Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation

Juliette Mattioli, Henri Sohier, Agnès Delaborde, Gabriel Pedroza, Kahina Amokrane, Afef Awadid, Zakaria Chihani, Souhaiel Khalfaoui

Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation*

Juliette MATTIOLI[1], Henri SOHIER[2], Agnès DELABORDE[3,2], Gabriel PEDROZA[4], Kahina AMOKRANE-FERKA[2], Afef AWADID[2], Zakaria CHIHANI[4] and Souhaiel KHALFAOUI[5,2]

[1]*Thales, France*
[2]*IRT SystemX, France*
[3]*Laboratoire National de métrologie et d'Essais LNE, France*
[4]*Université Paris-Saclay, CEA, List, France*
[5]*Valéo, France*

Abstract
The assessment of AI-based systems trustworthiness is a challenging process given the complexity of the subject which involves qualitative and quantifiable concepts, a wide heterogeneity and granularity of attributes, and in some cases even the non-commensurability of the latter. Evaluating trustworthiness of AI-enabled systems is in particular decisive in safety-critical domains where AIs are expected to mostly operate autonomously. To overcome these issues, the Confiance.ai program [1] proposes an innovative solution based upon a multi-criteria decision analysis. The approach encompasses several phases: structuring trustworthiness as a set of well-defined attributes, the exploration of attributes to determine related performance metrics (or indicators), the selection of assessment methods or control points, and structuring a multi-criteria aggregation method to estimate a global evaluation of trust. The approach is illustrated by applying some performance metrics to a data-driven AI context whereas the focus on aggregation methods is left as a near-term perspective of Confiance.ai milestones.

Keywords
Trustworthiness Assessment, Trustworthiness Attributes, Trustworthiness Metrics and Key Performance Indicators (KPIs), Multi Criteria Decision Aid, Data Quality, Robustness, Explainability

1. Introduction

Without an accompanying assessment of trustworthiness from the early stages of development, the deployment of an Artificial Intelligence (AI) component within a safety-critical systems such as in avionics, mobility, healthcare and defense becomes risky.

1.1. Trustworthiness definition

Trust is the willingness of one party to perform certain actions that are important to stakeholders (AI scientist, safety engineer, certification auditor, end-user, *etc.*) regardless of the other partys ability to monitor or control [2]. Trust is defined [3] as "*the degree to which a user or other stakeholder has confidence that a product or system will behave as intended*". But, the trust literature distinguishes trustworthiness (the ability, benevolence, and integrity of a trustee) from trust (the intention to accept vulnerability to a trustee based on positive expectations of his or her actions) [4]. Trustworthiness is represented as an objective aspect of trust estimated based on evidences or observations; whereas trust includes subjective aspects of a cognitive entity's opinion such as a human. We consider the following definition: trustworthiness (ISO/IEC DIS 30145-2) is the "*ability to meet stakeholders' expectations in a verifiable way*". Moreover, [5] identified nine characteristics that define AI system trustworthiness: accuracy, reliability, resiliency, objectivity, security, explainability, safety, accountability, and privacy.

1.2. Trustworthiness attributes

Trustworthiness is a complex concept which can be broken down into different attributes. In 1977, a FAA (Federal Aviation Administration) panel dedicated to how to certify aircraft as airworthy, explicitly linked the notion of trustworthiness to accounting. Then, security and dependability became key system attributes [6] to assess the trustworthiness of a computer-based system: Avizienis et al. [7] used dependability to represent the overall quality measure of a system based on four sub-attributes including security, safety, reliability, and maintainability.

SafeAI 2023: The AAAI's Workshop on Artificial Intelligence Safety
*Corresponding author: J. MATTIOLI
†These authors contributed equally.
juliette.mattioli@thalesgroup.com (J. MATTIOLI); henri.sohier@itr-systemx.fr (H. SOHIER); agnes.delaborde@lne.fr (A. DELABORDE); gabriel.pedroza@cea.fr (G. PEDROZA); kahina.amokrane-ferka@irt-systemx.fr (K. AMOKRANE-FERKA); afef.awadid@irt-systemx.fr (A. AWADID); Zakaria.CHIHANI@cea.fr (Z. CHIHANI); souhaiel.khalfaoui@valeo.com (S. KHALFAOUI)

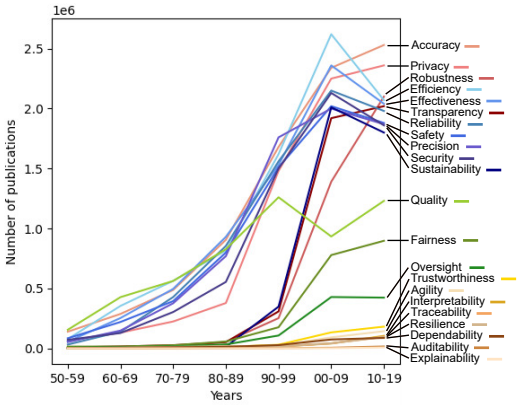

Figure 1: Evolution of the number of publications on Google Scholar between 1950 and 2019

