An End-to-End Method for Operationalizing Trustwothiness in AI-Based Critical Systems

Karla Quintero, Lucas Mattioli, Henri Sohier
IRT SystemX, France
email: {karla.quintero, lucas.mattioli, henri.sohier}@irt-systemx.fr

Juliette Mattioli
Thales, France
email: juliette.mattioli@thalesgroup.com

***Abstract*—This work presents one of the products of the Confiance.ai research program which addresses an end-to-end method for engineering trustworthy ML-based systems. The proposed methodology revisits software and systems engineering as it encompasses all development phases of the system while integrating the specificities related to the development of ML-based components within the system. The method leverages vastly researched and deployed standard procedures from design to validation and maintenance in order to provide rigor, structure and traceability when developing ML-models.**

Keywords-Trustworthy AI, safety-critical AI-based systems, end-to-end engineering of AI-based processes, trustworthiness attributes.

I. Introduction

Any technology, even Artificial Intelligence (AI), is developed to provide a service fulfilling some needs. In our context, an AI-based system is defined as a system that incorporates software-based AI components. AI-based critical systems, which can have severe consequences in case of failure, are considered to be "high risk" under the EU AI Act [1]. These systems can for example represent safety components of regulated products which are required to undergo a third-party conformity assessment. Examples of such systems can be found in the fields of transportation, healthcare, defense, and security in general. The deployment of such systems is contingent upon their demonstrated capacity to deliver the anticipated service in a secure manner, while meeting user expectations with regard to quality and continuity of service. Furthermore, users might consider as negative any surprising or unexpected actions from the system.

In order to characterize such systems with a view to quality assurance, [2] proposed considering several dimensions: the artifact type dimension, the process dimension, and the trustworthiness characteristics attributes that are relevant to software product or system quality. In addition, software quality is at the center of the SQuaRE (Systems and Software Quality Requirements and Evaluation) series of standards, and the specific nature of AI is addressed more specifically in order to offer a quality model for AI systems. Consequently, the design of AI-based critical systems necessitates the demonstration of their trustworthiness, as asserted by [3].

Trustworthy AI is based on three components [4], which should be met throughout the system's entire life cycle: firstly, it should be lawful, in that it complies with all applicable laws and regulations; secondly, it should be ethical, ensuring adherence to ethical principles and values; and thirdly, it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm. Thus, to support the industrial design of such systems, there is a requirement for Trustworthy AI Engineering, a new discipline that is an evolving multi-disciplinary field. The aim of this discipline is to ensure that an AI-based critical system (in the safety, mission and business domains) is valid, explainable, resilient, safe, secure, compliant with respect to regulation, standardization, and responsible practices (ethical and sustainable). When dealing with critical systems, several additional constraints must be considered. In the context of system design, there is a need to optimize processes, provide justification, replicate where possible, and implement improvements. However, it is also essential to ensure that the system meets the appropriate level of trustworthiness [5]. This includes robustness (defined as the ability of a system to withstand errors during execution and to cope with erroneous input), cyber-security, and dependability (including reliability, availability, maintainability, and safety properties), among others.

Thus, in the following, we will first remind the today context of AI regulation and standardization as *"trustworthiness is the ability to meet stakeholders' expectations in a verifiable way"*. Then, we present an end-to-end methodology to support "Trustworthy AI Engineering", which encompasses the entire lifecycle of AI-based systems, from Operational Design Domain (ODD) specification to maintenance. This methodology covers data engineering, algorithm design, development, deployment and monitoring. This systematic approach involves organizing multi-disciplinary and fragmented approaches to trusted AI and applying a continuous workflow approach. Measures to improve AI trustworthiness must be taken at every stage, such as data sanitisation, robust algorithms, anomaly monitoring and risk auditing.

II. Regulation and Standardization

Ensuring safety, reliability, availability and maintainability, means AI systems must perform and continue to perform as intended under sufficient conditions. Hazard analysis and risk assessment are tailored to the unique characteristics of AI. These include potential critical errors in training data or knowledge representation, and the ability of the AI model to generalize to unseen, operational data. The performance requirements on the AI algorithm are often driven by safety

