Towards the engineering of trustworthy AI applications for critical systems - The Confiance.ai program - Second Edition

Michel Morvan

Towards the engineering of trustworthy AI applications for critical systems / Second Edition

The Confiance.ai program

September 2024

Table of contents

Confiance.ai

Executive Summary

This white paper sums up the journey and findings of the program Confiance.ai, the technological pillar of the Grand Défi "Securing, certifying and enhancing the reliability of systems based on artificial intelligence" launched by the Innovation Council of the French Administration. The two other pillars of this state initiative focus on standardization (norms, standards and regulation toward certification) and application evaluation.

The active collaboration of over 50 partners including large-scale multi-sector industrial partners and research centers, for over four years, has addressed numerous challenges on the topic of engineering Trustworthy AI for critical systems as it aimed at the convergence of solvability of current industrial challenges and applicability of innovative research developments.

As the largest yet technological research program in the national AI strategy, Confiance.ai began in 2021 by a first year dedicated to covering the state of the art and pre-existing tools related to the integration and evaluation of data-driven AI. The following years focused on characterizing industrial use cases, developing technological components for assessing trustworthiness, and constructing numerous guidelines and an End-to-End method for the trustworthy design, integration, and evaluation of Machine learning (ML) components.

The previous white paper in 2022, provided initial results of the program including the first steps toward engineering trustworthy AI, use cases, a first version of a pipeline, a taxonomy and key attributes to characterize AI trustworthiness. As the program evolved, so did the initiatives toward regulation including the AI Act, making the program a bidirectional partaker on the process; this is: ensuring the production of methodological guidelines and digital components that incorporate state of the art developments and envisaged European constraints, as well as the contributing to these initiatives technologically, methodologically and in support of standards. As the AI Act, another element of rising interest during the course of the program is the topic of generative AI. Even though the subject itself was not addressed in the program, some results still hold in this field (e.g. image generation through diffusion models and experience on an NLP use case) and motivate the pursuit of this research in the initiatives ensuring the continuation of Confiance.ai.

This document is organized as follows:

- A first chapter for revisiting of the needs for trustworthy AI in critical systems through the user's lens as well as the challenges beyond the user;
- A glance on the two main gateways to Confiance.ai results: *the body of knowledge and the catalog;*
- The *End-to-End* methodology with a special focus on subjects related to the Operational Design Domain, the Intended Purpose and Assurance Cases;
- The trustworthy environment and Functional Sets with focus on 'Robustness', 'Data Lifecycle', and 'Explainability';
- The deployment of the End-to-End Method on use cases.

As a pioneer on engineering trustworthy AI, Confiance.ai presents in this document an overview of some of the results of this 4-year journey as a gateway for further exploration by both industry professionals and academic researchers. Additionally, the document presents the resulting initiatives inspired by the program which ensure the continuity of this work. ■

Fiction

Valenciennes, October 11, 2029: accident at the Pharma4.0 factory, a worker severely injured in the wrist. From our special correspondent.

An accident that will leave its mark. Yesterday morning, Mrs. D., an employee of the Pharma4.0 factory in Valenciennes, had her right wrist broken by an InCobot handling robot during an ordinary operation that until now had never caused any problems.

In this factory, the operation called "pick and place" of cough syrup bottles is performed jointly by human operators and robotic arms in the same work area, and this on many stations. Yesterday, one of the robots violently hit Mrs. D.'s right wrist during a routine operation, which caused the immediate stop of the line and a protest movement of all the workers, who did not return to work this morning. When asked, a trade union representative declared: “we don’t want to work again with these AI robots, we don’t trust them anymore”.

The cause of this accident can be traced back to the training method of the artificial vision device that equips the InCobot robotic arm. This arm, which weighs about 50 kilos, is equipped with a camera that observes its environment shared with the human operators, and detects the presence of a human hand nearby. The presence of a hand in the field of vision interrupts the movement of the robot, which waits to act until the space is free. The camera sends its video stream to a system trained by machine learning. This system is based on the generic "YOLO" (You Only Look Once) technology, widely used in computer vision, a neural network trained to recognize everyday objects, whose designers emphasize its generic character, and which is specialized by "transfer learning" by providing it with complementary images of the specific objects that one wishes to recognize.

In this case, Pharma4.0 had provided InCobot with images taken on the line containing numerous hand positions in all possible configurations, as well as those of hands protected by blue or pink gloves, as some operators found this more comfortable. The InCobots robotic arm was therefore able to recognize both bare hands and those equipped with these gloves. Unfortunately, yesterday, Mrs. D. was using yellow gloves that she had brought from home. She did not know that the system had not been calibrated for this type of equipment. When Pharma4.0 sent the training images to inCobot, the message indicated that the workers could wear gloves, but only images of pink or blue gloves were present in the transferred database.

The instructions posted in the factory lobby recommend the use of gloves provided by Pharma4.0, but without specifying a particular color. And so, the robotic arm, which had not "learned" to recognize yellow gloves, totally ignored the presence of Mrs. D's hand, which led to the accident we report.

Of course, one lesson to be learned is that it is absolutely necessary to perform a precise risk analysis integrating all possible context use and from that to monitor the system to deal with all of them and detect possible situations escaping from this operating domain. And obviously, that the artificial intelligence systems have been trained and validated with data representing all the operational conditions that may be encountered.

One can also ask the question of responsibility for this accident: was it Mrs. D., who was wearing "non-recommended" gloves but who could not have knowFn that this was a source of danger? Was it InCobot, the supplier of the robotic arm, who did not "program" its equipment well enough? Was it Pharma4.0, who commissioned the robot in the plant and did not provide training images for this situation (which could not easily be imagined, since the company provides gloves to the operators)? Was it the designers of the YOLO system, which was not as generic as they claim in their application document? In fact, this raises the crucial question of clear and specifications of AI systems from which the responsibility of all stakeholders will be clearly defined.
Moreover, the global issue of trust in AI applications is raised in this fictional example. If workers and, more generally speaking, users of AI applications do not have trust in these systems, they will reject them, despite millions of euros invested in their development. ■

Acronyms

List of Abbreviations:

AI	Artificial Intelligence
BPMN	Business Process Modeling Notation
FS	Functional Set
GDPR	General Data Protection Regulation
GSN	Goal Structuring Notation
HCI	Human Computer Interaction
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
IVVQ	Integration Verification Validation Qualification
KPI	Key Performance Indicator
ML	Machine Learning
ODD	Operational Design Domain
RUM	Robustness, Uncertainty quantification and Monitoring
UC	Use Case
UML	Unified Modeling Language
UX	User Experience
V&V	Verification and Validation
XAI	eXplainable AI

1. AI Risks, Trustworthiness and its Attributes

The development and adoption of AI are accompanied by an urgent need: to ensure reliability and trustworthiness in these systems. This chapter is dedicated to the imperative of a trustworthy AI, highlighting the risks associated with "untrustworthy" AI and the potential serious consequences. It explores the challenges and risks related to the adoption of AI, examines the concept of trust in AI, and explores user perspectives regarding interaction with these systems.

1.1 Challenges and Risks in AI Adoption

Trustworthiness is essential to ensure the adoption of Artificial Intelligence (AI) by users, regulators, and safety and quality managers. By rejecting AI, people fail to leverage its benefits, such as optimizing processes, improving decision accuracy and stimulating innovation. In the case of critical systems, the stakes are considerable, and so are the associated risks. This section presents some of these risks.

User-Related Risks

Verifiability and transparency are major considerations, especially when it comes to understanding how AI-based systems make decisions. This is particularly important in sectors where these decisions can have significant consequences, such as healthcare or justice. It is essential that AI decision-making processes are transparent enough to be understood and evaluated by users and stakeholders. For example, insufficient verifiability and transparency in AI-based systems, such as in the algorithms used for credit decisions, could lead to discrimination and losses of customer trust. It is therefore crucial to build trust and ensure that the decisions taken by AI are fair and ethical.

Data-related risks concern the quality, integrity, privacy, security, and management of the data used by AI. Inaccurate, incomplete or biased data can lead to incorrect or unfair decisions. In addition, data security is paramount to protect sensitive information from breaches and cyberattacks. Companies need to implement robust strategies to ensure data quality and security, while minimizing potential biases to improve the reliability and fairness of AI systems.

Beyond the User

Compliance refers to respect for existing laws and regulations. With the rapid development of AI, many jurisdictions are developing specific rules to govern its use, particularly in sensitive areas such as facial recognition or the collection of personal data. For example, in terms of regulatory compliance, a company using AI to process personal data must follow the appropriate protocols not to breach the General Data Protection Regulation (GDPR), and new regulations such as the AI Act. Companies therefore need to be vigilant in complying with these regulations to avoid legal penalties.

Secondly, *governance* is about how organizations manage and oversee their AI systems. This includes establishing internal policies or managing AI-related risks. A lack of adequate oversight of AI systems could lead to critical errors.

Finally, *ethical considerations* are crucial. They encompass transparency, fairness of algorithms, privacy and the social impact of AI. Companies must ensure that their AI systems do not perpetuate existing biases and respect the fundamental rights of individuals, while being aware of the overall societal impact of their technologies.

To tackle these challenges and minimize the risks associated with the adoption of AI, it is necessary to develop trustworthy AI and more specifically to define trustworthy AI characteristics. This is the subject of the next section.

1.2 Understanding Trust in AI

As discussed in the previous section, critical AI-based systems can present risks that require careful monitoring. Therefore, it is crucial to evaluate these systems according to specific criteria so that they can be qualified as '**Trustworthy AI**'.