To depict dynamic system security metrics, Pendleton et al. [8] presented a cyber-security metric framework based on the interactions between attackers and defenders. By taking into account a variety of probabilistic and/or analytical models to undertake a quantitative evaluation, Ramos et al. [9] primarily examined model-based network security indicators. Gol Mohammadi et al. [10] proposed a first list of software trustworthiness attributes. Fig. 1 shows the evolution of the number of publications on different attributes. [11] explores the potential benefits and challenges of quantifying AI risk by discussing how such an assessment could improve AI regulation.

These attributes must be mapped onto the AI processes and life-cycle, keeping track of their relations with the different stakeholders. Indeed, the definition of trustworthiness is up to individual interpretation and preference. For instance, to feel confident in the handling of safety-critical data, end-users can be concerned with usability.

1.3. Trustworthiness management

Trustworthiness does not only emerge from the product itself, but also from the process (how the product was made), the tools and infrastructure (with what), the people (by whom) as well as the governance (who decides). Trustworthiness can also be considered from a quality point of view or a risk point of view. In the former, the chances to meet the stakeholder expectations are maximized (by good practices and clear metrics). In the latter, the chances not to meet the stakeholder expectations are minimized (by identifying and mitigating potential issues). Thus, all trustworthiness attributes can generally be considered from a quality or risk point of view. Quality is at the center of the SQuaRE (Systems and software Quality Requirements and Evaluation) series of standards ISO/IEC 25000:2014 and AI quality is more specifically considered in ISO/IEC DIS 25059 (under development). The principles of risk management are explained in ISO 31000:2018 and AI risk is more specifically considered in ISO/IEC FDIS 23894 (under development).

1.4. AI heterogeneity

Trustworthiness assessment should be adapted to the different natures of AI systems. AI can follow three major paradigms: 1) **Data-driven AI**, which includes statistical and connexionist AI suitable for pattern matching and recognition, classification and forecasting problems; 2) **Knowledge-based AI** (also called Symbolic AI) relies on knowledge representations such as ontologies and conceptual graphs. Real world information is transformed into something understandable and usable by machines so that decisions can follow an organized path of planning, solution searching and optimization; and 3) **Hybrid AI** which is more than a combination of symbolic AI and machine learning approaches by encompassing any synergistic combinations of various AI techniques, which could be enhanced by a priori knowledge (such as mathematics, physics or geometry).

Not only should trustworthiness assessment be adapted to these paradigms, but also to the various AI functions such as anomaly detection (on images or time series), forecasting, decision making under uncertainty, planing and scheduling problems, optimization under constraints, *etc.* Trustworthiness assessment is field dependant (mobility, healthcare, finance, *etc*). Thus, an adaptation effort is highly requested.

1.5. Can trustworthiness be measured?

The quantification of AI-based system trustworthiness has become a hot topic [12]. From a strict metrological point of view, measurement is relative to a physical property which can be compared to a reference quantity of the same kind. Following this definition, trustworthiness cannot be "measured". However, each attribute related to trustworthiness can be represented on a scale (*e.g.*, number scale, nominal scale, ordinals scale) [1]. A trustworthiness metric can be defined as objective, mathematical measure of the AI-based component/system that is sensitive to differences in safety critical characteristics. It provides a quantitative measure of an attribute which the body of solution exhibit. For example, estimating the trustworthiness of a system can rely on performance and/or quality scoring (*e.g.*, for reliability: Fleiss Kappa score, goodness-of-fit tests, or for accuracy: precision, recall, F-score, *etc*). However, trustworthiness is not only based on objective attributes – for example, usability and interpretability are linked to human judgment; trustworthiness assessment should then also include rigorous methodological processes to manage subjectivity.

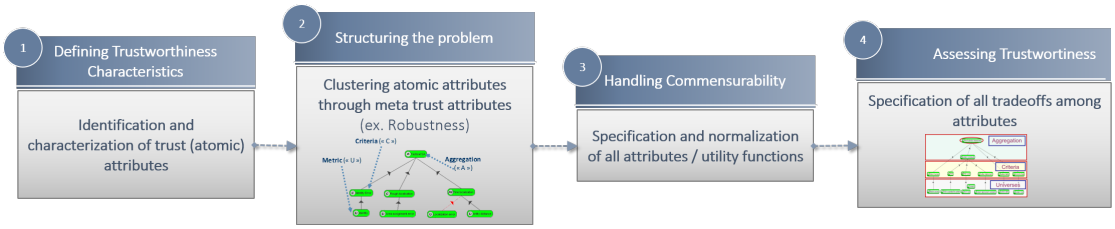

Figure 2: The unified approach based on MCDA

2. Unified approach to support trustworthiness assessment

Multi-Criteria Decision Aiding (MCDA) is a generic term for a collection of systematic approaches developed specifically to help one or several Decision Makers (DM) to assess or compare some alternatives on the basis of several criteria [13]. The difficulty is that the decision criteria are frequently numerous, dependent and sometime conflicting. For example, effectiveness may be conflicting with robustness, explainability, or affordability. The viewpoints are quantified through attributes (see §2.1).

