concludes the paper and opens up for future work.

2 CONTEXT AND MOTIVATION

Safety-critical systems, such as those used in avionics, mobility, healthcare, and defense, are designed to operate reliably and safely in dynamic environments where their failure could have severe consequences.

The adoption of Artificial Intelligence (AI) depends on their ability to deliver the expected service safely, to meet user expectations, and to maintain service continuity. Thus, such systems have to be valid, accountable, explainable, resilient, safe and secure,

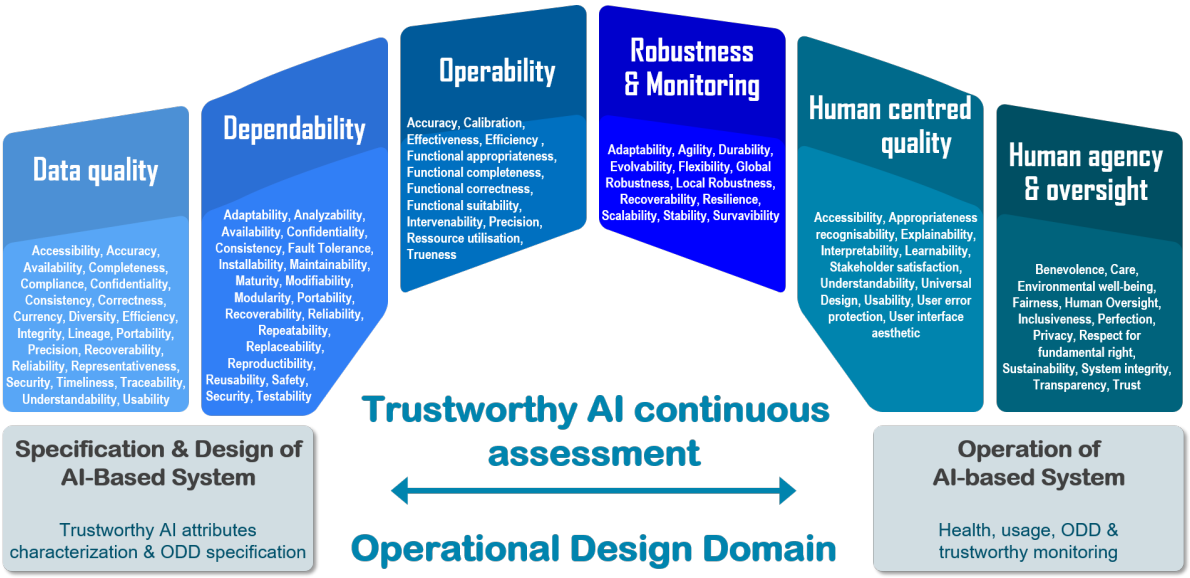

Figure 1: Trustworthy AI-based critical systems (Mattioli et al., 2023a).

compliant with respect to regulation and standardization (including ethics and sustainability).
Assessing the trustworthiness of AI becomes the cornerstone of successful improvement in the design and operation of critical systems. However, obtaining trustworthiness measures remains a challenging task.

On the one hand, measuring trust can help identify problems with the system before they become critical and allow for corrective action to be taken before a failure occurs. On the other hand, measuring trust can help to improve the design of critical systems.

By understanding the factors that contribute to user trust in AI systems, designers can create ones that are more reliable, safe and secure. AI trustworthiness characterization is multi-dimensional and multi-criteria as assessed by different stakeholders (Mattioli et al., 2023b) (regulators, developers, end-users). In this context, (Felderer and Ramler, 2021) proposed to consider three dimensions; the artifact type dimension (system, model and data perspective), the process dimension and the quality characteristic attributes, based on ISO/IEC 25023, that bear on software product or system quality.

AI-based systems, especially those using machine learning (ML), add a level of complexity to traditional systems, due to their inherent stochastic nature. Thus, to take into account the complexity of the ML-based systems engineering process, the set of trustworthiness properties illustrated in Figure 1 (Mattioli et al., 2023a), needs to be extended. Additionally, various experts and stakeholders are involved in the design of such systems.

Moreover, to assess AI trustworthiness, the choice of the relevant attributes is not easy, since the selection pertains to the context of application, which is modeled according to several elements (ODD, intended domain of use, nature and roles of the stakeholders...). The attributes can be quantitative (typically numerical values either derived from a measure or providing a comprehensive and statistical overview of a phenomenon) or qualitative (based on the detailed analysis and interpretation of a limited number of samples). Then once the list of relevant attributes has been defined, the aggregation of several attributes remains complex due to commensurability issues: indeed, this is equivalent with combining ”oranges and apples”, none of the attributes having the same unit. In addition, one aims at making trade-offs and arbitration between the attributes. This means that the value of each attribute should be transformed into a scale common to all attributes and representing the preferences of a stakeholder, and that the values of the scales for the different criteria should be aggregated. These elements constitute the main steps for solving the problem using a multi-criteria decision making (MCDA) approach.

MCDA is a generic term for a collection of systematic approaches developed specifically to help one or several decision makers to assess or compare some alternatives on the basis of several criteria (Labreuche, 2011). The difficulty is that the decision criteria are frequently numerous, interdependent/overlapping and sometime conflicting. For example, effectiveness may be conflicting with robust-

ness, explainability, or affordability. The viewpoints are quantified through attributes.

Aggregation functions are often used to compare alternatives evaluated on multiple conflicting criteria by synthesizing their performances into overall utility values (Grabisch and Labreuche, 2010). Such functions must be sufficiently expressive to fit the stakeholder's preferences, allowing for instance the determination of the preferred alternative or to make compromises among the criteria - improving a criterion implies that one shall deteriorate on another one.