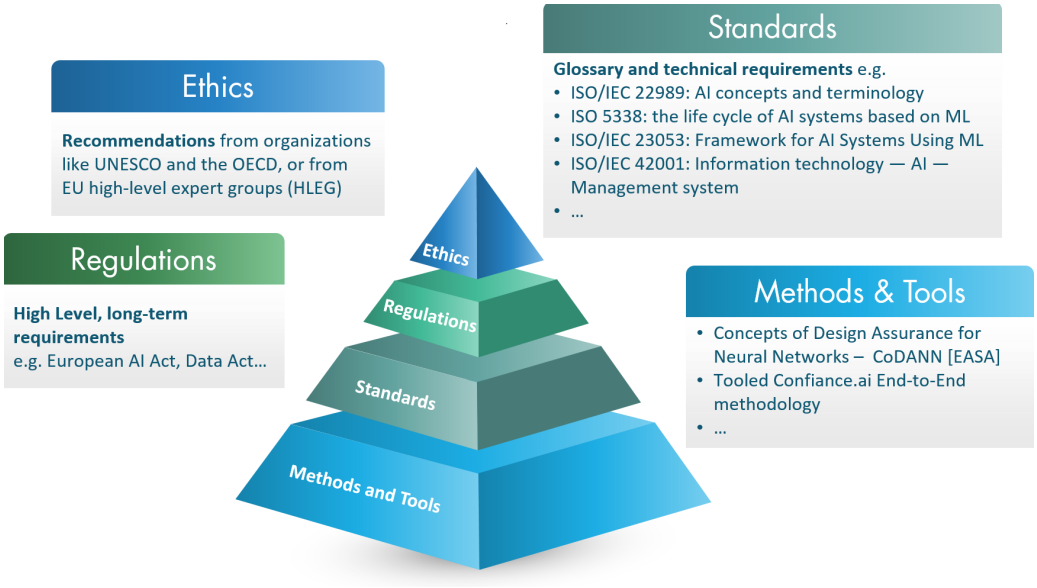

Figure 1. From ethics to the end-to-end methodology through regulation and standards

objectives to limit its worst credible approximation error to a given acceptable threshold.

However, trustworthiness is tightly related to accountability: accountability can be considered as a factor of trust or as an alternative to trust. Then, in [6], dependability is used to represent the overall quality measure of a system based on four sub-attributes including security, safety, reliability, and maintainability. Thereafter, security and dependability became key attributes for computer-based system trust [7].

In 2019, the U.S. National Artificial Intelligence Research and Development Strategic Plan [8] emphasized that: *"standard metrics are needed to define quantifiable measures in order to characterize AI technologies"*. More recently, [9] noted that "*significant work is needed to establish what appropriate metrics should be to assess system performance across attributes for responsible AI and across profiles for particular applications/contexts.*".

Governments are responding with regulations typically associated to human rights. In 2024, the European Union adopted the AI Act. These regulations set high-level, long-term requirements, sometimes building on recommendations from organizations like UNESCO [10] and the OECD [11], [12], or from High-Level Expert Groups (HLEG) [4].

These high-level requirements require to be operationalized for companies and developers. As shown in figure 1, standards and regulation frameworks define more detailed requirements but remain focused on *what* to do rather than *how* to do it, leaving the choice of a tooled end to end methodology to use for the development of AIs fulfilling these requirements.

The Assessment List for Trustworthy AI considers 7 pillars of trustworthiness: 1) human agency and autonomy, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non discrimination and fairness, 6) societal and environmental well-being, 7) accountability. This List is one of the basis of the AI Act [1] which requires companies to take measures to ensure that their products developed or deployed in the European Union are safe and comply with ethical principles.

In the aeronautic domain, EASA [13] proposes a model of trustworthiness based on: the characterization of the Machine Learning (ML) application (high-level function/task, concept of operations, functional analysis, classification of the ML application), safety assessment, information security management, and ethics-based assessment (which includes the 7 pillars of the ALTAI [14]).

The Fraunhofer [15] offered an analysis of the standard [16, Under development] on management system for AI, stating compliance to the standard can contribute to ensuring AI trustworthiness since it encompasses the pillars of the ALTAI, provided that a third-party verification has been performed and along with an adapted quality management system.

In the same period, the NIST produced an analysis of the components of trust [17] and highlighted several top level aspects for the design of a trustworthiness model, that should encompass the user experience, the perceived technical trustworthiness, the pertinence of each trustworthiness characteristic in the user's specific context of use...

Moreover, ETSI set-up in 2019 an Industry Specification Group on Securing AI (ISG SAI) from attack to resilience [18] providing existing and potential mitigation against threats for AI-based systems.

Robust security measures must protect AI systems from cyberattacks, data breaches, and unauthorized manipulation. These measures should include advanced threat detection and mitigation strategies and resilience mechanisms to operate securely in hostile environments. Cybersecurity should be embedded in the system and data pipelines. The lines between security and safety are not always clear when it comes to AI.

Incorrect outputs can be caused by malicious actions or natural events.

Ethical engineering focuses on the need for fairness, transparency, and accountability in AI. This involves ensuring that algorithms are unbiased, produce explainable results, and adhere to societal and legal values. The engineering of such systems requires ongoing review by engineers, ethicists and domain experts.