First, 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 (Operational Design Domain, intended domain of use, nature and roles of the stakeholders, *etc.*) 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 an MCDA approach.

Aggregation functions are often used to compare alternatives evaluated on multiple conflicting criteria by synthesizing their performances into overall utility values [14]. Such functions must be sufficiently expressive to fit the DM'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 [13].

Our approach is based on the following steps (see fig.2):

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.

2.1. Step 1: Semantic tree

Based on different sources (norms, standards, scientific communications, industrial and institutional reports, Confiance.ai reports, *etc.*), the characterization and evaluation of trust attributes focus on defining and structuring the attributes that constitute trust in the context of AI-based safety critical systems [15] going beyond a risk analysis as proposed in [16, 17].

Our problem of assessing Trustworthiness is decomposed in several sub-problems by introducing a hierarchy of an important number of specific criteria. This structuring phase aims to construct a tree representing a hierarchy of points of view in which the root represents the overall evaluation, and the leaves are the elementary attributes. In order to produce such a hierarchy, one shall succeed in grouping the criteria according to a classification that makes sense for the stakeholders. At the end of this step, one shall obtain the relevant criteria together with their organization in a tree. This first step has been captured in the mind-map of Fig.3.

The attributes are currently grouped according to the capabilities they characterize: technology, ethics, interaction and trust intermediaries (such as certification).

Technology is system-centric, it refers to the ability to verify that the AI-based component has valid and robust intrinsic properties such as accuracy, robustness, safety and security. Thus, AI-based systems should generate accurate output as consistent as possible with the ground truth. Additionally, AI systems should be robust to changes, specifically in complex, dynamic and uncertain real environments. Moreover, AI programs or systems must not harm any human being under any circumstances that prioritize user safety. In addition, the autonomy of trustworthy AI should always be under user's control. In other words, it has always been

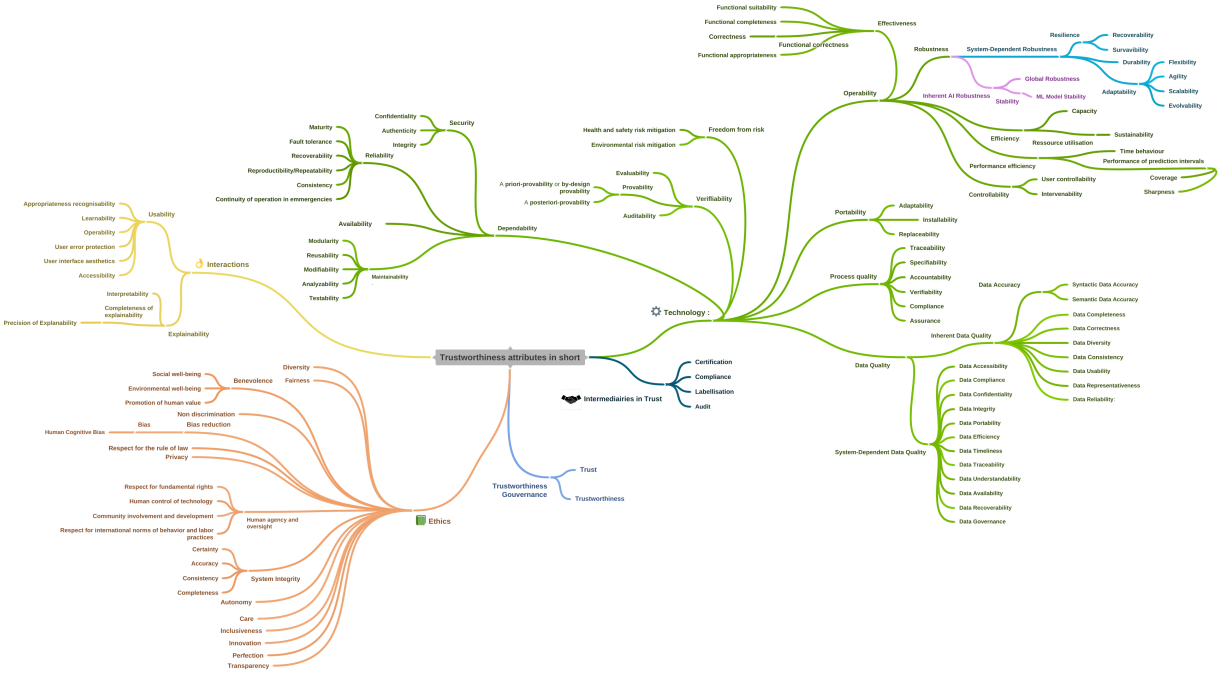

Figure 3: A first mindmap to structure the problem of AI-based system trustworthiness assessment

the human right to give the AI system decision-making authority or to revoke that authority at any time.