Trustworthy AI can be represented as a set of six higher-level requirements (see Figure 1): robustness; effectiveness; dependability (including safety and security), usability, human agency (including transparency, interpretability and explainability) and human oversight (including ethical issues). Trustworthiness does not concern only the system itself, but also other actors and processes that play their part during the AI lifecycle (engineers, operators, certification authorities, insurance companies...). Trustworthy AI characteristics can be defined as follows:

- **Robustness** describes the system's ability to maintain its desired performance and functionality even when faced with challenging conditions, such as dealing with uncertain or imprecise inputs;
- **Effectiveness** is a measure of its ability to perform the functions necessary to achieve goals or objectives;
- **Dependability** specifies the ability of a system to deliver a service that can be justifiably trusted;
- **Usability** describes the degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use;
- **Human agency** refers to the capacity of individuals to interact with, understand, and control AI systems, ensuring these technologies are transparent, explainable, and aligned with human intentions;
- **Human oversight** encapsulates the evaluation and guidance of AI systems to ensure their operation respects legal frameworks, fundamental rights, and general benevolence.

Ensuring the quality of AI systems demands a shared responsibility spread across the value chain. AI system design raises new challenges on the characteristics presented in Figure 1 which are sometimes called quality requirements or "-ilities".

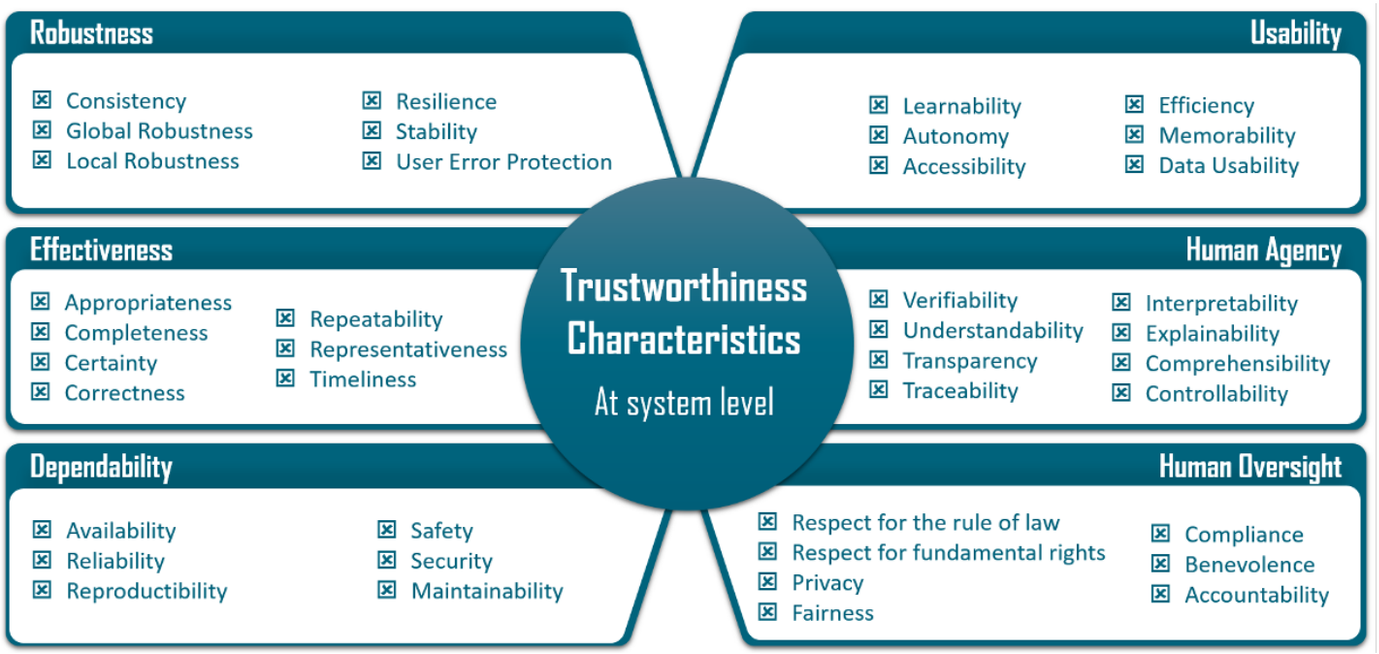


Figure 1: The AI Trustworthiness characteristics and sub characteristic (Mattioli, 2023)

These attributes can be associated to requirements on the system functions, the system performances, the development processes, the organization responsible for the system, the skills of the people within this organization, etc. The expected attributes depend on contextual elements such as the level of criticality of the application, the application domain of the AI-based system, the expected use, the nature of the stakeholders involved, etc. Hence, in some contexts, some attributes will prevail, and other attributes may be added to the list.
Trustworthiness characteristics can be assessed only if the Operational Design Domain (ODD) is clearly defined. *The ODD specifies the operating conditions under which a given AI- system is specifically designed to function as intended, i.e. in line with its intended purpose.* Many AI prototypes neglect to describe their ODD or leave it vaguely defined as the domain covered by the distribution of data used during training.

Assessments and audits may also be included in mandatory authorization and regulatory procedures. The European Commission's regulation indicates that such authorization procedures for AI will be introduced for the European market in the near future. As well as banning certain applications of AI, the directive requires high-risk systems to undergo a conformity assessment procedure. Last but not least, full trustworthiness in AI systems can only be established if all technical activities to establish trustworthiness are clearly defined for example by regulations, norms and standards to support the governance, processes of organizations and/or End-to-End methodology that use, develop and deploy AI.
Key Performance Indicators (KPIs) for measuring the quality of AI applications are important. However, obtaining trustworthiness measures remains a challenging task. On the one hand, measuring trust can help identify problems with the system before they become critical and allow for mitigation action to be taken before a failure occurs. On the other hand, measuring trust can help to improve the design of critical systems. By understanding the factors that contribute to user trust in AI systems, designers can create more reliable, safe, and secure systems. Another challenge in defining specific quality requirements for AI/ML applications is that different dimensions of trustworthiness cannot be assessed completely independently of each other. Instead, trade-offs must be made.
Some examples include:

- Increasing performance, such as the recognition performance of deep learning on image data, may come at the expense of traceability;
- Increasing transparency (for example, by revealing all hyper-parameters of a model) may lead to new attack vectors related to IT security.

To sum up, assessing trustworthiness in AI systems, through a thorough understanding and clear definition of the Operational Design Domain (ODD), as well as a rigorous assessment of trustworthiness characteristics, becomes key for an efficient design and operation of critical systems. This approach, requiring a balance between the different dimensions of reliability and adaptation to specific contexts of use, lays the foundations for a broader and secure adoption of AI, adapted to the needs and challenges of today's world. More details about the *Methodological Guideline for Trustworthy AI Assessment* are available in (Mattioli, 2023).

1.3 Users Perspectives and Interaction with AI

As discussed, the deployment of AI technologies raises various challenges, including the need for AI not only to be trustworthy but also understandable to a broader audience. Making AI algorithms

understandable by people is the goal of eXplainable AI (XAI). Users are provided with AI results completed with explanations, local or global descriptions helping to understand the model decision, in order to prevent confusion and understanding errors. Nonetheless, if XAI techniques are often understood and used by data scientists to study models' behavior, their adoption by end users requires further thoughts (Liano, 2020).
The challenge of unlocking XAI deployment to a broad audience lies in three layers:

- **Explainability** deals with the capability to provide the human with understandable and 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, etc.) 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, etc.) to be understood by a person according to its level of expertise or background knowledge.

A large body of work from XAI literature has thoroughly addressed the question of what characterizes an explanation. Recent work proposes to revisit this concept and to go deeper into interpretability and comprehensibility by taking inspiration from other fields such as psychology, epistemology and philosophy of science.
With the aim of accelerating their adoption and deployment, XAI systems must adapt their explanations to different stakeholders having their own background knowledge, skills, goals and interests. Interdisciplinary collaborations between data science and social sciences will pave the way to make AI systems understandable to a wider audience (Blanc, 2024). For more details on trying to assess mental models of XAI systems stakeholders using a semiotic-based framework, you can refer to (Dejan, Arlotti & Heulot, 2024).

This chapter underlines the critical necessity of developing trustworthy AI, particularly for integration into critical system. We began by highlighting the inherent risks and challenges associated with AI adoption, such as security, vulnerability, and reliability issues. To achieve trustworthy AI, a comprehensive approach is required. This involves revisiting and refining engineering methodologies, developing reliable software components, and experimenting with use cases in order to ensure they are fully addressed.

Confiance.ai aligns well with broader European efforts, which focus on establishing regulations and standards to ensure the development and deployment of trustworthy AI, such as the AI Act.

1.4 Contribution of Confiance.ai to the AI Act

The European approach to trustworthy artificial intelligence can be analyzed as consisting of three levels (see Figure 2). The highest level is regulation, applicable at long term, it sets the requirements namely for high-risk AI-based systems that will be deployed for the service of European citizens. The intermediate level, harmonized standards, are