MCDA provides a tool to specify the good compromises (Labreuche, 2011). Our approach is based on the following steps:

1. Step 1: Structuring attributes in a semantic tree;
2. Step 2: Identification of numerical evaluations;
3. Step 3: Adapting attributes for commensurability;
4. Step 4: Definition of an aggregation methodology to capture operational trade-offs and evaluate higher-level attributes.

Given this, to improve trustworthiness, assessment, measures and processes are needed. Moreover, context, usage, levels of safety and security, regulations, (ethical) standards (including fairness, privacy), certification processes, and degrees of liability should be considered. In addition to measures and processes, various techniques and methodologies such as testing, evaluation, and validation of the system's performance against specified criteria, expert review, and stakeholder participation are required for trustworthiness assessment in AI-based critical systems.

Such assessment should be ongoing, with regular updates and monitoring of the system's performance and compliant with standards and regulations. Besides trustworthy attribute definitions (Adedjouma et al., 2022), the current work focuses on some examples of associated metrics that help to identify potential areas for improvement. It is important to note in this context that such attributes have a different meaning depending on the stakeholder's profile. For instance, a system engineer, a safety engineer, a data engineer, and an AI scientist may all have distinct perspectives on accuracy.

3 A NEW AI TRUSTWORTHINESS META-MODEL

A trustworthy software is defined (Wing, 2021) by a combination of overlapping properties: reliability, safety, security, privacy, availability and usability. For a ML-based system, this translates and extends to accuracy, robustness, fairness, accountability, transparency, explainability and ethics. (Delseny et al., 2021) also considers auditability.

To capture the type of considered information and the different inter-relations needed to assess ML trustworthiness, we proposed a meta-model with concepts in different abstraction levels (see Figure. 2). The red part describes the way the tree of attributes is built. It highlights the abstract concepts central to trustworthiness assessment. An attribute which aggregates other attributes is called a macro-attribute (e.g. robustness, dependability, *etc.*). It is assessed with an aggregation method. An atomic attribute (leaf attribute) is assessed with a clear and actionable observable which can take different forms (metric, "expected proof").

The green part of Figure. 2 is the meta-model fragment with concrete concepts. These concepts represent the different possible subjects and relations between them. For example, the product is developed following processes as technical processes (through which the product must go: design definition, implementation, operation, ...), agreement processes (with external organizations: acquisition, supply), and management processes (supporting the development of the product: quality management, risk management, *etc.*). Risk and quality management ensures the compliance with the specification which includes the different expected trustworthiness attributes. Processes are applied with tools by people respecting a certain governance.

The blue part summarizes systems engineering key concepts more precisely part of the non-functional specification: they do not define what the system "does" or how the system works, but what the system "is". The attributes are also commonly referred to as "*-ilities*" as they often have this suffix. They can also be referred to as quality requirements. Whether a specification is functional or non-functional, it is influenced by stakeholders such as the user, the operator, the developer, *etc.*

As opposed to non-functional requirements which define what the system is, functional requirements define what the system does: should it move? roll? roll fast? under what conditions? From this point of view, the Operational Design Domain (ODD), which characterizes the conditions of operation of the system, can be considered part of the functional specification relating to trustworthiness attributes in different ways: 1) Having transparency or clear visibility into the ODD permits to understand the system's capabilities and limits (which is part of the AI Act's requirements); 2) The ODD is the domain to consider for the different operational trustworthiness attributes; 3)

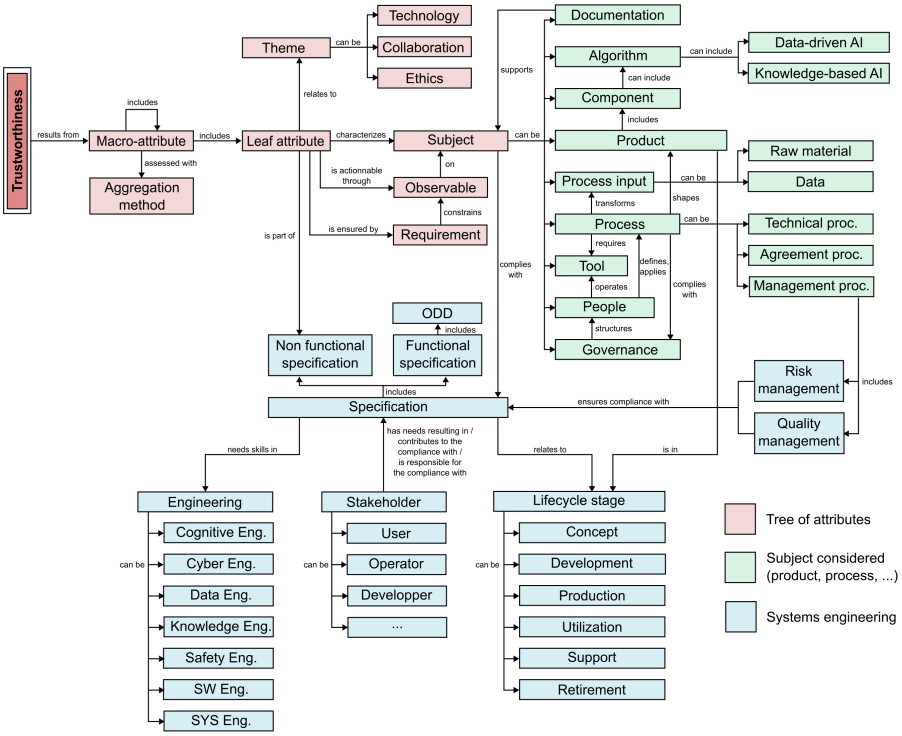

Figure 2: A new AI trustworthiness meta-model (Mattioli et al., 2023c)

The ODD has its own attributes (it should be complete, free of inconsistencies, human readable, *etc.*).