From **interaction's perspective**, trustworthy AI should possess the properties of usability, explainability and interpretability. Specifically, AI-based systems should not cease operation at inappropriate times (e.g. at times when the lack of output could lead to safety risks), and these programs or systems should be easy to use for people with different backgrounds. Trustworthy AI solutions should allow explanation and analysis by humans to reduce potential risks and harms and empower human users. In addition, trustworthy AI should be transparent so people can better understand its mechanism.

Ethics, strongly linked in Europe to the notion of fundamental rights, is notably put forward in the work of the AI HLEG (High-Level Expert Group on Artificial Intelligence) of the European Commission [18]. A system must, for example, be law-abiding, fair, accountable, environmentally friendly and compliant with the user privacy. Specifically, AI systems should operate in accordance with all relevant laws and regulations, as well as with the ethical principles of human society.

As some notions (e.g. explainability) concern several dimensions (ethics vs. interaction), Confiance.ai program made an arbitrary choice to be consistent with the methodology. Finally, the attributes for trusted ecosystem intermediaries focus on the relationships to third-parties, in particular quality assurance, audit and certification activities. All these properties apply to the AI-based component and the system that embeds AI, but they also apply to the quality of the data used for training connectionist AI and/or to the quality of knowledge representation and reasoning used in symbolic AI.

2.2. Step 2: Numerical evaluations

All nodes return a numerical evaluation. Specific Key Performance Indicators (KPI), metrics or evaluation methods are used to qualify the leaves of the tree according to the use cases. For example, data quality is a problem that has been studied for several decades now [19]. However, primarily the focus has been on the data in operational databases and data warehouses. Now, Data-driven AI is generating renewed interest in data quality, but there is yet no consensus on what comprises the data quality characteristics. Thus, [20] were among the first arguing that limiting quality to the level of accuracy is not enough, highlighting that the level of quality for given data can depend on its purpose. Its principles require an assessment of the various quality attributes as presented in §3.2, mainly in fig.6. Standards are currently being developed to define data quality attributes for ML (Machine Learning): ISO/IEC CD 5259-1 (terminology and principles) and ISO/IEC CD 5259-2 (data quality measures).

2.3. Step 3: Commensurability

Aggregating different attributes for a global assessment requires that they are commensurate. This implies that one shall be able to compare any numerical evaluation of an attribute with any numerical evaluation of any other attribute. In order to make the assessment "comparable," sound methods for normalization (to make the comparison between variables comparable) have to be

applied to single variables in order to first make them comparable, that is, transforming the various scales of variables into one unique scale. The numerical evaluation of the attributes is thus encoded in the $[0, 1]$ interval where the value 0 corresponds to the total absence of the property beneath a trustworthiness criterion, and value 1 corresponds to the complete satisfaction of the criterion. The normalized indicators could be aggregated using specific formulas (e.g. min/max, arithmetic mean, weighted sum, *etc.*). If one attribute is more "important" than another with respect to stakeholder preference, the former is assigned a stronger weight than the latter within the aggregation procedure.

2.4. Step 4: Aggregation and trade-off

The global assessment would be made on the basis of several trustworthiness attributes denoted by $N = \{1, ..., n\}$. The proposed approach provides a tool to identify best compromises from the stakeholder point of view. Each attribute $i \in N$ is quantified by a KPI – also called metric or Figure of Merit – represented by the set of its possible values X_i. The alternatives are characterized by a value on each attribute and can be fully described by elements of $X = X_1 \times ... \times X_n$. An alternative x can thus be represented by a vector $(x_1, ..., x_n) \in X$.

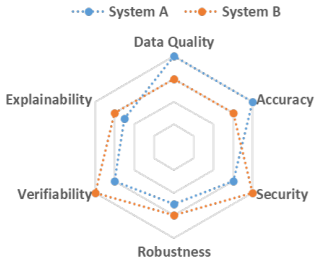

Figure 4: Radar chart is useful for comparing two systems which embed different AI approaches.

Radar chart is a visual method for comprehensive evaluation, particularly useful for holistic and overall assessment through multivariate data. However, this representation does not allow understanding the interactions and dependencies between attributes.