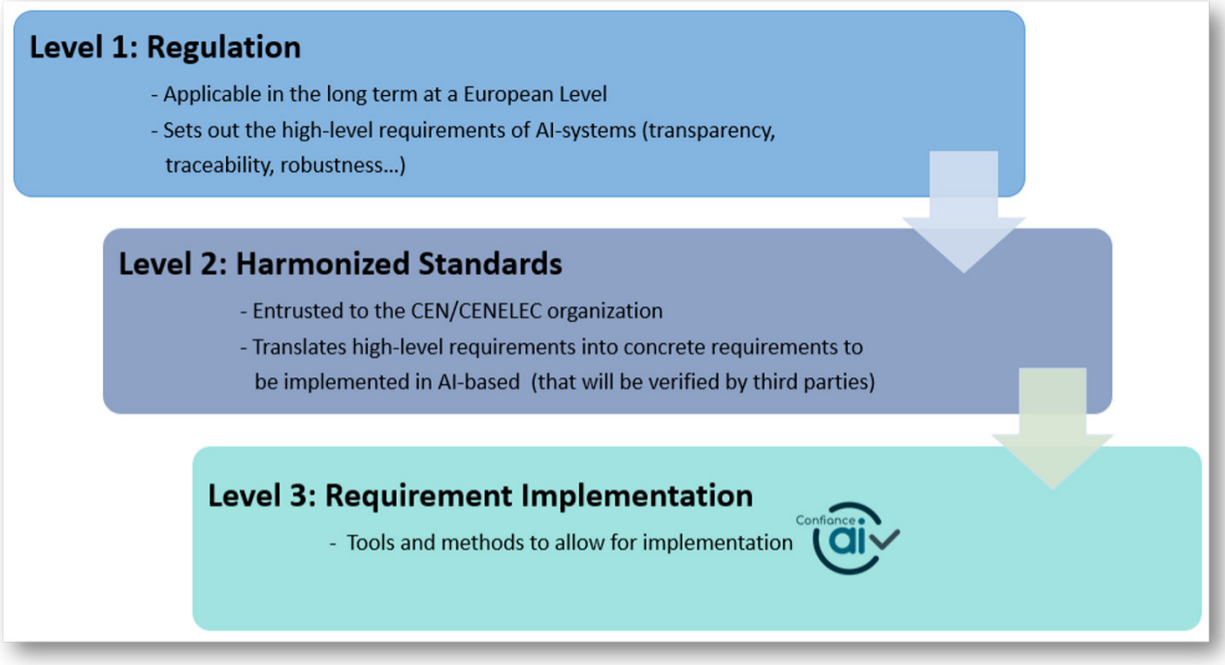


Figure 2: European approach to trustworthy artificial intelligence

to define concretely how the high-level requirements defined by regulation are to be operationalized by organizations and that will be verified by "notified bodies". The 3rd level covers the actual implementation of the requirements and the tools and methods to achieve these tasks. The contributions of Confiance.ai are placed in this 3rd operational level. Confiane.ai provides methods and tools to improve the trust in AI systems for critical applications, yet by extension it can also be applied to other non-critical applications.

The contributions of the Confiance.ai program to the AI Act are threefold in nature (see (Sohier, 2024) for details):

- Technological contributions, namely on three of the ten requests for standards made to CEN/CENELEC by the European Commission: "Robustness", "Accuracy", and "Data Quality". As an example on the latter, Confiance.ai has produced and evaluated around ten components and a dedicated platform allowing to improve quality of the input datasets for automatic learning systems. These tools (and some others addressing for example Explainability and Cybersecurity) are referenced in Confiance.ai catalogue of ressources (Sohier, 2024);
- Methodological contributions, a whole Body of Knowledge as described in section 2.1 displaying an End-to-end method and gathering methodological guidelines on many specific themes related to the trustworthiness of AI-based systems;
- Direct contributions to standards, as the Confiance.ai program was involved from the outset in the working groups set to produce the harmonized European standards. Among others, inputs on a taxonomy and attributes for trustworthy AI, support for initiatives for labeling AI products as well as companies designing them. ■

2. Confiance.ai Main Outcomes

The concrete outcomes of Confiance.ai are numerous and of different nature; going from documentary guidelines and methods, to software components and Functional Sets as component clusters fulfilling specific needs. After a rigorous process of documentation, evaluation and maturation, these outcomes have been systematically structured and released to the general public through the form of *'The Body of Knowledge'* and *'The Catalog'*.

2.1 The Body of Knowledge (https://bok.confiance.ai/)

The Body of Knowledge is one of the main results of the Confiance.ai program as it gathers a browsable version of the methodology known as *end-to-end* which covers the activities structuring the engineering cycle of a critical ML-based system. The Body of Knowledge is a compendium of expertise coming from multiple disciplines as it articulates the system level along with model and the data levels in the engineering process. The enrichment of this Body of Knowledge is continuous and expected beyond Confiance.ai.

The content provided in the Body of Knowledge is structured through the lens of an end-to-end engineering method and browsable through different roles in this process, namely through the scope of a: ML-algorithm Engineer, Data Engineer, Embedded Software Engineer, IVVQ Engineer or a Systems Engineer.

The Body of Knowledge displays the stages of the methodology from operational analysis and specification, down to development, and all the way up to validation and qualification. They can be navigated through each stage and according to each role, thus displaying the activities, sub-activities and workflow to be carried out when developing a trustworthy ML-based system. Figure 3 shows a view of the overall method on the Body of Knowledge, and the method itself is detailed in the next chapter.

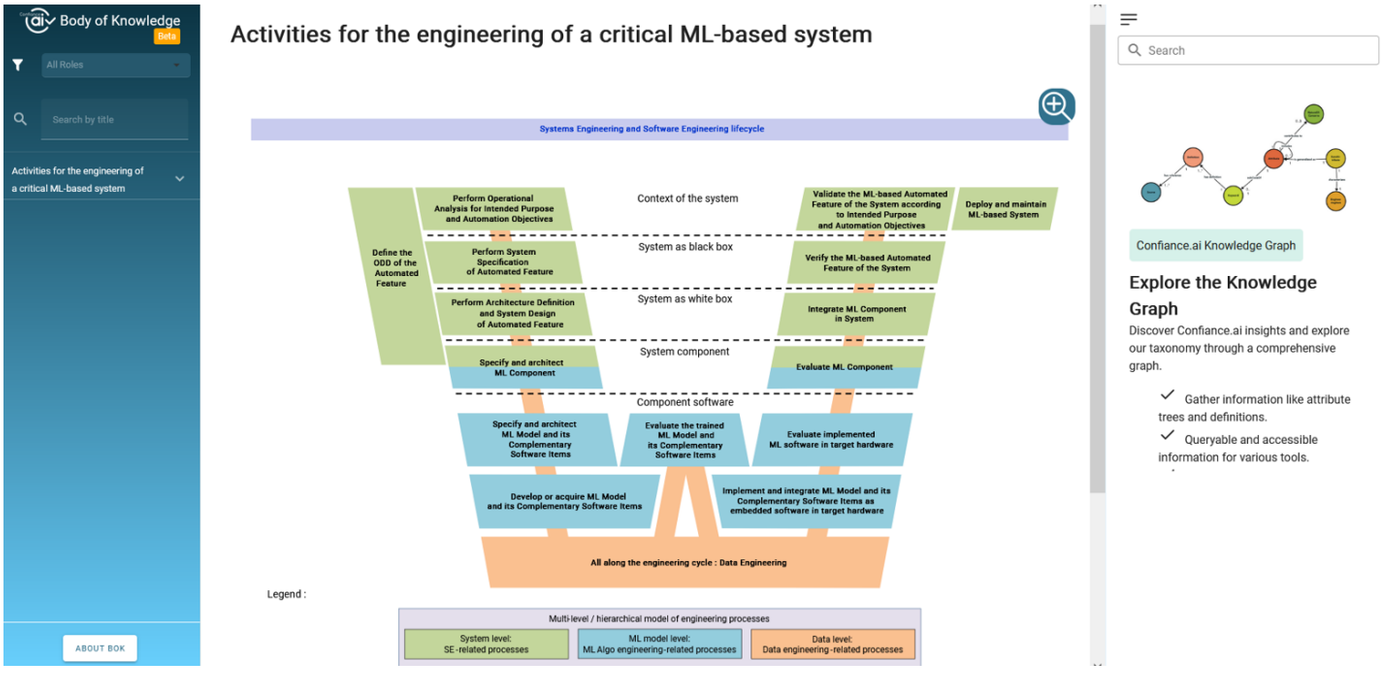


Figure 3: Simplified high-level view of the Body of Knowledge as a gateway to the End-to-End Method for engineering trustworthy ML-based systems

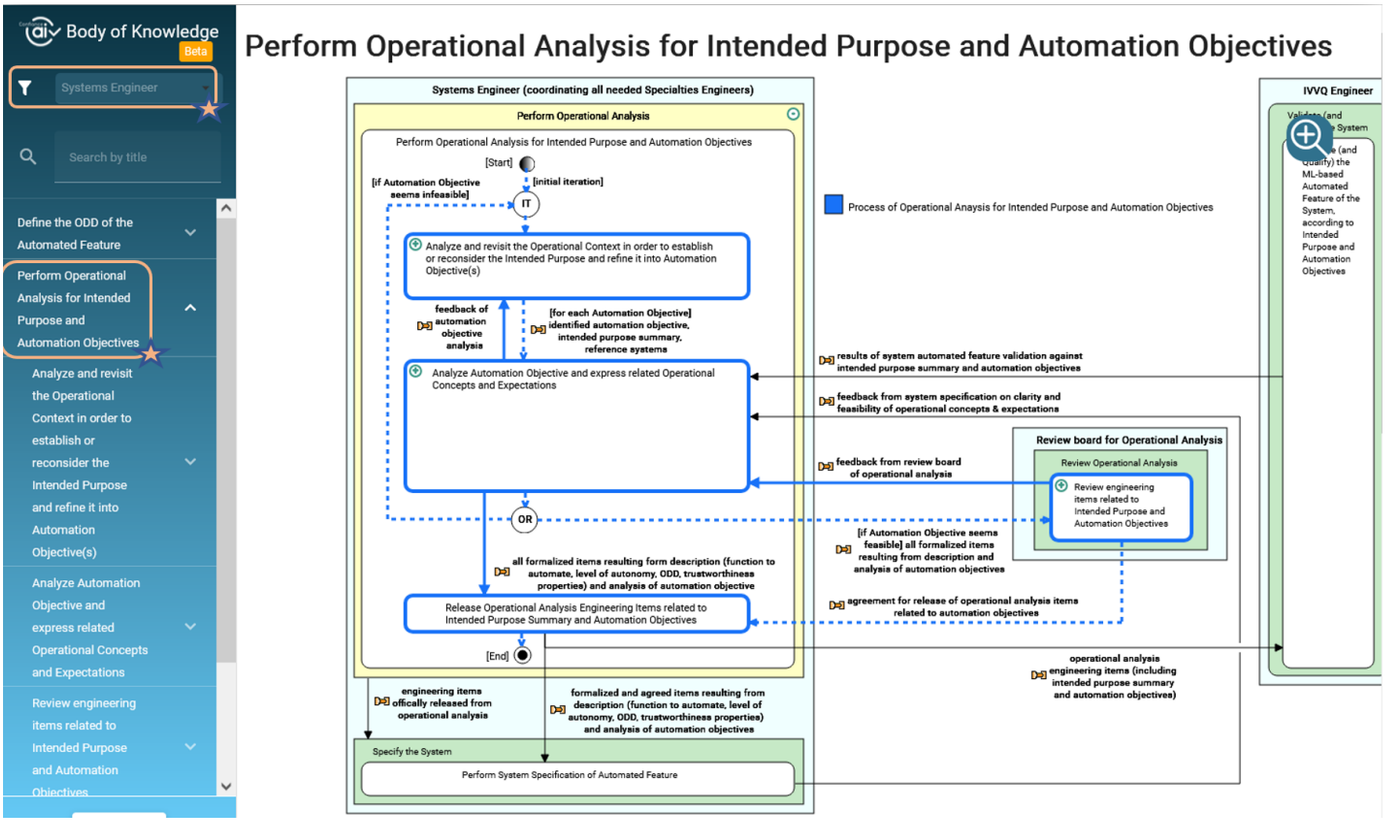


Figure 4: Navigating the Body of Knowledge through the role of a systems engineer and the activities to address when performing an operational analysis