However, it is imperative to recognize that the transfer of AI technology, particularly Machine Learning (ML), must align with specific standards and processes to ensure the successful transformation of research outcomes into industrial products that are fit for the intended purpose and meet customer needs. For instance, as data collection and analysis are pivotal for the development of any ML-based system, it is essential to prioritize the data quality. This necessitates adherence to compliance regulations (such as data privacy). Concurrently, operational requirements encompassing the maintenance must be addressed. Consequently, it is evident that the development and implementation of AI/ML systems is a multifaceted process involving both technical and business aspects, from problem conception to delivery to customers. Consequently, the development and operation of AI-based critical systems necessitates the utilization of an end-to-end tool-based AI engineering methodology, which will be subsequently delineated.

III. The end-to-end methodology

The version of the methodology presented herein has been produced as a result of the work within the Confiance.ai program [19] [20], [21] and the associated roadmap is nourished by industrial needs and the evolution of the state-of-the-art [22]. Namely, several industrial projects and research initiatives have derived from Confiance.ai, generating the emergence of an ecosystem for the engineering of trustworthy AI for critical systems. The proposed end-to-end methodology addresses the following challenges [23]:

- How can AI/ML models be designed to satisfy trustworthy attributes (explainability, robustness, accuracy, etc.)?
- How can these models allow a clear understanding of their behavior in the operational domain?
- How can AI/ML models be implemented and embedded on hardware, by making them fit to the target without discarding their trustworthy properties?
- Which data engineering methods should be applied to manage large volumes of data and account for the evolving operational domain?
- What kinds of verification, validation, and certification processes should be considered when dealing with AI/ML-based systems?

By addressing these challenges, the end-to-end methodology aims to answer the research question: How to ensure the reliability and trustworthiness of AI-based safety-critical systems? It is based on the premise that the development of ML-based critical systems should be structured with a trustworthiness imperative from the design phase, thereby providing precise requirements for integration, verification, and validation, as well as for proper deployment and maintenance [24]. It is a multi-domain collaboration that leverages concepts and procedures coming from different fields into the agnostic proposal of engineering trustworthy ML-based critical systems. The result is the formalization, through a common language, of the structure and workflow for all actors involved in the process of designing trustworthy ML-based critical systems, i.e. data engineers, systems engineers, safety engineers, software engineers, among many others.

The method addresses as a whole both the system engineering layer and the ML algorithm engineering layer. The system layer accounts for all underlying phases that should design and specify to further along verify and validate the overall system's objective and performance as carried out in classic systems engineering. The ML layer then covers all phases related to the ML component that inherit system requirements to then refined requirements specific to the ML-components to be developed. This process aims to ensure the compliance of the AI/ML components with the overall system requirements and intended purpose.

Developing ML-based systems can be visualized as a "W-shaped" life-cycle (see figure 2). This W-shape can be split into two parts. For AI systems, "intended goal"/"intended purpose" and "intended domain of use" are very high-level requirements that have to be translated into "engineering terms". The engineered "intended domain of use" is called Operational Design Domain (ODD). The ODD is the operational conditions for which an AI system is specified, designed, verified, assessed, operated, and disposed. ML engineering life-cycle begins with defining AI/ML algorithm requirements refined from system specification. This ML specification step includes the characterization of the ODD.

This engineering activity is a critical step that changes the way AI researchers and engineers work. It involves a detailed description of all possible operating conditions, called the system operating environment, to enable data collection and knowledge representation. The reliability of the AI-based system depends on the correctness and completeness of this description, particularly for rare events or combinations of conditions that could be unsafe. A system's validity is established by its intended use. The ODD description is developed using a combination of top-down and bottom-up approaches. ODD aligns data and functional intent, i.e. the data used for training and the resulting ML model(s) with their intended use, covering a wide range of conditions.

Data engineering is key. It involves the identification, collection, preprocessing and extraction of features from large datasets. These datasets are essential for designing and verifying ML models. This phase often involves advanced techniques. These techniques improve the representativeness, completeness and relevance of the dataset (minimizing the simulation-to-reality gap). Rigorous quality controls, guided by Data Quality Requirements (DQRs), ensure data inputs are accurate and consistent. During model design, engineers select appropriate learning algorithms and improve model architectures through training and evaluation cycles. Optimization strategies balance