The goal of MCDA is to define a numerical representation of the preferences of the stakeholders, expressed as a function $u : X \rightarrow [0, 1]$. The function will be used to compare each alternative or assess each alternative's level of satisfaction in order to provide the overall level of satisfaction for each. As mentioned before, the scale $[0, 1]$ can be interpreted as a degree of satisfaction. It is classical to write u in a decomposed way : $u(x) = F(u_1(x_1), ..., u_n(x_n))$, for all $x \in X$, where $u_i : X_i \rightarrow [0, 1]$ is a utility function (also called value functions) and $F : [0, 1]^n \rightarrow [0, 1]$ is called the aggregation function where $F(0, ..., 0) = 0$, $F(1, ..., 1) = 1$, and $F(x_1, ..., x_n) \leq F(y_1, ..., y_n)$ if $x_i \leq y_i, \forall i$. Moreover, the utility function normalizes the metrics and provides a measure of satisfaction for a single metric. The aggregation function takes a normalized score as input and returns an aggregate score. We recommend using the MACBETH approach [21, 22] to develop utility functions for resolving past difficulties related to interval scale construction and compatibility issues in a way that fully satisfies stakeholders. and mathematically significant.

The most widely used aggregation function is the weighted sum. It assumes the independence among the criteria. This is a major limitation as criteria often interact. We need to use other type of aggregation function such as the Choquet integral [23, 24], which is an extension of the weighted sum that is capable of measuring the influence of the importance of the individual criteria and the importance of the interrelationships among criteria.

Today, this step is work in progress.

3. Focus on Data-driven AI

In real-world industrial settings, the data-driven AI model is only a small part of the overall system and significant additional engineering and system functionalities are required to ensure that the model can operate in a reliable, predictable and scalable way with proper engineering of data and model pipelines, monitoring and logging, etc. While the necessity and usefulness of reasoning about trust assessment is obvious, obtaining trustworthiness scores remains a challenging task. To illustrate such issues, as stated previously, some aspects linked to trustworthiness are highly subjective or context dependent. For example, the notion of "data quality" (resp. "robustness", "explainability") requires having a knowledge of all induced attributes including those that are system dependent such as data availability, data portability, data precision, *etc.* (resp. adaptability, durability, resilience, *etc.*). The subjectivity or vagueness of the attribute definitions does not always represent a major hindrance to use them in operational settings, because skills and knowledge of AI and safety engineers may be enough to determine what may be appropriate thresholds and scores.

3.1. Classification performance

Classification is a prediction type used to give the output variable in the form of categories with similar attributes. Some of the popular metrics for the assessment of classification are Accuracy, Precision, Recall, F1 Score... **Confusion Matrix** is a core element that can be used to visualize the performance of the ML classification model, but it is a tool rather than a metric. By nature, it is a table

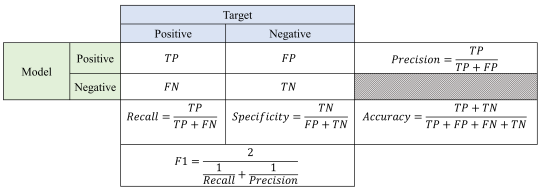

Figure 5: Performance metrics for classification problems

with two dimensions showing actual values and predicted values. Each row of the confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class. Each cell in the confusion matrix represents an evaluation factor. For example, for a binary classification of "positive" and "negative":

- True Positive (TP) signifies how many positive class samples your model predicted correctly.
- True Negative (TN) signifies how many negative class samples your model predicted correctly.
- False Positive (FP) signifies how many positive class samples your model predicted incorrectly.
- False Negative (FN) signifies how many negative class samples your model predicted incorrectly.

A **precision score** (see fig. 5) close or equal to 1 will signify that your model did not miss any true positives, and is able to classify well between correct and incorrect labeling of observed data. **Recall** is the proportion of actual positives that the model has correctly identified as such out of all positives. **Specificity** is the proportion of actual negatives that the model has correctly identified as such out of all negatives. A high **F1 score** symbolizes a high precision as well as high recall. It presents a good balance between precision and recall and gives good results on imbalanced classification problems. The **ROC Curve** is a plot which shows the performance of a binary classifier as function of its cut-off threshold. It essentially shows the TP rate against the FP rate for various threshold values. Selecting the most suitable evaluation metric strongly depends on the way how the stakeholder defines the criticality of the application.

3.2. Data quality

Data quality is at the center of the standard ISO/IEC 25012:2008. The standard distinguishes between "nherent data quality" and "system-dependent data quality". The former is intrinsic to the data and does not depend on the application (e.g. correctness). The later is application specific (e.g. accuracy). Data quality can also be considered for a complete data set (e.g. completeness) or for a unique value (e.g. currentness). For example, some data quality characteristics are described in fig. 6.

Traditionally, metrics for **data accuracy** are based on the rate of correct data items over an entire data set, using a 1 for an accurate data item, and a 0 otherwise: $data\_accuracy = \sum_{i=1}^{N} \alpha(d_i)/N$ where N is the number of data elements in the dataset, and $\alpha(d_i)$ is 1 if data element d_i is correct, and 0 otherwise.

