Leveraging Tropical Algebra to Assess Trustworthy AI

Juliette Mattioli[1*], **Martin Gonzalez** [2*], **Lucas Mattioli**[2*],
Karla Quintero[2*], **Henri Sohier**[2*]

[1] Thales, France
[2] IRT SystemX, France
[1] juliette.mattioli@thalesgroup.com - [2] {firstname.name}@irt-systemx.fr

Abstract

Given the complexity of the application domain, the qualitative and quantifiable nature of the concepts involved, the wide heterogeneity and granularity of trustworthy attributes, and in some cases the non-comparability of the latter, assessing the trustworthiness of AI-based systems is a challenging process. In order to overcome these challenges, the Confiance.ai program proposes an innovative solution based on a Multi-Criteria Decision Aiding (MCDA) methodology. This approach involves several stages: framing trustworthiness as a set of well-defined attributes, exploring attributes to determine related Key Performance Indicators (KPI) or metrics, selecting evaluation protocols, and defining a method to aggregate multiple criteria to estimate an overall assessment of trust. This approach is illustrated by applying the RUM methodology (Robustness, Uncertainty, Monitoring) to ML context, while the focus on aggregation methods are based on Tropical Algebra.

How Can We Assess the Trustworthiness of an AI-based System?

Artificial Intelligence (AI) technologies hold potential to improve products and services. However, failures of AI technologies - which can undermine trust in AI technologies and can hinder their use, mainly if they can cause harm and fail to meet the normative expectations of users. Thus, trustworthiness of AI is closely related to *accountability*. Indeed, this property can be seen as a factor of or alternative to trust (O'Neill 2014). However, in (Avizienis, Laprie et al. 2004), *dependability* is used to represent the overall quality measure of a system, based on sub-attributes including *safety*, *reliability* and *maintainability*. Subsequently, *security* and *dependability* became key attribute (Cho, Xu et al. 2019). Moreover, the Assessment List for Trustworthy AI (ALTAI 2019) considers 7 pillars of trustworthiness: 1) Human agency and autonomy, 2) Technical robustness and security, 3) Privacy and data governance, 4) Transparency, 5) Diversity, non-discrimination and fairness, 6) Societal and environmental welfare, 7) Accountability. Regarding risk assessment, the *Confiance.ai* program (www.confiance.ai/en) analyzes that the probabilistic nature of Machine Learning (ML) based systems must require new reliability analysis methodologies to assess the capability of such systems to comply with reliability requirements, as well as novel approaches to assess the dependability of such ML-based systems. This is due to the difficulty of properly defining the environment, context, outputs produced, internal state and associated risks to derive safety objectives and requirements. In addition, this difficulty raises a concern about the dependency between the ML uncertainties and their contribution to the overall level risk. About the resilience, which is one of the major stakes of certification, ML based systems present challenges regarding the definition of an "abnormal behavior", the system architecture on which one can rely to ensure the safe operations of the system, the monitoring of the system at runtime, and the identification of mitigation strategies. This can be due to the usually wider range of possible inputs, the difficulties to adopt classical strategies, and the ML specific vulnerabilities.

The success of AI/ML technology over recent decades has been significantly attributable to the utilization of accuracy-based performance measurements. By assessing task performance based on quantitative accuracy or loss, the process of training AI models becomes one that can be optimized. Conversely, predictive accuracy is frequently employed to demonstrate the superiority of an AI/ML product in comparison to other approaches. However, with the recent proliferation of AI/ML, the limitations of an accuracy-only measurement have become apparent with regard to a number of reliability characteristics, such as robustness. The objective of a trustworthiness assessment is to analyze and characterize the trust expectations associated with the specific objectives in question (Mattioli, Sohier et al. 2024). In conjunction with the Operational Design Domain (ODD) analysis process, it contributes to the definition of the system's observable and measurable conditions and properties.

Trustworthiness Characteristics

While most active academic research on trustworthy ML has focused on the algorithm properties, its systemic analysis has received very little attention (Mattioli, Le Roux et al. 2023). Trustworthiness characteristics need to be mapped onto the AI processes and lifecycle, while keeping track of how they are related to the different stakeholders. Typically, it is determined by the quantification of elementary scores on

[*]These authors contributed equally.