In contrast to non-functional requirements, which define what the system is, functional requirements define what the system does: does it move? does it roll? does it roll fast? under what conditions? From this point of view, the Operational Design Domain (ODD), which characterizes the operating conditions of the system/feature of interest, can be considered as part of the functional specification in relation to the reliability attributes in a number of ways: 1) the transparency of the ODD makes it possible to understand the limitations of the system (a requirement of the AI Act); 2) the ODD is the domain to be considered for the different operational reliability attributes; 3) the ODD has its own attributes (it should be complete, free of inconsistencies, human readable, *etc.*).

Thus, the trustworthy attributes can be assessed only if the ODD is clearly defined but many AI prototypes neglect to describe their ODD or leave it vaguely defined as the domain covered by the distribution of data used during training. In addition to the set of Requirements applicable to the System, one of the results of the System Specification phase is the ODD Definition/Specification, that aims at specifying the sub-domain where automation features are expected to operate according to their requirements, among the whole operational domain of

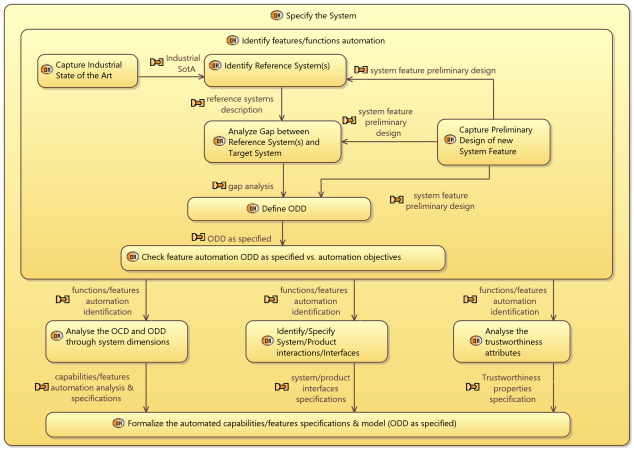

Figure 3: Engineering Activities for "System Specification" from AI perspective

the system/product. The diagram structures presented in fig. 3, the engineering activities needed to perform System Specification with AI/ML involvement in mind.

At every stage of the system lifecycle, from engineering and design to operation, trustworthiness relationships must be established and maintained. According to the seven pillars of reliability (High-Level Expert Group on Artificial Intelligence, 2019), *Confiance.ai* specifies AI reliability (Mattioli et al., 2023a)

by six macro-attributes: data/information/knowledge quality, dependability, operability, robustness, explainability/interpretability, and human control.

Through the system development, the desired ODD will be refined to fit the needs and constraints from different engineering. Here comes the question about where the ODD process stops. Figure 4 extends the ODD approaches with concepts related to ODD limits. As an example, the ODD is an important input for ML training, ML monitoring, etc. To be able to address the concerns of these engineering fields, one needs to define datasets, to define test scenarios, to identify the indicators to measure the algorithm performance, etc. A current expectation is that the ODD artefact must include all the features for deriving those elements (scenario definition, robustness metric, monitoring variables). However, following the definition of the ODD, it is not the case. Besides, that information is required by the engineering fields in general, regardless if the system is AI-based or not AI-based, which let induces that such information may come from another source.

4 AI-BASED SYSTEMS TRUSTWORTHINESS ASSESSMENT

4.1 Data Quality Assessment

In ML discipline, most of the research is focusing on model performance improvement more than on datasets (Mazumder et al., 2022). In the recent decade, ML techniques have advanced significantly and achieved a high maturity level (Adedjouma et al., 2022). Classical ML practices consist typically in using the existing datasets and in leveraging performances challenges through techniques complexity enhancement. In the other hand, data-driven AI takes a broader approach by placing a greater emphasis on the data itself (Jakubik et al., 2022; Jarrahi and Others, 2022). Instead of simply looking for patterns and relationships within the input features, data-driven AI involves collecting, processing, and analyzing large amounts of data to create more accurate and robust models (Mattioli and other, 2022).

Moreover, a real challenge today is to associate datasets to the Operational Design Domain (ODD)[1] from the operational level of the system definition. Indeed, these datasets include several factors such as user needs (Chapman et al., 2020) and related metadata. Moreover, (Mountrakis and Xi, 2013) highlights that dataset quality may have a more significant impact on performance than any model design choice. Many industrialization crisis often result from the data used to train the models instead of the model designs and architectures.

Without a systematic assessment of their quality, data-driven AI risks losing control of the various steps of data engineering such as collection, annotation and feature engineering. Doing without data quality assessment would result in assuming that data engineering can not be further improved and that problems will always be detected without systematic analysis. Thus, in a given end-to-end AI-based system process, the data quality assessment brings an evaluation of some ODD description aspects. These evaluation goes through a set of metrics, illustrated in Figure.1, such as data accuracy, data representativeness and data diversity.

Furthermore, to ensure conformity to the ODD specifications, well-founded metrics assess the reached data quality level. Both research and industrial practices have developed relevant data quality metrics in the AI-based system, such as accuracy and completeness. However, many of them still lack a sound foundation (Heinrich et al., 2018). Thereafter, a definition and a brief technical description of five metrics for data quality assessment are given.