The assessment of the **data timeliness** attribute [25] indicates whether the data was submitted in due time, respecting the data gathering deadline:

$$data\_timeliness = max\left(1 - \frac{\text{age of the data value}}{\text{shelf life}}\right)^{s}$$

where "age of the data value" represents the time difference between the occurrence (i.e., when the data value was created) and the assessment of timeliness of the data value; "shelf life" is defined as the maximum length of time the values of the considered attribute remain up-to-date, which can be determined through expert knowledge. Thus, a higher value of the parameter shelf life implies a higher value of the metric for timeliness, and vice versa. The exponent $s > 0$, which has to be determined based on expert estimations, influences the sensitivity of the metric to the ratio (age of the data value / shelf life).

The **data completeness** metric could be based on the Ge and Helfert's ratio [26], and 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 could be defined as: $data\_correctness = 1/(1 + d(\omega, \omega_m))$ where ω is the data value to be assessed, ω_ms the corresponding real value and d is a domain-specific distance measure such as the Euclidean distance or the Hamming distance. A larger difference between ω and ω_m s represented by a larger value of the distance function, which in turn leads to a larger denominator and thus a smaller metric value.

3.3. Robustness

The IEEE glossary of software engineering [27] defines robustness as "*The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions*". Related terms are thus error and fault tolerance, the first one regarding error in data inputs whereas the second at component level. In the present context, the robustness of an AI-enabled system essentially depends and is focused on the ML or DL (Deep Learning) components and the phase in the development cycle where the training model is designed and tested (*i.e.* the phase involving ML/DL algorithms, training and testing data sets). As it has been highlighted in [28], ML/DL models exhibit counter-intuitive properties like (1) the misclassification of (adversarial) perturbations that are statistically indistinguishable from the ones in the training data set (*e.g.* identically distributed noises) and (2) the misclassification of data subsets representing a semantic unit/object in the presence of minor perturbations that break data regularity whereas still being readable by humans (*e.g.* small square in stop signal).

Characteristic	Definition
Accuracy	The degree to which the data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use.
Correctness	Degree to which the data is free from errors.
Currentness	The degree to which data has attributes that are of the right age in a specific context of use.
Representativeness	Degree to which the data is representative of the statistical population.

Figure 6: Some data quality characteristics defined in ISO/IEC 25012:2008

Overall, a definition allowing to test the **robustness** of a model M is the measure of the impact of the minimum adversarial perturbation across many samples x. In addition, robustness can (should) be tested at two levels of possible perturbations as follows [29]:

Local robustness is satisfied by a single data input $x \in D$ of a model M and a given perturbation x' within a neighborhood δ iff $M(x)$ is identical to $M(x')$, in other words: $\forall x',\ d(x,x') \leq \delta \Rightarrow M(x) = M(x')$

Global robustness is satisfied by the set of data D of a model M, considering possible δ perturbations x' for all inputs $x \in D$, and exhibiting smooth convergence of $M(x')$ towards $M(x)$ during classification, in other words: $\forall x, x' \in D,\ d(x,x') \leq \delta \Rightarrow M(x) \rightarrow M(x')$.

If the model outputs $M(D)$ conform a dense set allowing a distance metrics $s(.)$, the convergence can be validated for a given $\varepsilon > 0$ satisfying $s(M(x), M(x')) < \varepsilon$. In practice, such post-condition could be difficult or unfeasible to verify depending upon the nature of $M(D)$. Further means are thus needed to understand how perturbations impact misclassification.

3.4. Explainability

Explainability is by far one of the most rich and complex to assess feature in recent research concerning ML/DL topics. Several reasons justify such fact, in particular because explainability aims to provide answer as to why ML/DL algorithms succeed or fail, which is rather a challenge given their heuristic nature and intricacy. Explainability is also a high-level demand in several domains aiming to transfer safety-critical human-based tasks to autonomous systems, e.g. for accountability purposes, a basic explainability requirement for a self-driving vehicle being to sufficiently characterize the contribution of the ML/DL network branches during pedestrian/obstacle detection. Last, yet not the least, explaining the behavior of ML/DL models can be tihgtly coupled to solve apparent conflicts/inconsistencies between certain attributes/features. To illustrate this point, we recall the counter-intuitive properties of a ML/DL model referred in [28] which are rather related to robustness, *e.g.* misclassification of statistically indistinguishable points, misclassification of the same points after adding new training data and updating model parameters, etc. Such apparently inconsistent outcomes make designers raise questions about ML/DL models' foundations and call for methods to characterize them. Thus, explainability is a term used to encapsulate and refer to all previous needs and aims and is therefore a qualitative attribute of AI-based systems. That said, there are no unified methods or scales to evaluate explainability. Recent surveys, as the one offered by [30], suggest that explainability can be decomposed by the methods used to evaluate it. A brief description of the main families found in literature is provided below.