Figure 4 shows a glimpse of the Body of Knowledge when navigating the first phase of the cycle (i.e. performing *Operational Analysis for Intended Purpose and Automation Objectives)* through the lens of a *Systems Engineer* profile, and looking at the specific activities within this phase. As an example, when performing operational analysis in order to include an ML-based component in the overall system, several engineering processes must be addressed. As shown in Figure 4, the operational context must be revisited to establish or reconsider the *intended purpose* and refine it into automation objectives, which will then have to be analyzed to take stock on their feasibility. This can include refinement iterations until a formalization can be made on *Automation Objectives* expressing the related *Operational Concepts* and *Expectations.* Once this goal is reached then the release phase can follow and will provide the inputs for the system specification of the automated feature.
(https://bok.confiance.ai/)

2.2 The Catalog (https://catalog.confiance.ai/)

The Catalog is a web application for browsing the results of the Confiance.ai program. It uses navigation and search functions (sorting, categories, etc.) to make it easier for users to navigate through the various results, which can take two distinct forms: they can be documentary or software, see Figure 5.

- Documentary when their form is exclusively literary: reports (studies or benchmarks), state of the art, PhD Dissertations or guidelines;
- Software from the moment they are supposed to be executed directly or through another application: a web application, a library, a plugin or a binary executable.

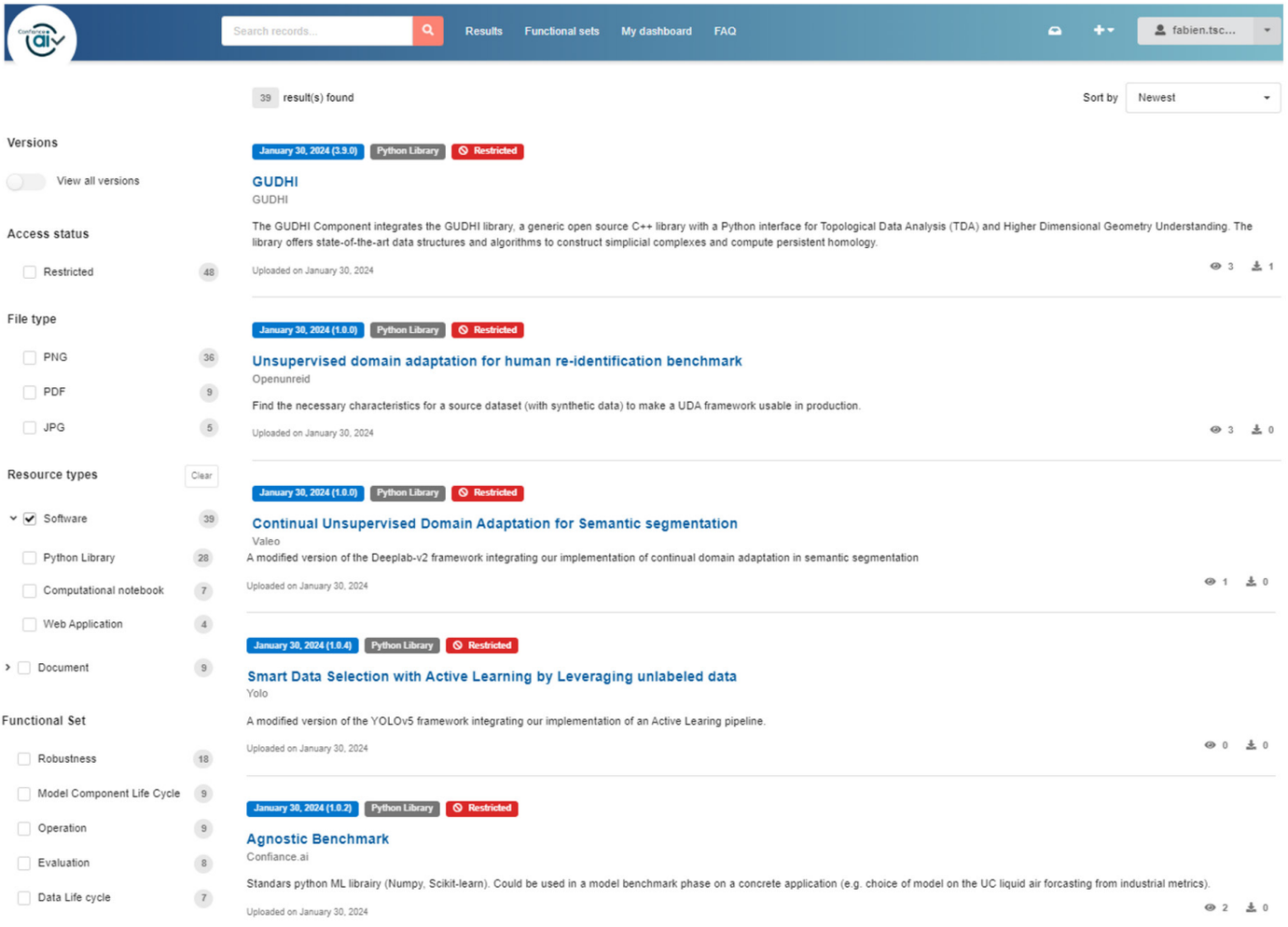


Figure 5: The result page of the Catalog

All the results of the program are gradually being integrated into the Catalog. This integration follows a process that includes the evaluation and maturation of the components. In fact, for a software component to be published in the Catalog, it must meet a certain number of criteria:

- Documentation, that allows its installation and execution;
- Packaging process, as python library or a docker container;
- Confiance.ai program use case application;
- Execution and integration in the Trustworthy Environment;
- The intellectual property and the license to which it is subject are identified.

Among these results, it is possible to find components that are the fruit of the research and development work of the Confiance.ai program itself, as well as components produced outside the program but evaluated within it. The Confiance.ai program aims not to duplicate existing and operating libraries and tools, but rather to identify, evaluate and when necessary, promote their relevance and value within their respective domain via the Catalog.

Chapters 3 and 4 lay out some of the main results leading to these outcomes. The first one addresses those related to the Body of Knowledge which includes the End-to-End methodology itself as a framework for engineering trustworthy AI-based systems and an overview of some specific topics of the method. The following chapter overviews results leading to the constitution of the Catalog which are broken down into the structure of the Trustworthy Environment, its contents, the segregation into component clusters known as Functional Sets for specific use, and finally, an overview of the intrinsic intertwining of Robustness, Uncertainty quantification and Monitoring aspects posed as the RUM Methodology.
(https://catalog.confiance.ai/) ■

3. End-to-End Method for Engineering Trustworthy AI-based systems

3.1 The End-to-End Approach: Structure and Methodological Drivers

The Need for an End-to-End Methodology

Trustworthiness of ML-based systems can only be ensured if considered and assessed at all stages of the system development cycle. Several disciplines are part-takers in this process to fulfill a global system purpose through proper workflows on each stage to integrate ML-related component requirements.

One of the objectives of the Confiance.ai program is to revisit the classic engineering disciplines (Systems Engineering, Software Engineering, Algorithm Engineering, Data Engineering, etc.) with regard to the challenges posed by the integration of AI into complex systems.
From the genesis of the program, it was clear that a methodology would be necessary for several reasons:

- for the multi-disciplinary interactions to take place and contribute to global processes for the development of AI-components,
- to ensure coherence and integration between these in terms of inputs/outputs from certain processes and disciplines to others,
- to ensure conformity of results and traceability of development of ML-components according to initial specifications,
- to allow for integration of ML-component development into a larger reference system, which follows on its own a pre-established development cycle,
- to provide a common reference to industrial partners applicable to safety-critical systems of different nature, on how to design, develop, integrate, deploy and maintain trustworthy ML-based systems.

Figure 6 shows an overview of the End-to-End method proposed by Confiance.ai. This overview combines, at system level, the classical "V" cycle and, at software level, the "W" cycle specific to Machine Learning. Naturally, this "V-W" cycle is not intended to be performed in one shot from left to right and from top to bottom: iterations between successive phases are always necessary.

In comparison to a classical (non-ML-based) systems, two new engineering domains have been integrated: ML Algorithm Engineering and Data Engineering.

Methodological Drivers

In order to design the engineering processes necessary to build trustworthy ML-based critical systems, the approach of Confiance.ai was based on a rigorous formalization of processes (through modeling, thus guaranteeing overall consistency) and interdisciplinary contributions (specialists from various fields were involved: Systems Engineering, Safety Engineering, ML Engineering, Data Engineering…).