Data completeness for ML datasets refers to the degree to which it contains the necessary information required to accurately model the underlying patterns by the learning algorithm. Measuring dataset completeness includes evaluation of the amount of missing items, outliers and errors. Completeness metric could be based on the Ge and Helfert's ratio (Ge and Helfert, 2006) defined as: $data\_completeness = \sum_{i=1}^{N} \gamma(d_i)/N$, where $\gamma(d_i)$ is 0 if d_i is a missing data, and 1 otherwise.

Data correctness refers to the accuracy of the data items to faithfully represent the real-world phenomena or objects they meant to capture. Dataset correctness could be defined as: $data\_correctness = 1/(1+d(\omega,\omega_m))$ where ω is the data value to be assessed, ω_m is the corresponding real value and d is a domain-specific distance measure such as the Euclidean or Hamming distance.

Data diversity is defined by the evaluation of the presence of all required information and quantifies how the dataset fits the environment and application domains described in the specifications. During ML model design, training and testing, the level

[1] An ODD is concept created initially for automated driving system (ADS) used to restrict where the ADS is valid (Gyllenhammar et al., 2020). In the current work, ODD is a restriction of the domain where an AI-based system acts safely.

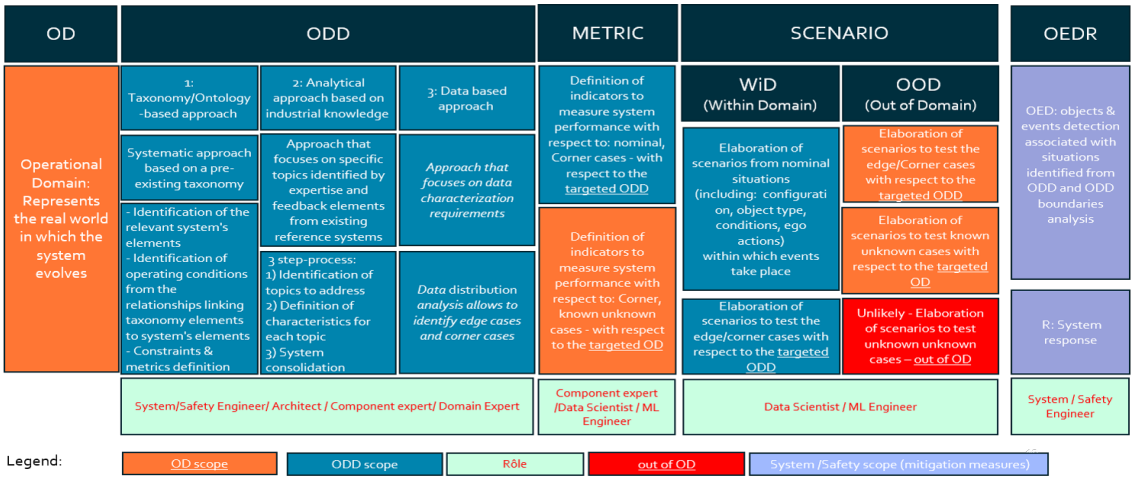

Figure 4: The ODD analysis process

of diversity should be equally distributed for the different data subsets being selected. This should ensure that the ML model is enough diversified so as to cover its domain of possible stimuli. According to (Gong et al., 2019), the only metric used for diversity for supervised ML is the Determinantal Point Process (DPP) introduced by (Kulesza et al., 2012). Then,(Dereziński, 2019) regularizes the DPP (R-DPP) to accelerate the training process. Moreover, other diversity indexes, used in biology and ecology, could be adapted for ML models such as Shannon entropy and mean proportional species abundance (Tuomisto, 2010).

Data representativeness (Mamalet et al., 2021) refers in statistics to the notion of sample and population. Transposed to AI, the sample corresponds to the data-set available for the development of the model (training, validation, testing), and the population corresponds to all possible observations in the field of application. Moreover, a dataset is representative when it describes the environment observations, and the distribution of its key characteristics is conform to the specifications need, requirements and the ODD of the targeted application. There are multiple existing methods to quantify the representativeness of datasets, stemming from statistics and ML fields. Indeed, Student, Chi-square and Kolmogorov-Smirnov tests may be applied to assess the goodness of fit of specified distributions. Furthermore, in case of large datasets, the confidence interval combined with the maximum entropy probability could be used to determine, in terms of dataset size and acceptance thresholds, the suitable dataset for ML need (Blatchford et al., 2021).

4.2 Operability Assessment

By assessing operability, we can ensure that we deliver value to end-users and avoid problems at design-time, where operability is the ability to keep such AI-based system in a safe and reliable functioning condition, according to predefined operational requirements. Thus, (AI-based system) operability is a measure of how well the system works in a production environment, for both end users and developers. Good operability induces diagnosis and recovery for AI maintenance. In an operational context, it is also defined as the degree to which a product or system is easy to use, monitor, control and maintain and to use. Thereby, accuracy, trueness and precision are required for AI/ML operability assessment and considered as different concepts when referring to measurements.

Generally speaking, accuracy refers to how close a measured value is in relation to a known value. However, the ISO (International Organization for Standardization) uses trueness for the above definition while keeping the word accuracy to refer to the combination of trueness and precision. On the other hand, precision is related to how close several measurements of the same quantity are to each other. Thus, (model) accuracy is a fundamental metric for evaluating AI-based critical system, to measure how well the system performs its intended function compared to its ground truth or expected output.

Traditional operability metrics for regression include Mean Squared Error (MSE) or Mean Absolute Error (MAE), while classification problems can be evaluated through precision, accuracy and recall (Davis and Goadrich, 2006). In classification, a confusion matrix (depicting the distribution of true/false

negatives/positives for each class) is a practical tool for visualizing of the errors, and allows the computation of most metrics: precision, recall, sensitivity, specificity, F1 score, Calibration measures how well the AI system's predicted probabilities match the true probabilities of the outcomes. This can be evaluated using various metrics, such as the Brier score or calibration plot.