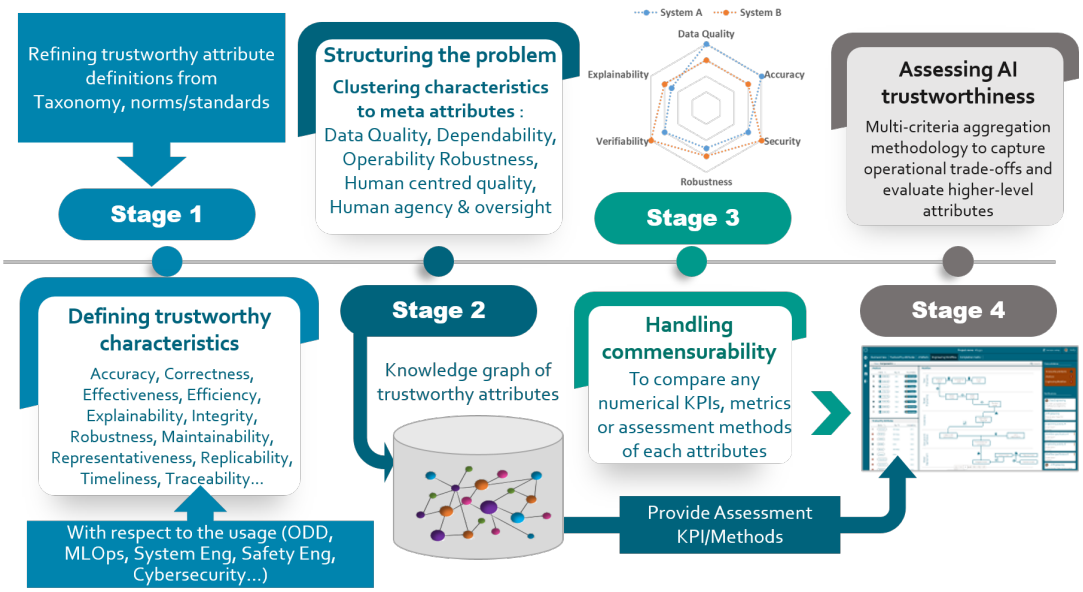

Figure 1: The unified approach to support trustworthy AI assessment

the ML component, *e.g.* for reliability: Fleiss Kappa score, goodness-of-fit tests, or for accuracy: precision, recall, F-score, *etc.*. However, as a system property, such trustworthy concept results from a combination of factors and depends on the context of application. This implies that the relative importance of each attribute can fluctuate depending on the circumstances wherein such system is operating (Braunschweig, Gelin, and Terrier 2022).

Trustworthiness at system level and at model level induces robustness, effectiveness and dependability. They relate to the ability to verify that the AI-based component is valid, effective and has robust intrinsic properties such as accuracy, safety and security. In complex, dynamic and uncertain real-world environments, AI systems should be particularly robust to change. Under all circumstances, AI-based solutions should not harm humans. The autonomy should always be under the control of the user. It is imperative that humans retain the right to grant or revoke AI systems' decision-making authority at any given moment. In particular, AI-based systems should not cease functioning at inappropriate times, for example when a lack of output could lead to safety and cyber security risks. Furthermore, they should be accessible and intuitive for users with diverse backgrounds.

Trustworthiness Assessment Unified Approach

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...) (Adedjouma, Adam et al. 2022). 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" (different benefits for the different stakeholders, different discretization of these benefits, different units, etc). 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 stakeholder preferences, and that the values of the scales for the different criteria should be aggregated.

Multi-criteria Decision Aiding (MCDA) is a scientific field that studies evaluation of a finite number of alternatives based on multiple criteria/attributes through aggregation operators which are mathematical objects that have the function of reducing a set of numbers into a unique representative (or meaningful) number used to compare, evaluate, and rank solutions. The difficulty behind MCDA problem solving is to elicit the expert's know-how. First of all which family of decision models is the best suited to the expert's knowledge? Then, what type of information is needed to determine the parameters of the chosen family of decision models? Most available tools solving MCDA problems, such as ELECTRE (Hashemi, Hajiagha et al. 2016) or MACBETH (Bana e Costa, De Corte, and Vansnick 2016), are based on a specific and restrictive model most often the weighted sum. For the weighted sum, the expert shall compare the importance of criteria. The relevance of the model in the context of the application is never checked out, and the expert may even not be aware of what the restriction on the model may imply for him/her. Consequently, the use of weighted sum may be misleading. To overcome such issue, the Confiance.ai approach (Mattioli, Sohier et al. 2023) is a variation of MCDA, based on the following steps (fig.1).

Step 1: Defining Trustworthiness Attributes: Based on different sources (standards, scientific communications, industrial and institutional reports...), the characterization and evaluation of trust focus on defining and structuring the attributes that constitute trust (Pons and Ozkaya 2019), going beyond a risk analysis as proposed in (Piorkowski, Hind, and Richards 2022). Definitions (what does the trustworthiness characteristics stand for?) adopted in Confiance.ai, motivations (why is the requirement relevant for trustworthiness?) and a short glimpse at assessment methods (how can we assess the level of satisfaction of the requirement?) are given for each of these trustworthy characteristics in their respective sections (Mattioli, Sohier et al. 2024). These various characteristics are categorized in terms of the artifact they describe: dataset or data item, ODD, ML model, AI-based system, etc.

Step 2: Structuring Attributes in a Semantic Tree: Subsequently, the issue of evaluating trustworthiness is broken down into a series of discrete sub-problems through the introduction of a hierarchical framework comprising a substantial number of specific criteria. The objective of this structuring phase is to construct a tree representing a hierarchy of points of view, with the root representing the overall evaluation and the leaves representing atomic attributes with one or more evaluable KPIs. In order to construct such a hierarchy, it is necessary to group the criteria according to a classification system that is meaningful to the stakeholders. Upon completion of this phase, the relevant criteria, along with their hierarchical organization, should be obtained.