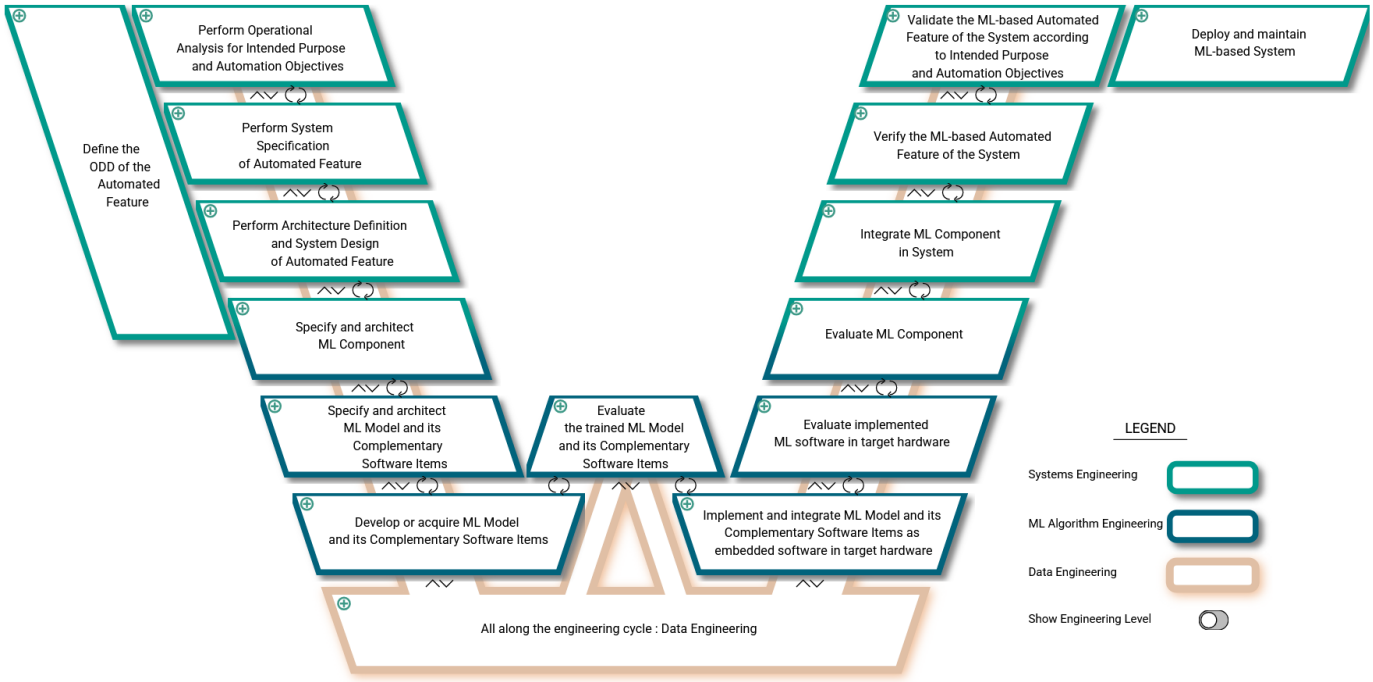

Figure 2. High-level view of the end-to-end methodology.

computational efficiency and performance.

The second "V" of the "W-shaped" life-cycle includes the implementation engineering processes performed on the target platform (e.g., specific hardware embedded in a ground or aerial vehicle). Validation and verification activities are driven by key trustworthiness properties, specified in low-level ML requirements. Validation activities ensure the correctness and completeness of ML requirements by verifying, analyzing and tracing them back to higher-level requirements. Verification activities include simulating extensively, testing edge/corner robustness, scenario-based testing, analyzing the ML model explainability and ODD coverage analysis. The first level of verification ends with a selected AI model, which meets all its requirements in the development (learning) environment and serves as a design specification, ready for implementation into software and/or complex electronic hardware elements in the second level of verification. figure 3 shows a high-level view of the verification phase of an ML-based automated feature and the interaction with specification and validation phases.

MLOps, or Machine Learning Operations, and AM/ML Engineering, while closely related, serve distinct roles within the machine learning lifecycle. MLOps focuses on the operationalization of machine learning models, ensuring that they are deployed efficiently and maintained effectively in production environments. In contrast, ML Engineering is primarily concerned with the development and the maintenance of an AI-based system. Thus MLOps emphasizes the operational aspects of machine learning, while ML/AI Engineering is centered on the overall lifecycle of the system covering all system engineering concerns (from specification to maintenance) which includes MLOps. MLOps involves collaboration between data scientists, ML engineers, and IT operations teams when AI/ML Engineering involves system ad software engineers, data scientists, safety and cyber-security engineers. The end-to-end methodology (see figure 2) supports all AI/ML engineering activities where MLOps covers ML algorithm engineering and data engineering.

IV. Trustworthiness Attributes and Assessment

Trustworthiness is fundamental for the successful development and adoption of AI-based critical systems. Thus, trustworthiness assessment [25] can be defined as the process of evaluating and determining the level of trustworthiness of a given characteristic, such as robustness [26], accuracy, reliability [20], or effectiveness, in the context of AI systems engineering.