Let us denotes by TP (True Positive) to represent how many positive class samples your model predicted correctly; TN (True Negative) to represent how many negative class samples your model predicted correctly; FP (False Positive) to represent how many positive class samples your model predicted incorrectly and FN (False Negative) to represent how many negative class samples your model predicted incorrectly.

- **Accuracy** measures how often the Model produces correct results where $Accuracy = (TP + TN)/(TP + FP + TN + FN)$.
- **Precision** measures the proportion of true positives out of all positive predictions: $Precision = (TP)/(TP + FP)$.
- **Recall** measures the proportion of true positives out of all actual positives: $Recall = TP/(TP + FN)$
- **F1_Score** is a combination of precision and recall providing a single score to evaluate the overall performance of the AI system: $F1\_Score = 2 * (Precision * Recall)/(Precision + Recall)$.
- **Specificity** is the proportion of actual negatives that the model has correctly identified as such out of all negatives: $Specificity = TN/(FP + TN)$
- **ROC Curve** plots the true positive rate against the false positive rate at various classification thresholds, and can be used to evaluate the overall performance of the AI system. The area under the ROC curve is often used as a metric, with higher area indicating better performance.

Closely related to accuracy, trueness, and precision is correctness. Indeed, the latter is defined by ISO-25010 as the *degree to which a product or system provides the correct results with the needed degree of precision*. In ML, correctness measures the probability that the ML system under test "gets things right". Let D be the distribution of future unknown data. Let x be a data item belonging to D. Let h be the ML model that we are testing. $h(x)$ is the predicted label of x, $c(x)$ is the true label. (Zhang et al., 2020) defined the model correctness $E(h)$ as the probability that $h(x)$ and $c(x)$ are identical, $E(h) = Pr_{x \sim D}[h(x) = c(x)]$. Note that there are many other metrics that can be used to assess the operability of AI systems, and the choice of metric(s) depends on the specific use case and application.

4.3 Dependability Assessment

As AI becomes prevalent in critical systems, their dependability takes on increasing importance. In systems engineering, dependability can be defined as the ability of a system to deliver a service that can be justifiably trusted (Avizienis et al., 2004). But, this concept has evolved to integrate other attributes: Availability readiness for correct service; Reliability continuity of correct service; Safety for absence of catastrophic consequences on user(s) and environment; Security availability for authorized users, confidentiality, and integrity; Confidentiality for absence of unauthorized disclosure of information; Integrity for absence of improper system alterations; and Maintainability for ability to undergo modifications, and repairs. Moreover, the requirements on the AI system cannot be described completely, and the system must function dependably in an almost infinite application space. This is where established methods and techniques of classical systems and software engineering reach their limits and new, innovative approaches are required. A core element to assess dependability is to provide assurance (Buckley and Poston, 1984) that the system as a whole is dependable, i.e., that risk of failures is mitigate to an acceptable level. In a data-driven AI component, the functionality is not programmed in the traditional way, but created by applying algorithms to data. One challenge is to make this (learning) assurance case (Byun and Rayadurgam, 2020) cleanly and to find appropriate evidence that demonstrates the dependability of the AI system.

- **Availability** and reliability are often used interchangeably but they actually refer to different things. Reliability refers to the probability of an AI-based component/system performing without failure under normal operating conditions over a given period of time. Thus, availability measure provides an indication of the percentage of the time that the system is actually available over the scheduled operational time. The first step in calculating availability is deciding the period we want to analyze. Then, it is calculated by dividing *Uptime* by the total sum of *Uptime* and *Downtime*: $Availability = Uptime/(Uptime + Downtime)$, where *Uptime* (resp. *Downtime*) represents the time during the system is operational(resp. isn't operational). *Downtime* has the biggest impact on availability and is one of key

KPIs for maintenance and in service support activities. Moreover, estimating AI-based software MTBF (Mean time between failure) is a tricky task. This interval may be estimated from the defect rate of the system or can also be based on previous experience with similar systems.

- **Reliability** is the probability that an asset will perform a required function under specified conditions, without failure, for a specified period. For AI systems, the definition of AI reliability is defined as (Kaur and Bahl, 2014) "*the probability of the failure-free software operation for a specified period of time in a specified environment*". Common measurements of reliability are MTBF and mean time to failure (MTTF). MTBF measures the average time between two consecutive failures, while MTTF accounts for the time elapsing from the beginning of operation to the detection of the first failure. Nevertheless, the measurement of the reliability of an AI algorithm is associated to its performance. Most data-driven AI algorithms are designed to solve problems of classification, regression, and clustering, (Bosnić and Kononenko, 2009) used prediction accuracy from ML algorithms as a reliability measure.
- **Repeatability** and **reproducibility** are also characteristics of dependability. In the context of AI engineering, repeatability measures the variation in various runs of test plan under the same conditions, while reproducibility measures whether an entire experiment can be reproduced in its entirety. This verification facilitates the detection, analysis, and mitigation of potential risks in an AI system, such as a vulnerability on specific inputs or unintended bias. Therefore, reproducibility is emerging as a concern among AI Engineers.

Moreover, depending on the AI methods used, functional safety can still be measured and verified, insofar as such properties can be formally defined. For example, ML dependability properties have to be entirely verified in the field of aviation following the "*EASA Concept Paper: guidance for Level 1 & 2 ML applications*".