Step 3: Adapting Attributes for Commensurability: A numerical evaluation is returned for all atomic attributes. Depending on the use case, specific key performance indicators (KPIs), metrics or evaluation methods are used to qualify the leaves of the tree. For example, data quality is a issue that has been studied for a couple of decades now (Mattioli, Robic, and Jesson 2022), but the focus was primarily on data in operational databases and data warehouses. Now, ML is generating renewed interest in data quality, but there is still limited consensus on what constitutes data quality characteristics. (Wang and Strong 1996) were among the first to argue that limiting quality to the level of accuracy is not enough, emphasizing that the level of quality for given data may depend on its purpose. Standards for the definition of data quality attributes for ML are currently being developed: ISO/IEC CD 5259-1 (terminology and principles) and ISO/IEC CD 5259-2 (data quality measures).

Then, as the KPIs are given in various units and can also be qualitative, they need to be normalized into a satisfaction scale $[0, 1]$ through a transformation canned criterion. That is, you must be able to compare each numerical score of one attribute with each numerical score of another attribute. To make the evaluation "comparable", sound methods of normalization (making comparisons between variables comparable) must be applied to individual variables to first make them comparable, that is, to transform different scales of variables into a single scale. The numerical evaluation of the attributes is therefore coded in the interval $[0, 1]$, where the value 0 corresponds to the total absence of the property under a reliability criterion and the value 1 corresponds to the complete fulfillment of the criterion.

Step 4: Assessing AI Trustworthiness: In order to make an overall assessment of AI trustworthiness, we need to build an aggregation function of the KPIs associated to the various attributes. In general, the most commonly used aggregating function is the weighted arithmetic means, but this assumes that the criteria are independent of each other. Since the criteria often interact (e.g. accuracy vs. robustness), this is a major limitation. So, we need to use a different type of aggregation function based on specific formulas (e.g. min/-max, Choquet integral(Grabisch and Labreuche 2010)...) to aggregate the normalized indicators. If one attribute is more "important" than another in terms of stakeholder preference, the former will be weighted more heavily than the latter in the aggregation procedure. Conflicting criteria have also to be handled by emphasizing the most important criteria while considering the overall impact of all criteria. This aggregating approach developed in this paper is based on Tropical Algebra to provide a balanced multi-attribute aggregator.

Genesis of the RUM Methodology: From ML Model to AI-Component Robustness Evaluation

In accordance with the ethical guidelines for trustworthy AI set forth by the EU, robustness represents a fundamental criterion for the development of reliable AI systems (ALTAI 2019). The IEEE defines the notion of robustness as "*the degree to which a system or component can function correctly in the presence of invalid input or stressful environmental conditions*". The majority of existing research on robustness in AI systems has concentrated on adversarial robustness, that is to say, the capacity to withstand adversarial attacks. However, research into non-adversarial robustness, defined as 'the ability to preserve model performance under naturally induced corruptions or alterations' in model inputs, has also attracted attention. In addition, erroneous outputs may result from deficiencies in algorithmic robustness, defined as the capacity of a machine learning algorithm to preserve its performance under all circumstances, including unexpected inputs, external interference, and harsh environmental conditions. In order to create trustworthy AI systems, **robustness** plays a crucial role. It refers to "*the ability of a system to continue to behave as it intends to behave, and to avoid causing harm, even under difficult or unexpected conditions*". A trustable AI model is resistant to human input errors, malicious attacks, unspecified model objectives, inadequate model training, and non-linearities manifestations. Consequently, assessing robustness is particularly important for high-risk systems before they are made available for users to access. Particularly in cases where lives are at stake, such as in aeronautics, automotive and health care industries, where incorrect decisions made by systems can pose a significant threat to human life. In these cases, it is vital that systems are designed and implemented to withstand input disturbances. So a global robustness KPI to assess ML model robustness should not only take into account various

robustness attributes combined with uncertainty quantification and monitoring measures.

Robustness Attributes

The **robustness** property of an AI-based system relates to the question whether or not the system can be trusted to do well its intended purpose on the envisioned ODD. Mainly this characteristic ensures that the system will keep its performance properties (as accuracy and functional suitability). The ODD is key when defining the robustness of an AI system since the latter requires the definition of the AI system's function as well as the environment in which it will operate.