Confiance.ai's End-to-End engineering method has been built through:

- consideration of drafts of standards such as ISO/IEC 5338 "*Information technology - Artificial intelligence - AI system life cycle processes*" and ARP 6983 "*Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI*", in order to structure the engineering phases and engineering items (objects) of the method, and to ensure compliance, by design with these future standards.
- analysis of the mature methodological and technological assets produced by Confiance.ai research teams, in order to:
  - demonstrate how Confiance.ai's results can help industrial users to meet the requirements of standards,

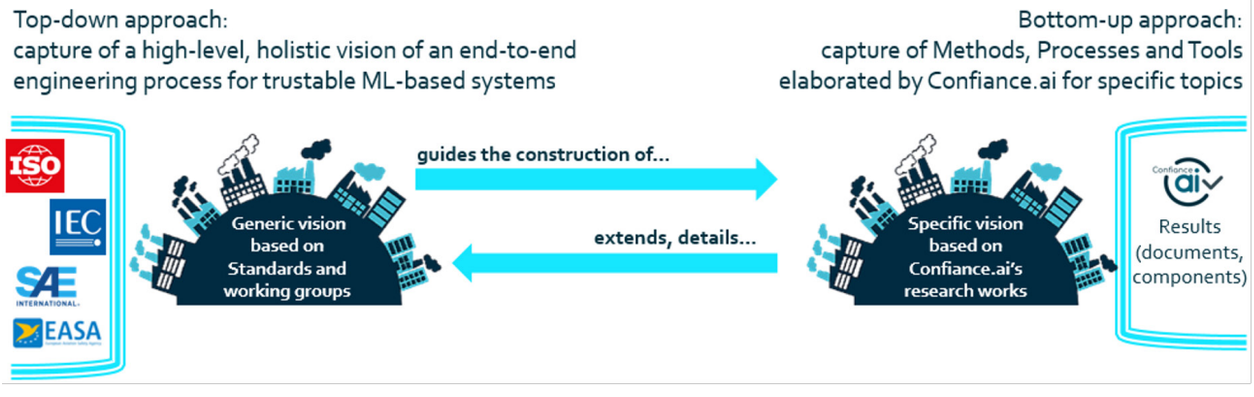


Figure 6: Overview of the Confiance.ai approach to build an End-to-End engineering method

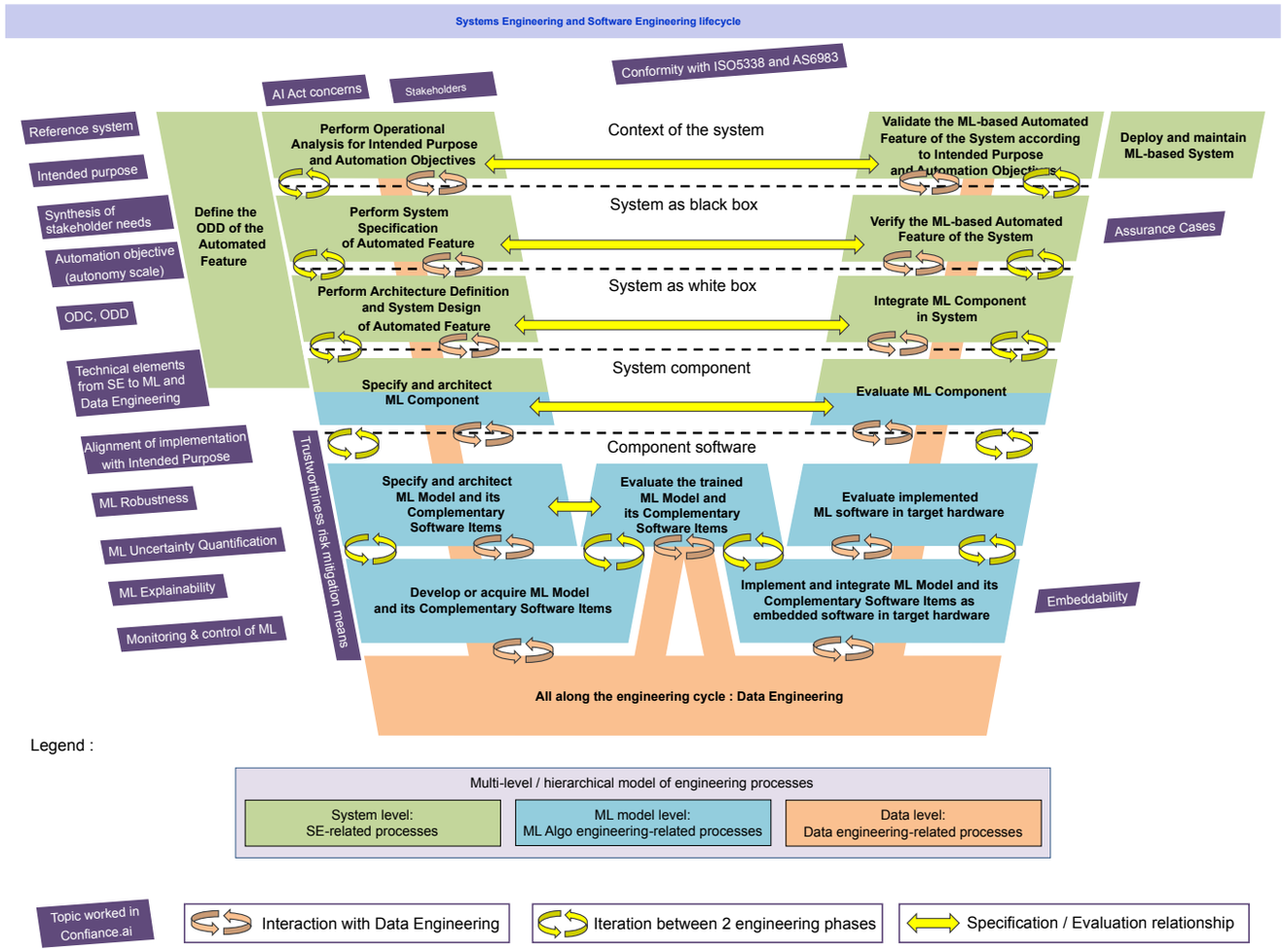


Figure 7: Overview of Confiance.ai End-to-End method for the engineering of critical trustworthy ML-based systems

- leverage the specificities and added value of Confiance.ai's within the context of the development of a critical ML-based system,
- favor the usability of Confiance.ai results as part of a structured development cycle.

Indeed, the different local methods and software components produced by Confiance.ai are each designed to meet a very specific goal, e.g. ML robustness, ML explainability, ML embeddability, generation of synthetic data, among others. However, they also need to be integrated and operated effectively by industrial users within a broader context of the engineering cycle of their products.

Confiance.ai's End-to-End engineering method, whose navigation is facilitated by Confiance.ai's Body of Knowledge (Confiance.ai, 2024a), intends to help users in the process of contextualization of Confiance.ai's results by fitting them into a consistent end-to-end process, (see Figure 7). For more details about this End-to-end method, readers can refer to (Robert, 2024).

The following sections explain three specific aspects of this engineering method: the design of the Operational Design Domain (ODD) and its impact on the overall engineering method, the Intended Purpose and its operationalization, and assurance cases as a way to build an IVVQ strategy.

Takeaways

- Confiance.ai has produced an End-to-End method seeking to aid industrial parties in the development of ML-components in coherence with an existing reference system.
- The End-to-End method details, high-level phases when developing ML-based systems, necessary processes and workflows per phase and interacting disciplines.
- The End-to-End method provides a framework of good practices when developing trustworthy ML-based systems based on existing norms and standards as well as on specific methods and components developed in the confiance.ai program

3.2 The Operational Design Domain (ODD) in the Engineering Method

Challenge

The ODD (Operational Design Domain) … how does it impact the engineering of trustworthy ML-based systems and how to unequivocally formalize it?

In practice, the scenario-space, i.e. the number of possible scenarios to be managed by an automated system, tends to be infinite. In the case of data-driven AI, it is impossible to ensure that the models will learn all possible scenarios only through the training data; this makes their safety evaluation challenging. A scenario-space must then be defined in which the automated system must operate safely without having to enumerate all different scenarios. The Operational Design Domain can support this definition of the scenario-space.

An objective of the Confiance.ai program was to revisit the existing engineering processes regarding the challenges posed by the AI integration into complex systems. In this case, the challenge regards the definition of the Operational Design Domain where a system is intended to operate.

Moreover, current approaches to define the ODD can be ambiguous, as the level of detail is defined according to the targeted audience, resulting in informal ODD descriptions, potentially incomplete and/or ambiguous.

The ODD plays a crucial role in defining the conditions and environments in which the AI system is expected to operate effectively and safely. A deep understanding of the ODD is essential to ensure that an AI system meets its intended purpose as well as its reliability expectations.

ODD Definition Process for an Automated Feature

Confiance.ai developed two distinct initial approaches for defining an Operational Design Domain (ODD): a taxonomy-based approach and an analytical approach.

The industrial partner Naval Group and the ODD team of Confiance.ai experimented on the definition of an ODD through two proprietary Use Cases, thus formalizing a unified process for ODD definition based on these two initial approaches. Figure 8 presents the process of defining an ODD via BPMN (Business Process Modeling Notation) process diagrams. The process is composed of five steps on the ODD: Scoping objectives definition, initialization, refinement, consolidation, and Business or operational relevance verification.