Visualization methods pursue the characterization of a ML/DL network by visual observation of the levels of activation/deactivation according to the input data and their influence in the classification performance, sensitivity, and other functional/structural properties. Representative instances in this family are:

Back-propagation helps to observe relevance of data in terms of the activation/deactivation gradients observed at different layers in the network during training, *e.g.* Activation Maximization [31], Deconvolution [32], Layer-wise Relevance Propagation [33].

Perturbation-based methods provide means to observe and compare its impact in the network w.r.t. non-perturbed input, *e.g.* Occlusion Sensitivity [32], Representation Erasure [34], Meaningful Perturbation [35].

Distillation methods aim to represent (distill) the knowledge encoded in the ML/DL network after training via a more human-readable format suitable for both user interpretation and logic/machine reasoning. Some representative instances in this family are:

Local Approximation methods mimic the input/output behavior of the target ML/DL model on smaller data sets, and using approximation functions, *e.g.* linear functions. Local Approximation methods are based upon the hypothesis that the ML/DL behavior can be better and more easily characterized on local areas rather than over the entire data set, *e.g.* LIME [36], Anchors [37].

Model Translation methods aim to mimic input/output behavior of the target ML/DL model however considering the whole data set over a symbolic model, *e.g.* Graph-based [38], Rule-based [39].

Intrinsic methods search to integrate the means for explainability as part of the design of the ML/DL model. The explainability of ML/DL networks should be intrinsic and thus input/output behavior should be explicitly justified by the ML/DL model itself. Representative instances in this family are:

Attention Mechanisms rely upon contextual vector and attention mechanisms used to learn a conditional distribution over data inputs which provide an interpretation on the behavior of the weights of the operations of activation and deactivation, *e.g.* Single Modal Weighting [40], Multimodal Interaction [41].

Joint Training consists in introducing an additional task in the ML/DL model, asides the original one, in charge of providing direct or indirect explanations for the main task behavior, *e.g.* Text Explanation [42], Explanation Association [43].

Being a qualitative feature, explainability in turn requires criteria to evaluate the quality of explanations. This presupposes a non-negligible intervention of humans in the assessment process. Some methods proposed to evaluate explanations can be found in [30].

4. Conclusion and perspective

This paper presented the method used in Confiance.ai to tackle the issue of trustworthiness assessment, in the context of safety-critical AI-based systems. Trustworthiness is a complex notion, combining subjective concepts, heterogeneity of granularity in the attributes composing it, and non-commensurability of the different attributes. The approach consists in defining the different attributes constituting the notion of trustworthiness, an exploration of each attribute to determine related KPIs or assessment methods, and the definition of an aggregation methodology based on a MCDA approach. Some of such KPI examples were illustrated in data-driven AI context. The work envisions the creation of a methodological framework for the assessment of trustworthiness that leverages expert knowledge (for example in the definition of thresholds), a modeling of the environment of the application (e.g. influence of the Operational Design Domain on the selection of attributes), and usability in an engineering process (each atomic attribute is linked to a method or metric), covering other AI paradigm in order to go beyond ML.

Acknowledgment

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

[1] J.-L. Adam, M. Adedjouma, P. Aknin, C. Alix, X. Baril, G. Bernard, Y. Bonhomme, B. Braunschweig, L. Cantat, et al., Towards the engineering of trustworthy AI applications for critical systems - The Confiance. ai program, 2022.

[2] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of management review 20 (1995) 709–734.

[3] ISO/IEC 25010, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models, 2011.

[4] J. A. Colquitt, S. A. Brent, L. A. Jeffery, Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance, Journal of applied psychology 92 (2007) 909.

[5] N. I. of Standards, Technology, Us leadership in ai: a plan for federal engagement in developing technical standards and related tools, 2019.

[6] J.-H. Cho, S. Xu, P. M. Hurley, M. Mackay, T. Benjamin, M. Beaumont, Stram: Measuring the trustworthiness of computer-based systems, ACM Computing Surveys (CSUR) 51 (2019) 1–47.

[7] A. Avizienis, J.-C. Laprie, B. Randell, C. Landwehr, Basic concepts and taxonomy of dependable and secure computing, IEEE transactions on dependable and secure computing 1 (2004) 11–33.

[8] M. Pendleton, R. Garcia-Lebron, J.-H. Cho, S. Xu, A survey on systems security metrics, ACM Computing Surveys (CSUR) 49 (2016) 1–35.

[9] A. Ramos, M. Lazar, R. Holanda Filho, J. J. Rodrigues, Model-based quantitative network security metrics: A survey, IEEE Communications Surveys & Tutorials 19 (2017) 2704–2734.

[10] N. Gol Mohammadi, S. Paulus, M. Bishr, et al., Trustworthiness attributes and metrics for engineering trusted internet-based software systems, in: International Conference on Cloud Computing and Services Science, Springer, 2013, pp. 19–35.