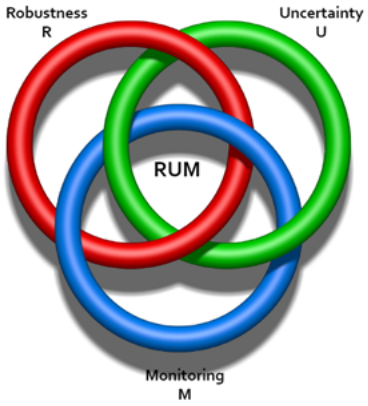

Figure 2: Interrelation between RUM elements

As an additional element to the robustness assessment framework, **uncertainty** is aimed at incorporating the consequences of unobserved, misunderstood, or elusive factors within a modeling scope. Understanding uncertainty is vital for making informed decisions, when deploying models in real-world applications. Applied to an AI/ML system, it primarily aims to express the consequence of uncertainties coming from different sources (noise measurement, irreducible task variability, lack of observation, modeling constraint, among others) on the AI system's output (for a modeling scope link our AI system). It contributes to robustness properties. Several different mathematical tools and approaches, coming from statistics and probabilities domains, could be used to produce uncertainty assessment such as Bayesian methods, dropout during training, ensemble methods, and probabilistic models. In a nutshell, uncertainty assessment aims to address and/or contribute to: (1) Modeling with uncertainty: Formalize the problem into a decision under uncertainty; (2) Express model risk: Express the uncertainty link to the system's decision/prediction by using a model able to produce uncertainty quantification by design or with overlays (model agnostic); (3) Model-Monitoring: by providing insight (as Out of distribution score) that qualifies the irrelevance of the model in answering a query (that may be due to lack of observations); (4) Model-based Data-Monitoring: by finely characterizing the abnormality of an observation by considering uncertainty (as for example task irreducible uncertainty); (5) Model Risk Aversion Assessment: Evaluate (and/or calibrate) the error risk more finely by going beyond the assessment of the average error thanks to a modeling with uncertainty.

Finally, **monitoring** is fundamental to handle uncertainty and failure mitigation since models can degrade over time due to changes in the environment, data, or underlying assumptions. Thus, monitoring ensures early detection of such issues, enabling timely intervention and model maintenance. It involves continuous tracking and evaluation of a ML model's performance and behavior over time. It helps identify potential issues, drifts in data distribution, or changes in model effectiveness. Performance metrics tracking, concept drift detection, anomaly detection, and model explainability methods could be used for ML monitoring. In summary, the following questions arise for a global assessment of the *robustness* of the ML model: (1) Robustness: How resilient is the model's performance when faced with different conditions or perturbations in the input data? (2) Uncertainty Quantification: How well does the model know what it doesn't know? (3) Monitoring: Is the model behaving as expected? Is its performance consistent with the intended objectives?

As monitoring (M), uncertainty quantification (UQ), and robustness(R) are interconnected concepts in ML, Confiance.ai proposes the "**RUM methodology**" that aims at providing means to assess AI component robustness beyond the robustness of the ML model. The following principles underpin the methodology: a joint monitoring of robustness attributes and uncertainty quantities; the robustness joint assessment of monitored observables and uncertainty quantities; and the uncertainty joint quantification of monitored observables and robustness attributes.

The RUM Methodology

The RUM methodology arises from the need to consider at a technical level, the interrelation between Robustness, Uncertainty Quantification and Monitoring. This relationship should be apprehended as displayed intuitively in fig. 2 as three 3D loops topologically linked where any two such loops are linked by a third one.

Robustness, Uncertainty and Monitoring methods can only successfully address their different challenges by working together. At a technical level, the RUM methodology provides means to articulate and characterize different ODD zones to better detect possible failure modes, assess possible trade-offs or overall system-level compensations. This is not possible if robustness, uncertainty quantification or monitoring are considered independently. Ensuring a model's robustness during development, quantifying uncertainties during inference, and monitoring its performance in production are continuous processes. A first architecture of this process is presented in fig. 3.

In this process, the displayed green rectangle articulates the difference between evaluating RUM of the resulting AI component and evaluating the quality of the Robustification protocol itself, the UQ module itself, and the monitoring system itself. Consequently, the action of *Robustness* on the RUM methodology implies evaluating both the robustness improvement and the evaluation of the UQ module - rather than only the central ML model, as well as the evaluation of

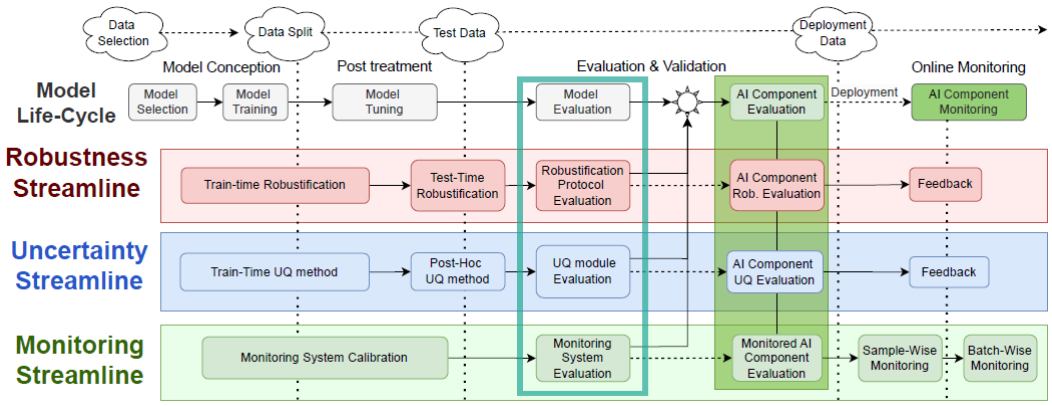

Figure 3: The RUM-process View

the monitoring system.