Nevertheless, it is very misleading to only judge how good an AI system is based on how accurate it is. It is also difficult to test and check the quality of software in the traditional way, and it is even difficult to measure test coverage at all. Trust and trustworthiness are complex, and so one of the main issues we face is to establish objective attributes such as accountability, accuracy, controllability, correctness, data quality, reliability, resilience, robustness, safety, security, transparency, explainability, fairness, privacy, and compliance with regulatory actors. We need to map these attributes onto the AI processes and its lifecycle and provide methods and tools to assess them. This highlights the importance of quality requirements, which are non-functional requirements and are particularly challenging in AI systems, although many of them can be considered in any critical system. Furthermore, this can also include risk and process considerations. The attributes

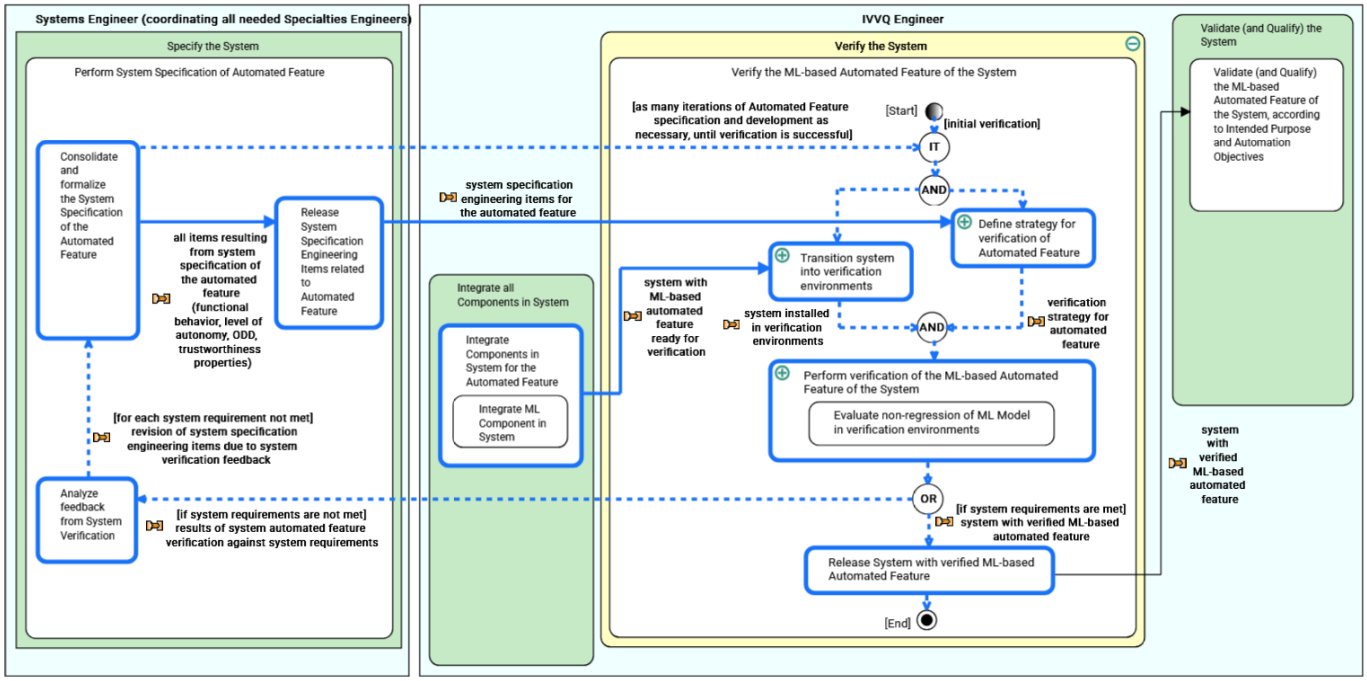

Figure 3. Verification Phase: verification of the ML-based automated feature of the system.

and values for these requirements depend on things like how important the application is, what the AI system is used for, how it will be used, and the people involved. So, in some situations, some attributes may be more important than others, and new attributes may be added to the list [27]. Clear specifications of the non-functional requirements will help clarify these conflicts and can also encourage innovation that solves some of these conflicts, allowing us to fulfill more of them at the same time.

Thus by leveraging system engineering best-practices, ML development workflows, and testing procedures, the end-to-end methodology ensures that trustworthiness attributes are embedded in every stage of the AI system life-cycle, from conception to maintenance. The Confiance.ai framework focuses on the following attributes:

- **Robustness**. Robust AI systems should be resilient to various perturbations (ie: variations in input data and operating conditions). This requires :
  - Adversarial robustness, ensuring the system is not easily manipulable by adversarial attacks.
  - OOD Robustness (Out-Of Distribution), the system must generalize well across different environment and be trained on diverse datasets.
  - Model monitoring, ensuring a continuous evaluation of the AI models, to detect performance degradation.