4.4 Robustness Assessment and Monitoring

AI-based critical systems should be robust, secure and safe throughout their entire life-cycle in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and without unreasonable safety risk. To this end, robustness is mandatory to ensure that an invalid input data will not lead to an unsafe state of the system. This can be reached "by-design" and it can also be monitored "in operations" to enable analysis of the AI system's outcomes and responses appropriate to the context.

Therefore, robustness and monitoring are two closely related topics in an AI-based system lifecycle. Robustness and stability are defined by (Mamalet et al., 2021) as an AI-based system's global robustness (out of distribution), the ability to perform its intended function in the presence of abnormal or unknown inputs; and local robustness (in distribution), the extent to which the system provides equivalent responses to similar inputs.

These definitions are made more precise by (SAE J3016, 2018) using the ODD concept. The global robustness is then called robustness and the local robustness is called stability where robustness is an AI asset's ability to maintain its expected/intended performance under well-characterized abnormalities or deviations in inputs and operating conditions under its ODD; and stability is the ability of an AI asset to maintain its expected/intended output(s) under well-characterized and bounded perturbations to its inputs and operating conditions within its ODD.

In addition, adversarial robustness refers to the ability of models to maintain their performance in the face of adversarial attacks and perturbations where perturbations are imperceptible, non-random changes to the input that alter a model's prediction, thereby maximizing its error (Kapusta et al., 2023). Some tooled methods dedicated to local robustness assessment are based on evaluation or (formal) demonstration such as:

- Non-overlapping corruption on a dataset provides an assessment of the robustness of a given AI-based model (Py et al., 2023);
- AI Metamorphism Observing Software (AIMOS) (Girard-Satabin et al., 2022) assesses metamorphic properties on AI models such as robustness to perturbations on the inputs but also relation between models' inputs and outputs;
- Time-series robustness characterization focuses and the assessment of the robustness w.r.t. perturbations on the inputs of regression models applied to time series;
- Adversarial attack characterization: (Kapusta et al., 2023) evaluates the impact and usability of adversarial attacks on AI models;
- Amplification methods evaluate the robustness of models using amplification methods on the dataset with noise functions.

Monitoring comprises methods for inspecting system in order to analyze and predict its behavior.

Enforcement, on the other hand, involves designing mechanisms for controlling and restricting the behavior of systems. Once an AI systems are deployed, we need tools to continuously monitor and adjust them. Thus, the main objective of online monitoring of AI models is to identify the output that does not fulfill the expectations by detecting any deviation in operation from the specified expected behavior, or from a predefined set of trustworthy operational properties (Kaakai and Raffi, 2023). In Confiance.ai program, (Adedjouma et al., 2022) addresses both concepts, monitoring and enforcement, by combining several monitoring timescales (Present Time, Near-Past and Near-Future Monitoring) - with a rule-based approach to compute the final "safe output".

4.5 Explainability Assessment

The need to explain AI algorithms gave rise to the field of Explainable AI. In the literature, several studies argue that explanations positively affect user trust (Biran and Cotton, 2017) and inappropriate trust impairs human-machine interaction (Ribeiro et al., 2016). For example, in data-driven AI, explainability is a main property to bring trust to models, given the black box nature of AI. This property is related to the notion of explanation as an interface between humans and AI. It involves AI systems that are accurate and understandable to people (Philippe et al., 2022). However, explanations do not necessarily have to provide accurate information about the algorithm of the ML process. In our study, we consider three key dimensions of explainability:

- Interpretability: assesses how easily human experts can understand the internal workings of an AI system; interpretable explanations need to use a representation that is understandable to humans, regardless of the actual features used by the model (Ribeiro et al., 2016). In the context of ML systems, interpretability is defined as the ability to explain or to present in understandable terms to a human (Doshi-Velez and Kim, 2017);
- Fidelity measures how well the explanations provided accurately reflect the AI system behavior (Yeh et al., 2019). Fidelity metrics measure the efficiency of the methods to explain models. Fidelity is also defined (Plumb et al., 2020), when the explainer's output space is $(\varepsilon_s, (\varepsilon_s := (g \in G | g : X \to Y))$, the explanation is defined as a function $g : X \to Y$, and it is natural to evaluate how accurately g models f in a neighborhood $N_x : F(f,g,N_x) := E_{x' \sim N_x}[(g(x') - f(x'))^2]$ which refer to the neighborhood-fidelity (NF) metric. This metric is sometimes evaluated with N_x as a point mass on x, this version is called the point-fidelity (PF) metric.
- Usefulness: evaluates how effectively the explanations support human decision-making and action. This last dimension is qualitative. In the evaluation context, some questions can be asked to the user (Tambwekar and Gombolay, 2023): Using this explanation would be useful for me? Using this explanation will improve my effectiveness. Using this explanation will improve my performance.
- Faithfulness: measures the degree to which an interpretation method accurately reflects the reasoning of the model it interprets. It is important to note that explanations provided by an unfaithful method can conceal any biases that exist in the model's judgments, which may result in unwarranted trust or confidence in the model's predictions. Faithfulness is calculated using the following formula (Du et al., 2019): $Faithfulness = 1/N \sum (y_{x^i} - y_{x^i|a})$, where y_{x^i} is the predicted probability for a given target class using the original inputs, and $y_{x^i|a}$ is the predicted probability for the target class for the input with significant sentences/words removed. According to (Arya et al., 2022), faithfulness is the inverse of the Pearson Product-Moment correlation and ranges from -1 to 1. A negative correlation of 1 indicates a perfect correlation, a positive correlation of -1 indicates the inverse, and 0 indicates no correlation. Faithfulness is calculated as follows: $Faithfulness = -\sigma_{xy}/(\sigma_x + \sigma_y)$, where σ_x^2 (resp. σ_{xy}^2) represents the variance of x (resp. the co-variance of (x,y)). This metric can be interchangeable with Fidelity metric in some methodes.
- Monotonicity: applies only to some explainable methods. It consists in progressively adding the values of x to a null vector, then looking if the probability of predicting the correct class with it is increasing (Ribeiro et al., 2016); The interest in studying monotonicity in the context of MBSE lies in its ability to enhance the understanding and analysis of complex systems.
- Sensitivity: measures the degree of explanation changes to subtle input perturbations using Monte Carlo sampling-based approximation (Yeh et al., 2019).