A holistic approach that integrates robustness, uncertainty quantification, and monitoring is essential for building resilient and trustworthy machine learning systems, particularly in applications where accuracy, reliability, and interpretability are critical. The principle of RUM is the integral consideration of these 3 axes and the means to deploy it will depend on the specific constraints and characteristics of the AI component and context of operation. Specific trade-offs and aggregations stem from the consideration of the RUM method; these trade-offs are then RUM attributes per se.

A Focus on "Robustness Protocol Evaluation"

Various methods can be applied to evaluate the robustness of an AI/ML-system:

- Evasion attacks (Croce and Hein 2020) such as FGSM attack (Fast-Gradient Sign Method) or PGD attack (Projected Gradient Descent), which consist in perturbing the model inputs by including well calculated noises.
- Patch attacks (Brown et al. 2017) which can be printed and positioned in the physical world on an object or in the environment.
- Environmental perturbations: distortions of the inputs which mimic real-world changes occurring on the operational domain (various noises, blur, rain ...)

Confiance.ai's robustness assessment consists on the workflow in Fig. 4, based on:

- **Robustness Evaluation Protocols:** These transform the theoretical evaluation concept into an algorithmic protocol. For instance, crafting adversarial attacks over the test sets realizes the idea of worst-case In-Distribution analysis while Monte-Carlo sampling is used to estimate a theoretically continuous region around a sample to be certified. It usually serves as a design guide to craft defense approaches.
- **Robustification Methods:** These are tightly bounded with the evaluation protocols from which they were crafted, and their goal is to improve that protocol's metrics. For instance, adversarial training is crafted to improve a model's performance in the context of worst-case evaluation (i.e. adversarial attacks).
- **Protocol Analysis:** In addition to the answering of various questions, protocols are themselves implementations of theoretical methods and must be subject to a validation/certification process. For example, it is assumed that the worst-case evaluation is performed with the strongest available adversarial attacks within a predefined cybersecurity breach context. Thus, an overestimation of robust accuracy will be generated due to a poor quality protocol rather than a poor defence mechanism if the protocol only uses weak adversarial attacks against a well-defended model.
- **Defense Analysis:** This goes beyond metric improvement. The process is closely related to protocol analysis. For example, it is a fact that too strong adversarial training overfits and poorly generalises to adversarial threats not seen when being trained. As a consequence, the worst-case analysis protocol needs to be enriched to account for such a strength-generalisation trade-off.

Other works have focused on attacking the AI-system through attacking the UQ module. We can cite (Ledda, Angioni et al. 2023) who focused on a specific adversarial scenario in which the attacker is interested in manipulating the uncertainty estimate, regardless of the correctness of the prediction. Its aim is to undercut the use of UQ techniques for ML models when their results are consumed by a downstream module or by a human. Moreover, ML Watermarking (Kapusta, Mattioli et al. 2024) consist in perturbing the model's behavior by a set of legitimate back-doors in order to enable model identification.

A First Global Assessment Model

The assessment process commences with the identification of the desired outcomes and the resulting effects, with a particular focus on the robustness of the ML-based system. Subsequently, the functions to be performed and the associated operational requirements are determined. In our context, RUM criteria represent robustness requirement. In order to achieve this objective, we propose the use of Tropical Algebra, which allows us to leverage the mathematical properties in order to aggregate a variety of KPIs. To demonstrate this, we will concentrate on the issue of robustness in

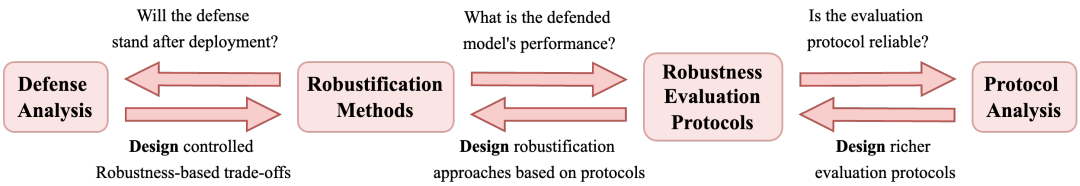

Figure 4: Retroactive action of RUM on Robustness Enhanced Analysis

accordance with the RUM methodology. It is customary to express the function μ in a decomposed manner, namely as $\mu(x) = F(\mu_1(x_1), ..., \mu_n(x_n))$ for all $i \in X$, we have that x_n is a member of the set of values for which we have the function value of the form: The function $X_i \rightarrow [0, 1]$ may be defined as a utility function, which is also referred to as a specific normalized KPI function. The aggregation function, F, is a function that takes as input a vector of values in the interval $[0, 1]^n$ and outputs a value in the interval $[0, 1]$, which is based on Tropical Algebra.