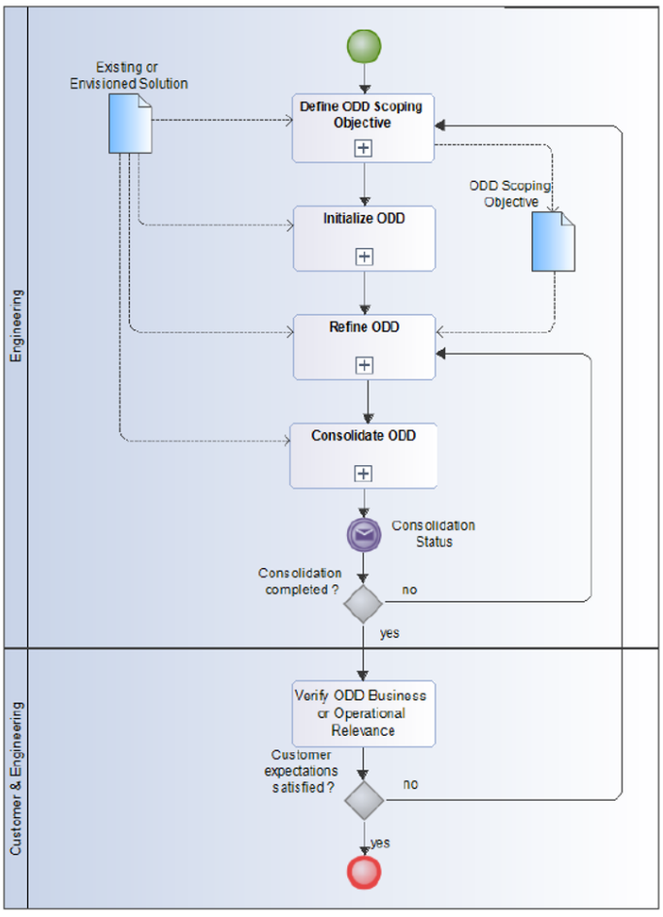


Figure 8: Overall ODD definition process

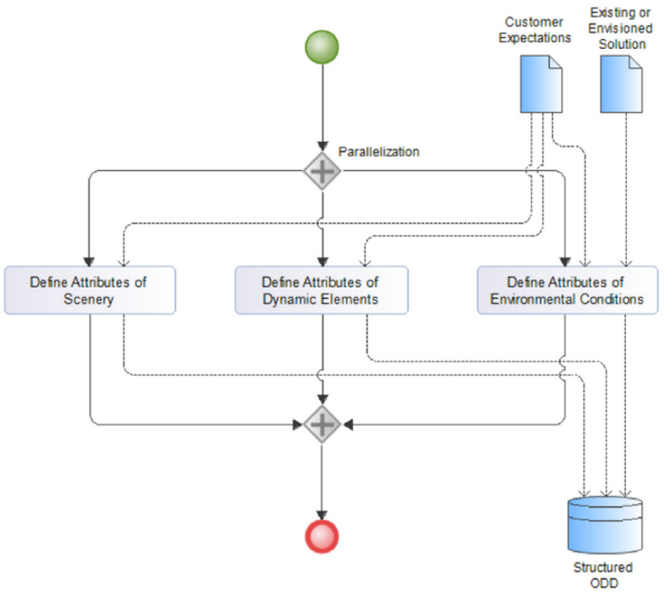


Figure 9: ODD initialization

The ODD initialization (Figure 9) borrows elements from the taxonomy-based approach where a hierarchical structure of attributes is defined. This step captures all attributes from the customer expectations, (i.e. customer needs and requirements) as well as environmental condition attributes that are considered or imposed by the existing or envisioned solution.
On the other hand, ODD refinement borrows elements from the analytical approach for refining the ODD previously initialized, (Adedjouma, 2023).

A rigorous and detailed ODD definition process is significantly important for the development of reliable and effective AI systems. Each step of the process contributes to building a robust ODD, aligned with the system's objectives and adapted to its operational environment.

ODD Engineering Process through the Design Lifecycle

Confiance.ai puts forward the notion that once the ODD is properly unequivocally defined and structured at a high level, it can be refined to be of use on different stages of the engineering lifecycle.
Confiance.ai has proposed an approach to refine an ODD from the early engineering phases to reduce ambiguity and incompleteness until a machine-readable stage where it can be used to support engineering activities such as safety analysis or V&V.

Figure 10 displays the overall process for refining the ODD through the engineering lifecycle. The process comprises 6 main steps that can be linked to different engineering levels defined in the Confiance.ai End-to-End method, (Confiance.ai, 2024a).
The first 3 steps pertain to the formal definition of the ODD as described in the previous section. In the case in which the system level ODD satisfies customer expectations, subsequent refinements can be pursued at the lower-level engineering phases to consider specific constraints pertaining to the related engineering phase. Details are presented on (ADEDJOUMA, 2023) for each refinement procedure of the ODD to ensure overall consistency and system reliability; the link to the other trustworthiness attributes is also addressed.

Takeaways

- ML-based systems inherently carry uncertainty as their performance depends on the training data, which must encompass all situations the system might encounter. The definition of the ODD of an ML-based system can tackle part of this challenge.
- A rigorous and detailed ODD definition and refinement process is fundamental for the development of reliable and effective ML-based systems.
- Confiance.ai proposes a definition and refinement of ODD for ML system features in order for them to operate correctly within the specified domain, recognizing that no guarantees can be provided outside their ODD.

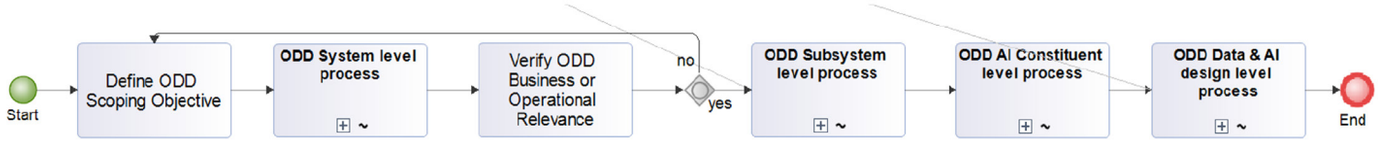


Figure 10: ODD refinement process through engineering lifecycle

3.3 Operationalizing the Intended Purpose

Challenge

The AI Act and the Intended Purpose, what can be the impact for the engineering cycle of ML-based systems?

Intended Purpose Definition

The AI Act defines the Intended Purpose of AI-based systems as *"the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation"*.

AI development is often based on technological-driven or data-driven approaches. With the AI Act regulation, development needs to consider the global added value for the end user, and a proper understanding of what the AI-based system can achieve. In this context, the Intended Purpose is a means of communication between stakeholders and end users.

The Intended Purpose is still relatively new in the AI field. However, we can envision that it still relies on four main pillars:

- Intended Population: *who will be subject to the use of the system?*
- Intended Users: *who will use the system?*
- Intended Use Environment: *what will be the operating conditions of the system?*
- Structure and Function of the component

Confiance.ai has produced engineering items that could allow addressing some of these pillars since the AI Act does not explicit it today.

The *Intended Population* and *Intended Use Environment* could be dealt with thanks to the Operational Design Domain (ODD), originating from the automated driving field (SAE J3016), which presents a voluntary restriction of the operating conditions under which an automated system is designed to function. In the meantime, in the same field, the Object & Event Detection and Response (OEDR) could cover some of the Structure & Function of the component by stating how and when the AI-based system should react to identified situations and objects in its environment.

The intended purpose is a formalization that states what the ML-based component is meant to do, how it intends to do it, and for whom it will do it. Confiance.ai considers that this entails that it is to be translated into a set of requirements to be considered:

- at the beginning of the engineering cycle, from the operational analysis,
- throughout the cycle to ensure its consideration and implementation in all phases (considering feedback loops iterations and adapting it if necessary),
- and, finally at the end of the cycle to validate conformity with what is stipulated at the beginning of the cycle.

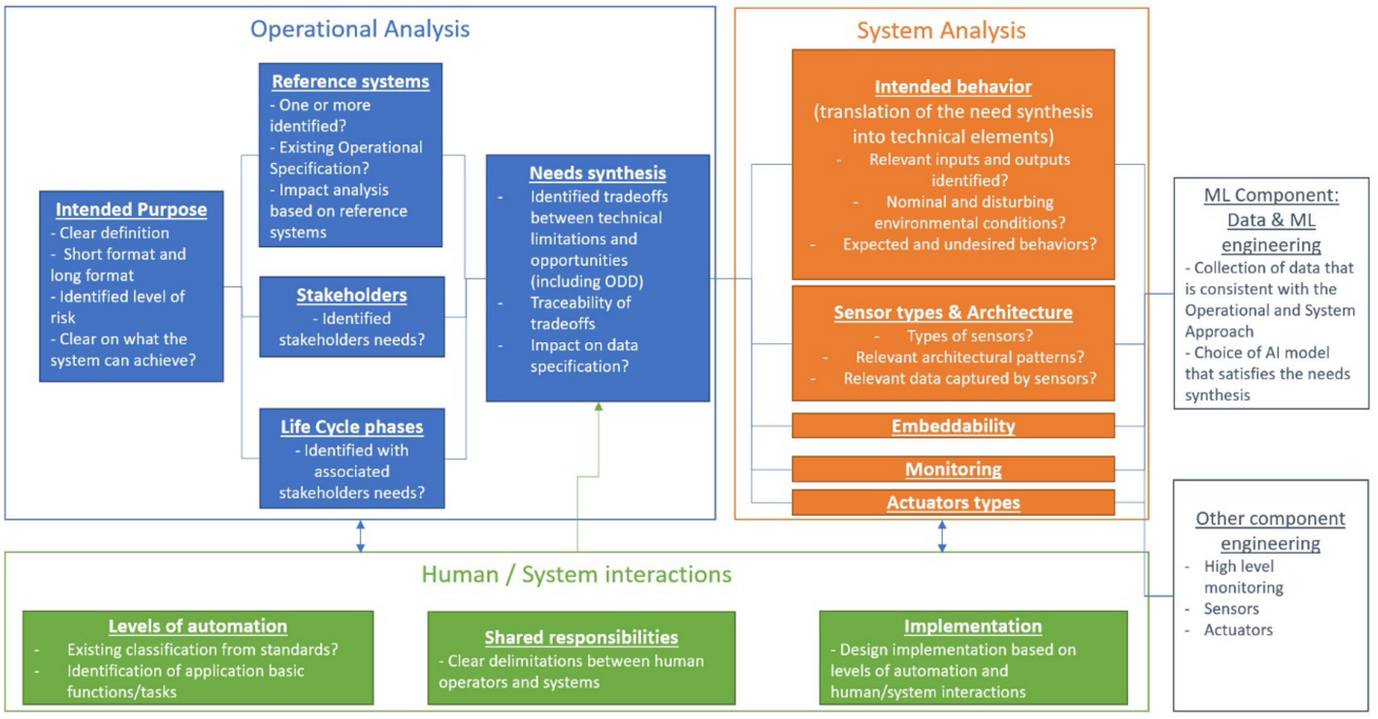


Figure 11: The Intended Purpose as a key driver at operational analysis and system analysis within the End-to-End approach of AI-based systems

Confiance.ai puts forward a methodological approach, depicted in high-level on Figure 11, where the Intended Purpose is managed as a design objective based on reference systems. The motivation being that the ML-based component shares properties with other non-AI-based pre-existing systems, called reference systems, where operational specifications are already available and structured.