4.6 Human-centered Quality & Human Oversight Assessment

To ensure trustworthy AI, it is important to go beyond the AI model itself (inputs, features and outputs) and consider dynamics of the model interacting with the overall system, including end-users. Human-centered quality involves meeting requirements for "*usability, accessibility, user experience, and avoiding harm from use*".

From such perspective, trustworthy AI should be both usable and explainable, meaning that it should not stop working at inappropriate times (which could create safety risks) and should be user-friendly for individuals with diverse backgrounds. Moreover, trustworthy AI must allow for human explanation and analysis to mitigate risks and empower users, as well as transparent to promote understanding of its workings mechanism. Human agency and oversight means that AI systems shall be developed and used as a tool that serves people, respects human dignity and personal autonomy, and being under human control and oversight. In that context, ethics guidelines for trustworthy AI were written by High-Level Expert Group on AI (High-Level Expert Group on Artificial Intelligence, 2019). The guidelines have 4 ethical principles: (1) Respect for human, (2) Prevention of harm, (3) Fairness, (4) Explainability; and seven key (ethical) requirements, among it, we can mention:

- Privacy: For IEEE-7000 privacy means that collection with unsolicited surveillance, processing with unexpected and unsolicited personal data aggregation, and the dissemination of personal information is carried out in such a way that it preserves the self-determination of the person with regard to information (breach of confidentiality, disclosure) and that any form of invasion is prevented (intrusion against the will). In practice, collection implies that data acquired are cleaned of private information. Once stored, the cleaned data may still fall within the scope of privacy when crossed with other data. Privacy rules must be explicit and respected throughout the data life cycle. When data is crossed and processed, information must be anonymized. This implies that data remains coherent, and that representativeness, diversity, and completeness are preserved. In (Fjeld et al., 2020) eight principles of privacy are highlighted: control over the use of data, ability to restrict data processing, right to rectification, right to erasure, privacy by design, and recommends data protection laws, and privacy (other/general).
- Respect for fundamental rights: During human-machine interaction, the machine is perceived as a) attentive, by replying in a reasonable amount of time, and b) responsive, by respecting user privacy, with appropriate decision criteria, transparency, fairness and politeness. For instance, fairness is unsatisfied when biases were introduced during model training in case of unreliable sources or distribution shifts over the time; transparency is neither met when data for learning were suppressed nor traced. Like the other data quality attributes, the definition of respect requirements with their thresholds is essential to regularly assess data and dataset quality - what must be considered at the beginning of the development of an AI system (High-Level Expert Group on Artificial Intelligence, 2019).

5 CONCLUSIONS AND PERSPECTIVES

This paper highlights the importance of assessing AI trustworthiness in the context of Model-based System Engineering (MBSE) for the development of AI-based systems. The complexity and uncertainties associated with AI necessitate a comprehensive evaluation of trustworthiness attributes and corresponding evaluation metrics. The state of the art review presented in this paper provides insights into the various trustworthiness attributes that need to be considered when assessing AI systems. These attributes include data quality, robustness, and explainability, among others. Each attribute plays a crucial role in ensuring the reliability, safety, and ethical implications of AI systems.

The review is based on a thorough analysis of academic and industrial literature conducted within the Confiance.ai research program. This ensures that the findings are grounded in both theoretical and practical perspectives, making them relevant and applicable to real-world scenarios.

By considering the trustworthiness attributes and evaluation metrics identified in this review, MBSE practitioners can effectively assess the trustworthiness of AI-based systems. This assessment is essential for mitigating risks, addressing uncertainties, and building confidence in the deployment and utilization of AI technologies.

However, it is important to note that the field of AI trustworthiness assessment is rapidly evolving, and new attributes and evaluation metrics may emerge in the future. Therefore, future research will focus on keeping up with the advancements in AI technology and hence extending this work to include other trust-

worthiness attributes and metrics.

ACKNOWLEDGEMENTS

This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute within the Confiance.ai Program (www.confiance.ai).

REFERENCES

Adedjouma, M., Adam, J.-L., Aknin, P., Alix, C., Baril, X., Bernard, G., Bonhomme, Y., Braunschweig, B., Cantat, L., Chale-Gongora, G., et al. (2022). Towards the engineering of trustworthy AI applications for critical systems - the confiance.ai program.

AI, U. L. I. (2019). A plan for federal engagement in developing technical standards and related tools.

Arya, V. et al. (2022). AI Explainability 360: Impact and design. In *Proceedings of the AAAI Conf.*, volume 36 (11).

Avizienis, A. et al. (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. on dependable and secure computing*, 1(1):11–33.

Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13.

Blatchford, M. L., Mannaerts, C. M., and Zeng, Y. (2021). Determining representative sample size for validation of continuous, large continental remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 94:102235.

Bosnić, Z. and Kononenko, I. (2009). An overview of advances in reliability estimation of individual predictions in ML. *Intelligent Data Analysis*, 13(2):385–401.