Tropical Algebra Brief Introduction

"Tropical Algebra" is a relatively new mathematics field (Gaubert and Max-Plus 1997). Tropical Algebra is based on $(max, +)$-algebra and $(min, +)$-algebra.

In $(max, +)$-algebra, the addition is replaced by maximum: $a \oplus b = \max(a, b)$, and the multiplication is replaced by addition: $a \otimes b = a + b$. Two null elements are defined: $0_{\oplus} = -\infty$ for $\oplus$, while $0_{\otimes} = 0$ for $\otimes$. Similarly, in $(min, +)$-algebra the tropical sum of two numbers returns the minimum among the two numbers while the tropical product of two numbers returns the sum of those two numbers: $a \oplus b = \min(a, b)$, and $a \otimes b = a + b$.

The tropical operation $\oplus$ and $\otimes$ are associative and commutative, and multiplication is left and right distributive over addition. This algebraic structure differs from classical structures, like fields, in that addition is idempotent: $a \oplus a = a, \forall a$. As with conventional algebra, it is possible to extend the $(max, +)$ (resp. $(min, +)$) algebra to analyze matrices. Indeed, given two matrices A and B with the same dimensions, then: $A \oplus B = C$, where $C_{ij} = A_{ij} \oplus B_{ij}$. If A and B are conformable for multiplication, then $A \otimes B = C$, where $(A \otimes B)_{ij} = \max_k(a_{ik} + b_{kj})$, (resp. $(A \otimes B)_{ij} = \min_k(a_{ik} + b_{kj})$). A complete detailed description and analysis of the mathematics behind such algebra also call "Tropical Algebra" can be found in (Cuninghame-Green 2012; Olsder, Quadrat et al. 1992), a review of the basic concepts in (Gaubert and Max-Plus 1997).

The Tropical Algebra formalism is particularly well suited to multi-attribute aggregation. Concretely, Tropical Algebra:

1. Is adept at handling non-linear aggregation, which is often required in trustworthiness assessment where attributes may not combine linearly.
2. Provides flexibility in modeling complex relationships between trustworthiness attributes, allowing for more accurate modeling of their respective assessment.
3. The operations are computationally efficient, making it suitable for large-scale problems.
4. Can handle a large number of attributes without significant loss of performance.
5. Is robust against variations in KPIs, ensuring stable assessment even under uncertainty.
6. Maintains consistency in aggregation, which is crucial for reliable assessment.

Accordingly, the Tropical Algebra structure $(S, \oplus, \otimes)$ is an idempotent semi-ring (a.k.a dioid (Gondran and Minoux 2008)) with S denoting a set of elements of the concerned trustworthiness dimension. Such aggregation operators satisfy the following axioms: identity when unary, boundary conditions, non decreasing. Besides these basic properties such operators are interesting because they are monotone, symmetric, associative, idempotent (Labreuche 2016). Then aggregate trustworthiness KPIs using Tropical Algebra specifies a global robustness score $\mu_{\mathcal{R}}$ to combine the normalized metrics.

For example, if we have three normalized robustness KPIs μ_{R_i} with $i = 1, 2, 3$, the global robustness score can be computed as: $\mu_{\mathcal{R}} = \mu_{R_1} \oplus \mu_{R_2} \oplus \mu_{R_3}$. Alternatively, we can use a weighted sum approach: $\mu_{\mathcal{R}} = \omega_1 \otimes \mu_{R_1} \oplus \omega_2 \otimes \mu_{R_2} \oplus \omega_3 \otimes \mu_{R_3}$ where $\omega_1, \omega_2, \omega_3$ are weights assigned to each metric based on their importance. The resulting $\mu_{\mathcal{R}}$ will give a single value that represents the overall robustness of the model. A higher $\mu_{\mathcal{R}}$ indicates better robustness. Concretely, the $(max, +)$ algebra is often used for systems where synchronization and timing are crucial, such as train scheduling. Consider a simple example with 3 trains on a single track, where each train needs to wait for the track to be free before proceeding. We denote the times at which each train departs as T_1, T_2 and T_3. The initial departure times are t_1,t_2,and t_3 respectively, and each train might experience a certain delay. Let's assume that the train 1 departs at t_1 with no delay. The train 2 departs at t_2 but can only leave after the train 1 plus a delay d_{12}. Finally, the train 3 departs at t_3 but can only leave after the train 2 plus a delay d_{23}. We can then express the dependencies using $(max, +)$ algebra as follows: $T_1 = t_1$; $T_2 = max(t_2, T_1 + d_{12})$; $T_3 = max(t_3, T_2 + d_{23})$.

Moreover, in the context of safety-critical systems, it is important to evaluate robustness with respect to a range of attacks rather than just against one. This is why we propose to illustrate using a toy case that employs the $(min, +)$ algebra, corresponding to a worst-case aggregation.