  Two types of strategies for robustness by design can be distinguished: empirical robustness and formal robustness.
  - Empirical methods emphasize on uncertainty quantification and adversarial robustness of ML Models, like the adversarial training method.
  - Formal methods aim to design neural networks with exact robustness guarantee such that, under some constraints on the norm of the perturbation added to the input, the class of the input remains the same for the ML Model. Lipschitz method is one example of formal methods advocated as enablers for robustness by design.
- **Explainability, Interpretability and Comprehensibility**. Trustworthy AI should be transparent and its decisions should be interpretable where
  - Explainability deals with the capability to provide the human with relevant information on how an AI application is coming to its result.
  - Interpretability relates to the capability of an element representation (an object, a relation, a property...) to be associated with the mental model of a human being. It is a basic requirement for an explanation.
  - Comprehensibility refers to the capability of an element representation (an object, a relation, a property...) to be understood by a person according to its level of expertise or background knowledge.

  This requires:
  - Post-hoc explainability tools, to provide insights into model decisions.
  - Model simplification strategies to enhance interpretability.
  - Human-in-the-loop validation to ensure AI decisions align with expert knowledge.

  There is a profusion of methods, tools, and solutions available, each with its own set of advantages, drawbacks, and trade-offs [28]. The many different approaches show how tricky it is to make sure that AI and machine learning models can explain their predictions and decisions. Choosing the

right way to make models explainable is a technical and strategic decision. It depends on the unique needs and limits of the people it will be used by, the specific example it will be used for, and the wider situation in which the AI system will be used. What works for a medical diagnosis model may not work for the aeronautic domain, and what regulators expect can be very different from what end-users or business stakeholders expect. The Confiance.ai program provides a "Methodological Guideline for Explainability" (https://catalog.confiance.ai/) which is designed to be a complete guide to help people use AI. It will explain why explainability is important, highlight the many available methods, and offer guidance on selecting the most suitable approach based on the specific situation.

- **Fairness and Bias Mitigation**. AI models should be free from discriminatory biases. This involves:
  - Bias detection and correction techniques, in the data processing and model training phases.
  - Regulatory alignment with fairness standards (eg: GDPR, AI Act).
- **Safety and Security**. An AI-based system must meet rigorous safety and security requirements:
  - Safety analysis and certification based on standards.
  - Cybersecurity counter-measures, integrated on the AI pipeline.

The end-to-end methodology integrates those attributes throughout the AI system life-cycle, namely in:

- **Operational Design Domain (ODD) definition**
  - Define the operational boundaries where the AI system is expected to function reliably.
  - Establish clear environmental constraints for the AI-system's development.

  The ODD is a description of measurable foreseeable operating conditions within which a system/component shall operate. A traceability property shall be assured between the different levels of ODD (system, subsystem or component).
- **Systems Engineering**
  - Ensure AI system-level requirements are defined in alignment with overall system objectives.
  - Align AI-based system requirements with preexisting system engineering standards and certification guidelines.
- **Data Engineering**
  - Rely on a robust data pipeline to guarantee data integrity, consistency, and traceability across the engineering cycle.
  - Implement bias mitigation strategies at the data collection and processing stages.
  - Use adaptive data augmentation strategies to improve data diversity and model generalization to distribution shifts and operational scenarios.
- **ML Algorithm Engineering**
  - Use ML robustness techniques, designed to handle perturbation and adversarial outputs.
  - Incorporate explanability techniques to have understandable decisions.
  - Apply Uncertainty quantification techniques to asses the model's confidence.
- **Verification and Validation**
  - Perform extensive simulation-based testing to asses performances under edge cases.

In addition, measuring how trustworthy AI systems are is tricky. The ideas behind them are complicated, the characteristics they produce are different, and you can't always compare them. The Confiance.ai program proposes an innovative way to measure trustworthiness using (max,+) algebra [29] based on a complete hierarchical model that brings together different properties, such as how strong, effective, dependable, easy to use and human agency, and human oversight) into a single assessment method. This offers advantages over traditional weighted averaging methods by better handling extreme values and preserving sensitivity to critical indicators, while maintaining sensitivity to critical indicators to provide detailed, understandable assessments of AI-based system trustworthiness.

V. Conclusion and future works

The Confiance.ai program has evolved since its kick-off in 2021, with a first year dedicated to covering the academic and industrial state of the art related to ML-based system design. Subsequent years (2022-2023) were dedicated to the accurate characterization of industrial use cases, the development and evaluation of technological components to address specific aspects of reliability, and the construction of an end-to-end method revisiting all stages of the engineering cycle for the design, integration, and evaluation of ML components. The last year (2024) encompasses the evaluation of this end-to-end method, the completion and dissemination of key results, and the guarantee of their continuation and sustainability under the aegis of a new research initiative currently under construction. To facilitate the adoption of the tool-based methodology by industry, several implementations of the 2023 version have been carried out on use cases.