Braunschweig, B., Gelin, R., and Terrier, F. (2022). The wall of safety for AI: approaches in the confiance.ai program. In *SafeAI@ AAAI*, volume 3087 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Buckley, F. J. and Poston, R. (1984). Software quality assurance. *IEEE Trans. on Software Engineering*, 1(1):36–41.

Byun, T. and Rayadurgam, S. (2020). Manifold for machine learning assurance. In *ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results*, pages 97–100.

Chapman, A. et al. (2020). Dataset search: a survey. *The VLDB J.*, 29(1):251–272.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Delseny, H., Gabreau, C., Gauffriau, A., Beaudouin, B., Ponsolle, L., Alecu, L., Bonnin, H., Beltran, B., Duchel, D., Ginestet, J.-B., et al. (2021). White paper machine learning in certified systems. *arXiv preprint arXiv:2103.10529*.

Dereziński, M. (2019). Fast determinantal point processes via distortion-free intermediate sampling. In *Conf. on Learning Theory*, pages 1029–1049. PMLR.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Du, M. et al. (2019). On attribution of recurrent neural network predictions via additive decomposition. In *The WWW Conf.*, pages 383–393.

Felderer, M. and Ramler, R. (2021). Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session). In *International Conf. on Software Quality*, pages 33–42. Springer.

Fjeld, J. et al. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*.

Ge, M. and Helfert, M. (2006). A framework to assess decision quality using information quality dimensions. In *ICIQ*, pages 455–466.

Girard-Satabin, J. et al. (2022). CAISAR: A platform for Characterizing Artificial Intelligence Safety and Robustness. In *AI Safety workshop of IJCAI-ECAI*.

Gong, Z. et al. (2019). Diversity in machine learning. *IEEE Access*, 7:64323–64350.

Grabisch, M. and Labreuche, C. (2010). A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175(1):247–286.

Gyllenhammar, M. et al. (2020). Towards an operational design domain that supports the safety argumentation of an automated driving system. In *10th European Congress on Embedded Real Time Systems (ERTS)*.

Heinrich, B. et al. (2018). Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)*, 9(2):1–32.

High-Level Expert Group on Artificial Intelligence (2019). Assessment list for trustworthy artificial intelligence (altai). Technical report, European Commission.

INCOSE, T. (2007). Systems engineering vision 2020. *INCOSE, San Diego, CA, accessed Jan*, 26(2019):2.

Jakubik, J. et al. (2022). Data-centric artificial intelligence. *arXiv 2212.11854*.

Jarrahi, M. and Others (2022). The Principles of Data-Centric AI. *arXiv 2211.14611*.

Kaakai, F. and Raffi, P.-M. (2023). Towards multi-timescale online monitoring of ai models: Principles and preliminary results. In *SafeAI@ AAAI*.

Kapusta, K., , et al. (2023). Protecting ownership rights of ml models using watermarking in the light of adversarial attacks. In *AAAI Spring Symposium - AITA: AI Trustworthiness Assessment*.

Kaur, G. and Bahl, K. (2014). Software reliability, metrics, reliability improvement using agile process. *Int. J. of Innovative Science, Engineering & Techno.*, 1(3):143–147.

Kulesza, A. et al. (2012). Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Labreuche, C. (2011). A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence*, 175(7-8):1410–1448.

Mamalet, F. et al. (2021). White Paper Machine Learning in Certified Systems. Research report, ANITI.

Mann, C. (2009). A practical guide to sysml: The systems modeling language. *Kybernetes*, 38(1/2).

Mattioli, J. et al. (2023a). An overview of key trustworthiness attributes and kpis for trusted ml-based systems engineering. In *AI Trustworthiness Assessment (AITA) @ AAAI Spring Symposium*.

Mattioli, J. et al. (2023b). Towards a holistic approach for ai trustworthiness assessment based upon aids for multi-criteria aggregation. In *SafeAI @ AAAI*.

Mattioli, J. et al. (2023c). Towards a holistic approach for ai trustworthiness assessment based upon aids for multi-criteria aggregation. In *SafeAI@ AAAI*.

Mattioli, J., Le Roux, X., Braunschweig, B., Cantat, L., Tschirhart, F., Robert, B., Gelin, R., and Nicolas, Y. (2023d). Ai engineering to deploy reliable ai in industry. In *AI4I*.

Mattioli, J. and other (2022). Empowering the trustworthiness of ml-based critical systems through engineering activities. *arXiv preprint arXiv:2209.15438*.

Mazumder, M. et al. (2022). Dataperf: Benchmarks for data-centric ai development. *arXiv:2207.10062*.

Mountrakis, G. and Xi, B. (2013). Assessing reference dataset representativeness through confidence metrics based on information density. *ISPRS journal of photogrammetry and remote sensing*, 78:129–147.

Philippe, D., David, V., Alice, P., Antoine, C., Antonin, P., Caroline, G., and Allouche, T. (2022). Explainability benchmark v2 - the confiance.ai program.

Plumb, G., Al-Shedivat, M., Cabrera, Á. A., Perer, A., Xing, E., and Talwalkar, A. (2020). Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems*, 33:10526–10536.

Py, E. et al. (2023). Real-time weather monitoring and desnowification through image purification. In *AAAI Spring Symposium - AITA: AI Trustworthiness Assessment*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

SAE J3016 (2018). Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems.

Tambwekar, P. and Gombolay, M. (2023). Towards reconciling usability and usefulness of explainable ai methodologies. *arXiv preprint arXiv:2301.05347*.

Tuomisto, H. (2010). A diversity of beta diversities: straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33(1):2–22.

Wing, J. M. (2021). Trustworthy ai. *Communications of the ACM*, 64(10):64–71.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. (2019). On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32.

Zhang, J. M., Harman, M., Ma, L., and Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*.