Toy Use-Case: MNIST

To illustrate the use of Tropical Algebra in the context of aggregation of metrics, we trained a toy-model for image classification on the MNIST dataset (Deng 2012). Our aim was to answer the following question: given a list of var-

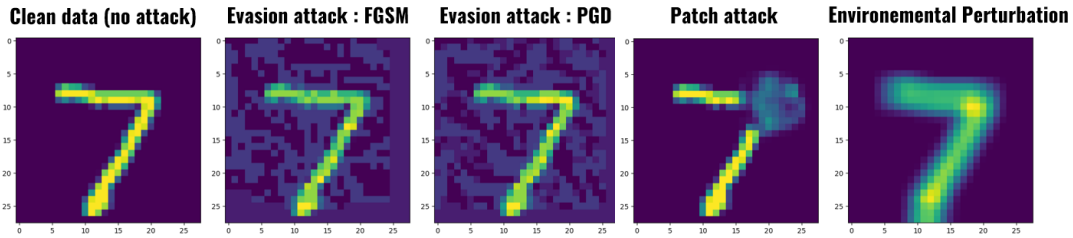

Figure 5: Various perturbations on a single image of the MNIST dataset

ious robustness attributes and prescribed multiple test scenarios, which attribute should be improved in priority for the model? To answer it, we first gathered all the existing robustness metrics of our use-case, then associated to them a number of scenarios prescribed by a choice of weights and make a majority vote to determine which is the leading metric determined by the $(min, +)$ calculus for each scenario.

Firstly, we evaluated it's clean accuracy, then we evaluated the model's accuracy when presented with various attacks and perturbations. Namely, we used a FGSM attack (Goodfellow, Shlens, and Szegedy 2014) (with $\varepsilon = 0.1$) to generate the first attack. For the second attack, we used PGD (with $\varepsilon = 0.1$). For the third perturbation, we used a Gaussian blur (with $\sigma = 3$) and the last one is based on a patch attack (see fig. 5). We obtained the following results:

No attack	FGSM attack	PGD attack	Patch attack	Environmental Perturbation
0.99	0.77	0.44	0.94	0.96

Table 1: Model accuracy (as a probability between 0 and 1) w.r.t. various attacks on MNIST.

The process of selecting weights and thus indicating preferences is a complex one. Therefore, when utilizing weights to represent the relative importance of the objectives, transforming the functions so that they all have similar magnitudes and do not naturally dominate the aggregate objective function can assist in accurately setting the weights to reflect preferences. A sensitivity analysis can be employed to compute such weighting parameters based on the ODD, where the relative importance of RUM criteria in the $(min, +)$ aggregator is represented by the vector $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5)$. The coefficient ω_i is a measure of the importance of criterion i. In order to construct such an operator, it is possible to utilize the two standard levels of 0 and 1. In this context, a value of 0 (respectively, 1) indicates that criterion i has not been met (respectively, has been met in a completely satisfactory manner). The importance of criterion i, denoted by ω_i, is then defined as the added value on the overall score going from the lower level 0 to the upper level 1 on criterion i, with the value on the other criteria being fixed. In order to illustrate the concept, three distinct operational scenarios can be conceived, each emphasizing a particular robustness attribute:

1. Scenario 1: Worst-Case In-Distribution Robustness. In this scenario, we would focus on the model's robustness against the strongest possible adversarial attacks within the distribution (here PGD). Therefore, we could assign the following weights to our metrics: $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) = (0.0, 0.0, -0.1, 0.0, 0.0)$;
2. Scenario 2: Average-Case Out-of-Distribution Robustness. This scenario targets the model's robustness to average-case out-of-distribution (OOD) perturbations. The assigned weights for this scenario could be: $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) = (0.0, 0.0, 0.0, 0.0, -0.25)$.
3. Scenario 3: Worst-Case in-ODD Robustness for deployment. In this final scenario, we concentrate on the model's robustness in the face of the worst-case in-ODD conditions in deployment. In this particular scenario, we could choose the following weights: $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) = (0.0, 0.0, 0.0, -0.2, 0.0)$;

We would thus obtain three different aggregators of the robustness, one for each scenarios:

1. Scenario 1: $\mu_{\mathcal{R}} = \min(0.99 - 0.0, 0.77 - 0.0, 0.44 - 0.1, 0.94 - 0.0, 0.96 - 0.0) = 0.34$;
2. Scenario 2: $\mu_{\mathcal{R}} = \min(0.99 - 0.0, 0.77 - 0.0, 0.44 - 0.0, 0.94 - 0.0, 0.96 - 0.25) = 0.44$.
3. Scenario 3: $\mu_{\mathcal{R}} = \min(0.99 - 0.0, 0.77 - 0.0, 0.44 - 0.0, 0.94 - 0.2, 0.96 - 0.0) = 0.44$.

Here, the majority vote on the aggregation's operator is obvious: - for this toy-case - the PGD attribute is the leading attribute that causes the lack of robustness. This information should be use by the AI scientist to refine the ML design in order to increase the robustness of the model.