[11] L. L. Pipino, Y. W. Lee, R. Y. Wang, Data quality assessment, Communications of the ACM 45 (2002) 211–218.

[12] B. Braunschweig, R. Gelin, F. Terrier, The wall of safety for AI: approaches in the confiance.ai program, in: Proceedings of the Workshop on AI Safety 2022 (SafeAI 2022), volume 3087 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[13] C. Labreuche, A general framework for explaining the results of a multi-attribute preference model, Artificial Intelligence 175 (2011) 1410–1448.

[14] M. Grabisch, C. Labreuche, A decade of application of the Choquet and Sugeno integrals in multi-

criteria decision aid, Annals of Operations Research 175 (2010) 247–286.
[15] L. Pons, I. Ozkaya, Priority quality attributes for engineering AI-enabled systems, arXiv:1911.02912 (2019).
[16] High-Level Expert Group on Artificial Intelligence, Assessment List for Trustworthy Artificial Intelligence (ALTAI), Technical Report, European Commission, 2019.
[17] D. Piorkowski, M. Hind, J. Richards, Quantitative ai risk assessments: Opportunities and challenges, arXiv preprint arXiv:2209.06317 (2022).
[18] A. HLEG, Assessment list for trustworthy artificial intelligence (altai) for self-assessment, High Level Expert Group on Artificial Intelligence. B-1049 Brussels (2020).
[19] J. Mattioli, P.-O. Robic, E. Jesson, Information quality: the cornerstone for AI-based industry 4.0, Procedia Computer Science 201 (2022) 453–460.
[20] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of management information systems 12 (1996) 5–33.
[21] C. A. Bana e Costa, J.-C. Vansnick, A theoretical framework for Measuring Attractiveness by a Categorical Based Evaluation TecHnique (MACBETH), in: Proc. XIth Int. Conf. on MultiCriteria Decision Making, 1994, pp. 15–24.
[22] C. A. Bana e Costa, M. Oliveira, A multicriteria decision analysis model for faculty evaluation, Omega 40 (2012) 424–436.
[23] G. Choquet, Theory of capacities, in: Annales de l'institut Fourier, volume 5, 1954, pp. 131–295.
[24] L. Sun, H. Dong, A. X. Liu, Aggregation functions considering criteria interrelationships in fuzzy multi-criteria decision making: state-of-the-art, IEEE Access 6 (2018) 68104–68136.
[25] C. Batini, M. Scannapieco, Data and Information Quality, Springer International Publishing, 2016.
[26] M. Ge, M. Helfert, A framework to assess decision quality using information quality dimensions., in: ICIQ, 2006, pp. 455–466.
[27] ANSI/ IEEE Std 729-1983, Ieee standard glossary of software engineering terminology, 1983.
[28] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv:1312.6199 (2013).
[29] J. M. Zhang, M. Harman, L. Ma, Y. Liu, Machine learning testing: Survey, landscapes and horizons, IEEE Transactions on Software Engineering 48 (2022) 1–36.
[30] N. Xie, G. Ras, M. van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, CoRR abs/2004.14545 (2020).
[31] D. Erhan, Y. Bengio, A. C. Courville, P. Vincent, Visualizing Higher-Layer Features of a Deep Network, Technical Report, Départment d'Informatique et Recherche Opérationnelle, Univ. of Montreal, 2009.
[32] M. D. Zeiler, R. Fergus, "visualizing and understanding convolutional networks", in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), "Computer Vision – ECCV 2014", Springer International Publishing, 2014, pp. 818–833.
[33] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognition 65 (2017) 211–222.
[34] J. Li, W. Monroe, D. Jurafsky, Understanding neural networks through representation erasure, CoRR abs/1612.08220 (2016).
[35] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3449–3457.
[36] M. T. Ribeiro, S. Singh, C. Guestrin, "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, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144.
[37] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018, p. 1527–1535.
[38] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, S.-C. Zhu, Interpreting cnn knowledge via an explanatory graph, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018, pp. 4454–4463.
[39] M. Harradon, J. Druce, B. E. Ruttenberg, Causal learning and explanation of deep neural networks via autoencoded activations, CoRR abs/1802.00541 (2018).
[40] L. A. Hendricks, Z. Akata, M. Rohrbach, et al., Generating visual explanations, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, 2016, pp. 3–19.
[41] P. Anderson, X. He, C. Buehler, et al., Bottom-up and top-down attention for image captioning and visual question answering, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
[42] H. Liu, Q. Yin, W. Y. Wang, Towards explainable NLP: A generative explanation framework for text classification, CoRR abs/1811.00196 (2018).
[43] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, K. Sycara, Transparency and explanation in deep reinforcement learning neural networks, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, 2018, p. 144–150.