These experiments have demonstrated the importance of integrating diverse tools and methods to address expectations regarding trusted ownership, as illustrated by the following two examples: In a use case involving autonomous driving, the analysis of dataset diversity reveals a limited presence of night-time images, prompting the generation of synthetic night-time data. This data exhibits a 'domain gap' and undergoes "domain adaptation" prior to integration into the model training data. These tools, instrumental in the construction of data sets, will also be reused in the supervision stage of the use case. In an aeronautical use case called LARD for "Landing Approach Runway Detection" [30] and represented figure 4, a data quality supervision module is incorporated to consolidate the confidence score of an ML model (see figure 4). In this example, local image quality estimators (e.g. level of blur, brightness) are taken into account in the detection zone of the landing strip that is being detected. The combination of these indicators with the other indicators intrinsic to the model facilitates the establishment of a level of confidence for

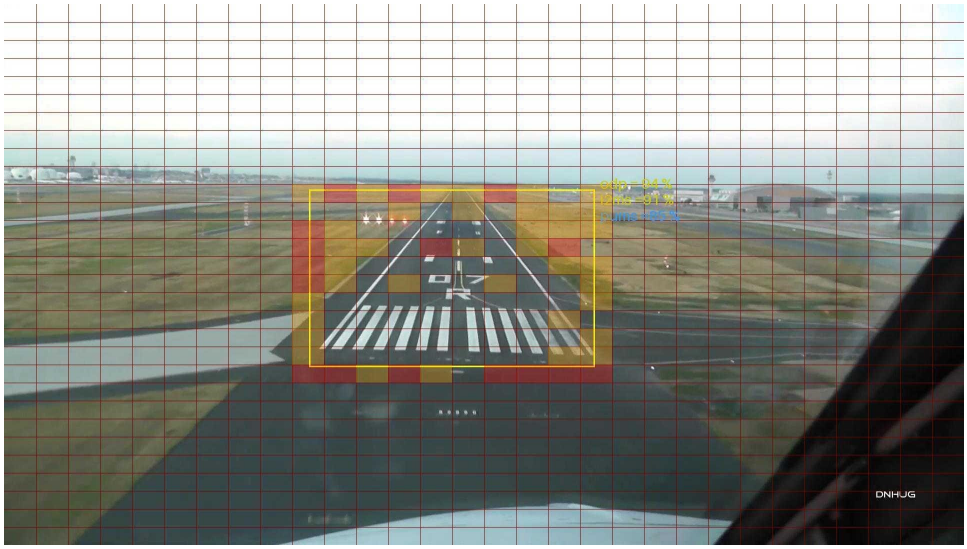

Figure 4. Example of the implementation of a supervision tool on the LARD use-case

the system component. In addition to providing a numerical value, this implementation serves as a tool to facilitate the interpretation of model and data errors.

The Confiance.ai program is opening up two major outcomes to the community as a "digital common good". First, it provides a body of knowledge describing an end-to-end method of AI engineering. This makes it possible to characterize and qualify the trustworthiness of a data-driven AI system and integrate it into industrial products and services. Second, this method is applicable to any sector of activity. A catalog of developed and/or mature technological components to increase the level of trust in AI integrated into critical systems.

The Body of Knowledge (BoK) is one of the main outcomes because it provides access to a navigable version of this end-to-end' methodology that covers the activities structuring the engineering cycle of a critical system based on ML (https://bok.Confiance.ai/). This compendium of expertise from multiple disciplines is a corpus that articulates the system level with the model and data levels in the engineering process. It is continuously updated and expanded and is expected to continue beyond the program. The content provided in the body of knowledge is structured with an end-to-end engineering method in mind and can be navigated through different roles in this process, namely through the field of application of different engineering profiles: These roles include, but are not limited to, the following: machine learning (ML) algorithm engineer, data engineer, embedded software engineer, IVVQ (Integration, Validation, Verification and Qualification) engineer or system engineer.

The following simplified high-level view of the BoK is presented as a gateway to the end-to-end method for engineering trustworthy ML-based systems.The body of knowledge presents the stages of the methodology, from operational analysis and specification of the function of the system that one wishes to automate through the use of ML technology, to verification/validation/qualification, including the development and implementation of the ML model. The navigation through each stage and according to each role facilitates the visualization of the activities, sub-activities and workflow to be carried out when developing a reliable ML-based system.This corpus is thus a compendium of expertise from multiple disciplines because it links the system level with the model and data levels in the engineering process.It is continuously updated and expanded, and this is planned beyond the program.

The catalog (https://catalog.Confiance.ai/) is a web application that allows users to consult the results of the Confiance.ai program. It employs filtering and search functions (sorting, categories, etc.) to facilitate navigation through the various results, which can be either documents or software. Results categorized as 'documentary' are exclusively of a literary nature, including reports (studies or benchmarks), state of the art, doctoral theses or good practice guides.'Software' results are components intended to be run directly or through another application, such as a web application, a library, a plugin or a binary executable.

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] European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, 2021.

[2] M. Felderer and R. Ramler, "Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session)", in *International Conference on Software Quality*, Springer, 2021, pp. 33–42.