Conclusion and Perspectives

This paper presents the method employed by Confiance.ai to address the issue of trustworthiness assessment. The concept of trustworthiness is inherently complex, encompassing subjective perceptions, a diverse range of granular attributes, and a lack of comparability between different attributes. The methodology entails defining the various attributes that comprise trustworthiness, investigating each attribute to ascertain pertinent KPIs or evaluation techniques, and formulating an aggregation methodology based on an MCDA approach. The RUM methodology, as applied to machine learning, exemplifies our approach to the global assessment of robustness, utilising an aggregation operator based on Tropical Algebra.

By defining a hierarchical aggregation operator μ, this approach can then be generalized to all the trust at-

tributes (Mattioli, Sohier et al. 2024):

$$\mu = \bigoplus_j \omega_j \otimes \left(\bigoplus_i \omega_{i,j} \otimes \mu_{i,j} \right)$$

where j represents the 6 macro-attributes: robustness, effectiveness, dependability, usability, human agency and human oversight; and i their respective atomic attribute KPIs. Specification of the weights $\omega_{i,j}$ depends on the ODD which captures the applicative importance and dependency between each attributes. Future work aims at creating a methodological framework for reliability assessment that takes advantage of expert knowledge (e.g. in defining thresholds), modelling the application environment (e.g. the influence of the operational design domain on attribute selection), and ease of use in an engineering process (each atomic attribute is associated with a method or metric), covering other AI paradigms to go beyond ML.

Acknowledgments

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.; et al. 2022. Towards the engineering of trustworthy AI applications for critical systems. Technical report, The Confiance.ai program.

ALTAI. 2019. Assessment List for Trustworthy Artificial Intelligence (ALTAI). Technical report, High-Level Expert Group on Artificial Intelligence, European Commission.

Avizienis, A.; Laprie, J.-C.; et al. 2004. Basic concepts and taxonomy of dependable and secure computing. *IEEE transactions on dependable and secure computing*, 1(1): 11–33.

Bana e Costa, C.; De Corte, J.; and Vansnick, J. 2016. On the mathematical foundations of MACBETH. In *Multiple criteria decision analysis*, 421–463. Springer.

Braunschweig, B.; Gelin, R.; and Terrier, F. 2022. The wall of safety for AI: approaches in the Confiance.ai program. In *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022)*, volume 3087 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Brown, T.; Mané, D.; Roy, A.; et al. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Cho, J.; Xu, S.; et al. 2019. Stram: Measuring the trustworthiness of computer-based systems. *ACM Computing Surveys (CSUR)*, 51(6): 1–47.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Cuninghame-Green, R. 2012. *Minimax algebra*, volume 166. Springer Science & Business Media.

Deng, L. 2012. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6): 141–142.

Gaubert, S.; and Max-Plus. 1997. Methods and applications of (max,+) linear algebra. In *STACS 97: 14th Annual Symposium on Theoretical Aspects of Computer Science*, volume 14, 261–282. Springer.

Gondran, M.; and Minoux, M. 2008. *Graphs, dioids and semirings: new models and algorithms*, volume 41. Springer Science & Business Media.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

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: 247–286.

Hashemi, S.; Hajiagha, S.; et al. 2016. Multicriteria group decision making with ELECTRE III method based on interval-valued intuitionistic fuzzy information. *Applied mathematical modelling*, 40(2): 1554–1564.

Kapusta, K.; Mattioli, L.; et al. 2024. Protecting ownership rights of ML models using watermarking in the light of adversarial attacks. *AI and Ethics*, 4(1): 95–103.

Labreuche, C. 2016. On capacities characterized by two weight vectors. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 16th International Conference, IPMU*, 23–34. Springer.

Ledda, E.; Angioni, D.; et al. 2023. Adversarial Attacks Against Uncertainty Quantification. In *Proceedings of the IEEE/CVF Int. Conference on Computer Vision*, 4599–4608.

Mattioli, J.; Le Roux, X.; et al. 2023. AI engineering to deploy reliable AI in industry. In *5th IEEE Conference on Transdisciplinary AI (TransAI)*, 228–231.

Mattioli, J.; Robic, P.; and Jesson, E. 2022. Information Quality: the cornerstone for AI-based Industry 4.0. *Procedia Computer Science*, 201: 453–460.

Mattioli, J.; Sohier, H.; et al. 2023. Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation. In *SafeAI 2023-The AAAI's Workshop on Artificial Intelligence Safety*, volume 3381.

Mattioli, J.; Sohier, H.; et al. 2024. An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering. *AI and Ethics*, 4(1): 15–25.

Olsder, G.; Quadrat, J.; et al. 1992. Synchronization and Linearity.

O'Neill, O. 2014. Trust, Trustworthiness, and Accountability. In Morris, N.; and Vines, D., eds., *Capital Failure: Rebuilding Trust in Financial Services*, 0. Oxford University Press. ISBN 978-0-19-871222-0.

Piorkowski, D.; Hind, M.; and Richards, J. 2022. Quantitative AI Risk Assessments: Opportunities and Challenges. *arXiv preprint arXiv:2209.06317*.

Pons, L.; and Ozkaya, I. 2019. Priority Quality Attributes for Engineering AI-enabled Systems. *arXiv:1911.02912*.

Wang, R.; and Strong, D. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4): 5–33.