As displayed in the diagram, the first step is to define what the Intended Purpose of the AI-based system should be. Its design often results from the automation of specific functions to achieve stakeholders' needs. This automation can be performed based on what we consider as reference systems, systems that share the same application context (applicative reference systems) or that share the same technology (technological reference systems). Studying the gaps between a newly AI-based system and their reference system can help in several ways:

- We can benefit from a previously built Intended Purpose that is considered mature and upon which we can expand on;
- It enables to envision a preliminary objective for the system design;
- It helps classifying the system relatively to other systems and benefits from the familiarity of end users with their reference systems.

Confiance.ai addresses two major axes to explicit the role of the intended purpose throughout the engineering cycle of the ML-based system: its role on operational design and its role on system design, readers can refer to (Bohn, 2024) (Mantissa & Bohn, 2024) for details.

As an overview, at an operational level, the resulting Operational Specification shall achieve the synthesis of stakeholders' needs for each lifecycle phase of the ML-based system. The Operational Specification, associated with the system ODD shall guarantee the Intended Purpose. The following phase in the method is then system design, which shall offer technical considerations at system level to ensure that the implementation will be in accordance with the defined Intended Purpose. That is the way to link the Intended Purpose and the Design Intent. It gives designers constraints on the system scope, and can help in defining evaluation objectives. Going beyond the initial Intended Purpose is a risk of function or pursuing multiple purposes without clear delimitations, where neither the designers nor the end-users can fully apprehend the full scope of what the AI-based system is capable of. This can lead to potential hazards and misuses from the end users.

Moreover, the design process supports the expected collection of evidence used to validate that the system achieves its Intended Purpose. See (Mantissa & Bohn, 2024) for details about the System design for the Intended Purpose of ML-based systems.

Takeaways

- The Intended Purpose is a key pillar in the design of AI-based systems, it should guide development and ensure coherence of expectation between users and stakeholders of what the system can and cannot do.
- The notion of the intended purpose is not mainstream in industry today and the AI Act does not provide methods to build it.
- Confiance.ai provides methods and assets on how to start operationalizing it for ML-based systems.

3.4 Assurance Cases (AC)

Challenge

How to provide proper justification on the trust we can have on an ML-based system and how to trust this argumentation?

Assurance Case Definition

An Assurance case is a set of structured claims, arguments, and evidence that provides confidence that an AI system will possess the particular qualities or properties that need to be assured.

An Assurance Case (AC) provides a structured argument to justify certain claims about the system, based on evidences concerning both the system and the environment in which it operates. In the AI domain, the following challenges arise:

- **System Complexity:** The AI components they contain are usually difficult to understand and analyze, making it challenging to develop a comprehensive Assurance Case.
- **Heterogeneity of evidence:** Assurance Cases must typically rely on a variety of evidence, including formal proofs, informal arguments, and empirical results. This heterogeneity of evidences is difficult to integrate and to reason about in a consistent manner.
- **Scalability and maintainability:** Assurance Cases can become very large, complex, and difficult to maintain, especially in the current context where new ML methods and techniques are emerging at an ever-increasing pace.
- **Human factors:** Assurance Cases are ultimately about convincing stakeholders that a system meets certain requirements. These arguments must therefore be understandable for target audience, including technical experts, non-technical users, and regulators.

Confiance.ai responds to these challenges advancing towards a globally accepted, well-argued IVVQ strategy for AI components.

Assurance Case Development

Assurance Case development in the Confiance.AI program relies on the simultaneous combination of two main approaches: starting from high level properties expressed on an engineering item to develop an argument in a top-down fashion, or starting from the available methods and tools to provide evidence in a bottom-up approach.

• Assurance Case Development: Top-down approach

In a top-down approach of Assurance Cases development, starting from a property on a specific item of interest, the goals are decomposed until they are sufficiently simple to be answered with a specific method or tool, which can be linked to a specific V&V activity. The steps to this approach are detailed in the work of (Jenn, 2023) and (Jenn, 2024) in the Confiance.ai program. In the workflow, for example, evidential steps are reached when solutions to the contextualized specific goals and; conversely; specific reasoning steps are necessary for goal refinement. The approach is considered as "top-down", since properties are refined progressively down to the point where the final goals are simple enough to be verified or proven. It implies the existence of some verification artifact (e.g. (a demonstration, a test result).

• Assurance Case Development: Bottom-up approach

The **bottom-up** method starts from available methods and tools that can be provided as solutions or evidences, and going up in the argument to try to link them to higher-level properties. The steps to this approach can also found in (JENN, 2023) and (JENN, 2024); it includes, among others, claim deduction based on available solutions and how to link them to GSN[1] goals.

• Assurance Case Development: A mixed approach

In practice, building an Assurance Case actually combines top-down and bottom-up reasoning. In particular, having a ready-made list of solutions can be used as building blocks to build part of the argumentation "bottom-up" (in the same way as having a list of software building blocks can be used to make appropriate design choices, etc.). On the other hand, when no predefined solution exists, it will be required to produce specific evidences and to define the associated V&V activities, which also implies verifying that they are feasible and applicable in the current industrial context.

Assurance Case Evaluation

In order to consolidate and ensure the validity of the argument, it is recommended to perform a critique of the assurance case product. This is also the procedure followed by external actors when reviewing such an assurance case. This verification can be done by trying to identify all scenarios that could invalidate the reasoning (the potential defeaters) and justify why those scenarios are not possible or prevented. This information may be kept outside the Assurance Case but could be important for the future reviewer.

An evaluation methodology is then required. This methodology should be based on concepts coming from the ACs literature as well as from the recognized literature of other domains (e.g. usability testing). The figure 12 presents an overview of the ACs evaluation process proposed within the Confiance.ai program. ■

Takeaways

- Assurance Cases for ML-based systems are a fundamental rigorous formalization of claims and argumentations of the system's capabilities, useful for internal as well as external review.
- A mixed bottom-up / top-down approach is detailed by Confiance.ai
- The AC itself must be trusted and therefore evaluated. Confiance.ai puts forward a 6-step process to tackle this challenge.

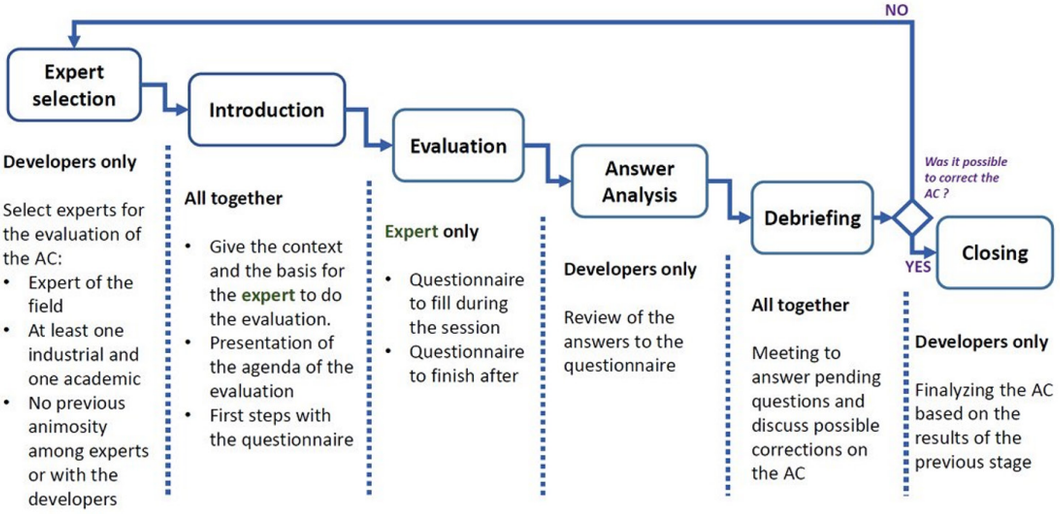


Figure 12: Assurance Case evaluation method (Jenn, 2024)

1. Goal Structuring Notation https://scsc.uk/r141C:1?t=1

4. The Trustworthy Environment and Functional Sets

Trustworthy Environment Definition

The Trustworthy Environment designates a modular set of components which, arranged in compliance with provided guidelines and documentations, can be used to instantiate interoperable tool chains whose execution enables the design, development, integration and maintenance in operational conditions of trustworthy AI components within AI systems.

The Trustworthy Environment is intended to be a simple and effective solution to enable the adoption of trustworthy AI by industries. It is designed to enable the gradual addition of artificial intelligence into industrial existing engineering processes. This is by revisiting existing concepts and methods without the need to overhaul long-deployed and operational engineering environments and processes.
Thanks to its modular nature and ability to integrate existing engineering environments, deploying the Trustworthy Environment can be operationalized in several ways:

- It can be taken as a whole by building a full engineering workbench tool chain. This would be the recommendation for anyone wishing to start from a blank environment. This is rather relevant for testing purposes.
- It can be merged with industrial native components and used as an engineering workbench orchestrator; a suitable choice if the existing environment covers only a part of the end-to-end AI trustworthy process.
- Only a selection of relevant components may be deployed and used directly in the industrial workbench, which makes it possible to benefit from the added value of the Trustworthy Environment, even if a complete engineering workbench is already operational.