[3] H. Liu *et al.*, "Trustworthy AI: A computational perspective", *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 1, pp. 1–59, 2022.

[4] HLEG, *A definition of AI: Main capabilities and scientific disciplines*, Definition developed for the purpose of the deliverables of the High-Level Expert Group on AI, 2018.

[5] M. Adedjouma *et al.*, "Engineering dependable ai systems", in *2022 17th Annual System of Systems Engineering Conference (SOSE)*, IEEE, 2022, pp. 458–463.

[6] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing", *IEEE transactions on dependable and secure computing*, vol. 1, no. 1, pp. 11–33, 2004.

[7] J.-H. Cho *et al.*, "Stram: Measuring the trustworthiness of computer-based systems", *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–47, 2019.

[8] NSTC, *The national artificial intelligence research and development strategic plan: 2019 update*. National Science and Technology Council (US), 2019.

[9] E. Schmidt *et al.*, "National security commission on artificial intelligence (ai)", National Security Commission on Artificial Intelligence, Tech. Rep., 2021.

[10] UNESCO, "Recommendation on the Ethics of Artificial Intelligence", Tech. Rep. SHS/BIO/PI/2021/1, 2022.

[11] OECD, "Recommendation of the Council on Artificial Intelligence", Legal Instruments, May 2019.

[12] OCDE, *G7 Hiroshima Process on Generative Artificial Intelligence (AI)*. 2023, p. 37. DOI: https://doi.org/https://doi.org/10.1787/bf3c0c60-en.

[13] EASA, *Concept Paper First Usable Guidance for Level 1 Machine Learning Applications*, 2021.

[14] P. Ala-Pietilä *et al.*, *The assessment list for trustworthy artificial intelligence (ALTAI)*. European Commission, 2020.

[15] M. Mock *et al.*, "Management system support for trustworthy artificial intelligence", 2021.

[16] ISO/IEC DIS 42001, *Information technology — Artificial intelligence — Management system*, 2022.

[17] B. Stanton, T. Jensen, *et al.*, "Trust and artificial intelligence", *preprint*, vol. 10, 2021.

[18] ETSI, *Securing Artificial Intelligence (SAI); Mitigation Strategy Report*, 2021.

[19] B. Braunschweig, R. Gelin, and F. Terrier, "The wall of safety for ai: Approaches in the confiance. ai program", in *Workshop on Artificial Intelligence Safety (SAFEAI)*, 2022.

[20] J. Mattioli *et al.*, "AI engineering to deploy reliable AI in industry", in *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, IEEE, 2023, pp. 228–231.

[21] R. Gelin, "Confiance. ai program software engineering for a trustworthy ai", in *Producing Artificial Intelligent Systems: The Roles of Benchmarking, Standardisation and Certification*, Springer, 2024, pp. 11–29.

[22] A. Awadid *et al.*, "AI Systems Trustworthiness Assessment: State of the Art", in *Workshop on Model-based System Engineering and Artificial Intelligence-MBSE-AI Integration 2024*, 2024.

[23] A. Awadid, X. Le Roux, B. Robert, M. Adedjouma, and E. Jenn, "Ensuring the reliability of ai systems through methodological processes", in *2024 IEEE 24th International Conference on Software Quality, Reliability and Security (QRS)*, IEEE, 2024, pp. 139–146.

[24] A. Awadid, B. Robert, and B. Langlois, "Mbse to support engineering of trustworthy ai-based critical systems", in *12th International Conference on Model-Based Software and Systems Engineering*, 2024.

[25] B. Braunschweig *et al.*, "AITA: AI trustworthiness assessment: AAAI spring symposium 2023", *AI and Ethics*, vol. 4, no. 1, pp. 1–3, 2024.

[26] K. Kapusta, L. Mattioli, B. Addad, and M. Lansari, "Protecting ownership rights of ml models using watermarking in the light of adversarial attacks", *AI and Ethics*, vol. 4, no. 1, pp. 95–103, 2024.

[27] J. Mattioli *et al.*, "An overview of key trustworthiness attributes and kpis for trusted ml-based systems engineering", *AI and Ethics*, vol. 4, no. 1, pp. 15–25, 2024.

[28] S. Naveed, G. Stevens, and D. Robin-Kern, "An overview of the empirical evaluation of explainable ai (xai): A comprehensive guideline for user-centered evaluation in xai", *Applied Sciences*, vol. 14, no. 23, p. 11 288, 2024.

[29] J. Mattioli, M. Gonzalez, L. Mattioli, K. Quintero, and H. Sohier, "Leveraging tropical algebra to assess trustworthy ai", in *Proceedings of the AAAI Symposium Series*, vol. 4, 2024, pp. 81–88.

[30] M. Ducoffe *et al.*, "Lard–landing approach runway detection–dataset for vision based landing", *arXiv preprint arXiv:2304.09938*, 2023.