This latter approach generally appears to be the best fit to industrial constraints. Therefore, in order to simplify its implementation, the Trustworthy Environment can also be approached through the prism of Functional Sets.

On the need for Functional Sets…

It stems from the industrial need to gather, test, and deploy coherently and consistently a set of tools and/or methods on trustworthy AI around a specific topic at the core of an operational engineering workbench.

Functional Set Definition

A Functional Set is a set of libraries, web applications and methods dedicated to a particular theme of Trusted AI (e.g. robustness, uncertainty, data lifecycle, explainability, embarcability, monitoring, end-to-end engineering...). The consistency of a Functional Set is based on a central user guide (head documentation) that enables users to find their way around the topic and how to address it.

As an example, someone specifically interested in the *robustness* of AI-based systems might consider the Functional Set on *robustness* as an entry point rather than tackling the issue through the Trustworthy Environment as a whole. A total of nine Functional Sets are available as results of Confiance.ai. Six of these are process-oriented. They approach the question from an engineering point of view, for instance: how to correctly manage the data lifecycle in the design process of an AI-based system or how to consider the end-to-end approach of such a system. The remaining three address essential issues of trust in artificial intelligence: robustness, explainability and uncertainty. Below is a list of all the Functional Sets with a short description of each one:

- **End-to-End Functional Set:** Contains methods and tools needed to identify the relevant reference implementation for a given use case, and then to implement it via a selection of tools, methods, and characterization elements (e.g. ODD, Assurance Cases), (Adedjouma, 2023) & (Robert, 2024).
- **Data Lifecycle Functional Set:** Covers the lifecycle of the data divided into five phases: Data Orientation, Data Architecture & Design, Data Implementation, IVVQ, and Deployment. (Langlois, 2024).
- **Model Component Functional Set:** Covers the lifecycle of an AI Model & Component: Specification, Development, Evaluation, Implementation and integration.
- **Deployment Functional Set:** Processes (methods and tools) that covers the integration of a ML model & component within a system. It can also be seen as embarcability.
- **Operation Functional Set:** Contains tools and methods that covers the AI-based system working in operation.
- **Evaluation Functional Set:** For tools and methods allowing to evaluate an AI component (Mattioli, 2023).
- **Robustness Functional Set:** For tools and methods that contribute to demonstrate robustness properties inside systems integrating AI components (Khedher, 2024).
- **Uncertainty Functional Set:** For tools and methods for quantifying uncertainties, and their contribution to trustworthy properties within a system integrating an AI component.
- **Explainability Functional Set:** Covers tools and methods that contribute to provide explainability properties inside a system integrating AI component. (Poche, 2023).

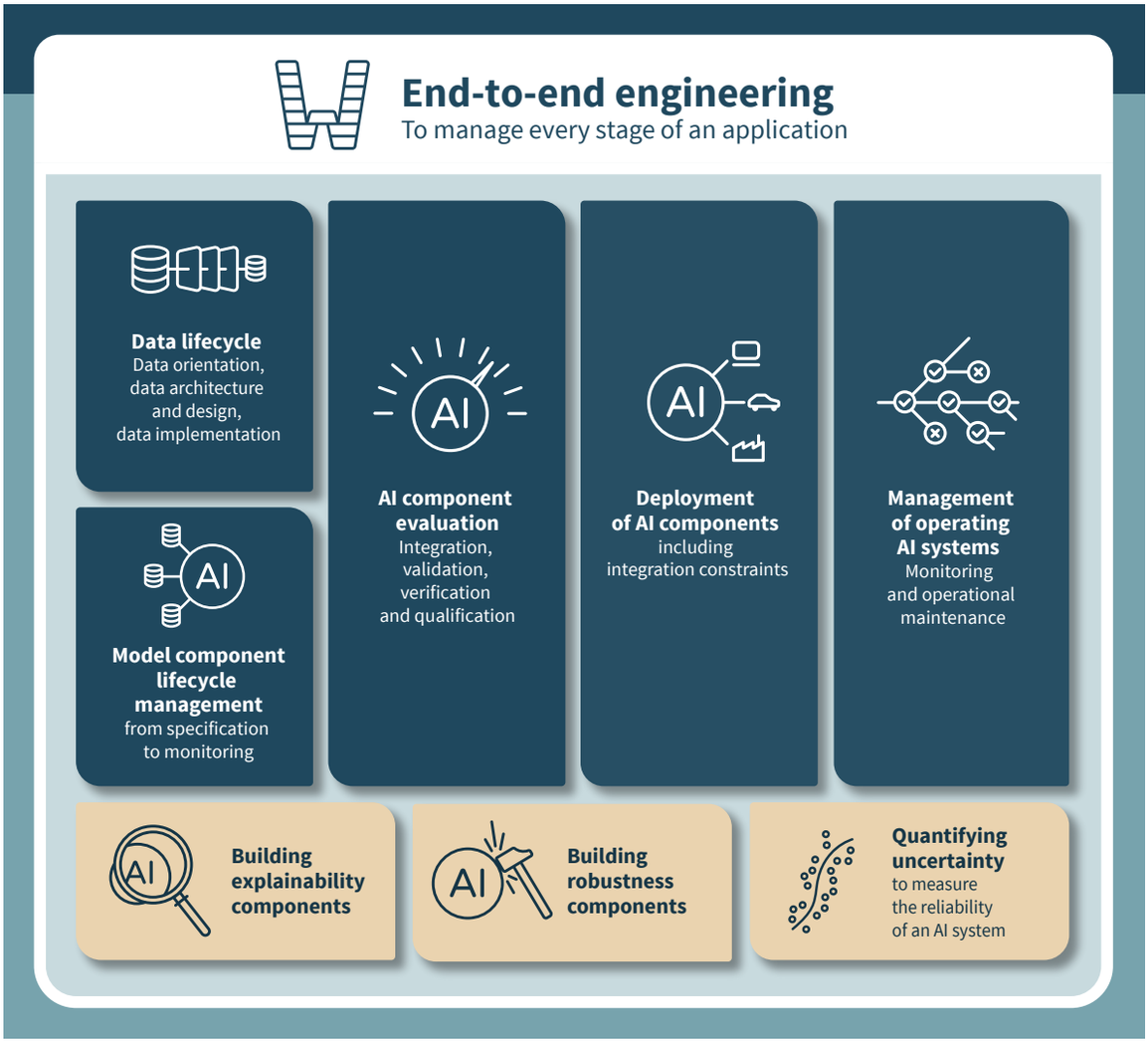


Figure 12: Integration of the Functional Sets produced by Confiance.ai

Figure 12 illustrates the articulation of different Functional Sets that have been defined and developed within the program. In the following, a glimpse on some of the program's most comprehensive Functional Sets is presented.

4.1 Functional Set 1: "Robustness"

Robustness Definition

The robustness of a system is its ability to maintain its desired performance and functionality even when faced with challenging conditions, such as dealing with uncertain or imprecise inputs.

Robustness plays a vital role in creating trustworthy AI systems. Although it is a broad term applicable across various systems, in this section, our discussion narrows down to AI-driven systems, with a particular emphasis on neural networks. **It refers to the ability of a system to maintain its intended behavior and avoid causing harm even under challenging or unexpected conditions.** Evaluating robustness is especially important for high-risk systems before they are deployed for user access. Incorrect decisions made by systems can pose a significant threat to human life, especially in cases where lives are at stake such as self-driving, robotics and cybersecurity. In these cases,

it is essential that systems are designed and implemented in a way to be able to withstand input disturbances.
Consider an autonomous vehicle approaching a roundabout with a STOP sign. Dust covering the sign causes the traffic sign detection system to misinterpret it as a YIELD sign, allowing the vehicle to proceed through the roundabout dangerously. This scenario underscores the need for robust traffic sign detection systems that can withstand unforeseen conditions like dust. Assessing robustness involves testing the system on diverse scenarios to identify potential errors and ensure safe operation. In fact, in this example, robustness evaluation is crucial for building reliable autonomous vehicles that navigate roads responsibly.

Given the inherent danger of non-robust systems, the primary objective for users should be to develop an AI-based system that is resilient to input perturbations. It is important, for these users, to provide a formal guarantee that the developed AI-model is robust. These formal guarantees of robustness will increase users' trust in using the system in a secure manner.
In order to help users assess the robustness of their AI-based systems, a robustness Functional Set (FS) is designed and implemented with 3 main functionalities which can operate independently depending on the user's needs, see Figure 14. However, Confiance.ai puts forward a general usage pipeline.

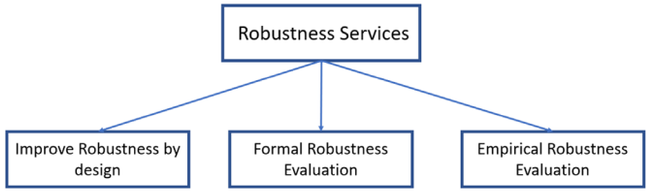


Figure 14: Functionalities or services in the Robustness FS

The user disposes of a FS offering a wide range of techniques for each of the three functionalities. To help them make their choice, a guide is provided that describes the compatibility of the different techniques with the AI model types (TensorFlow, PyTorch, Onnx, etc.) and data types (Tabular, Images, Time series, etc.).

Formal Robustness Evaluation Definition

The formal robustness evaluation seeks to provide a mathematical guarantee that a system will maintain its desired behavior even when subjected to any perturbation within a certain range of perturbations.

Empirical Robustness Definition

Empirical robustness evaluation is the study of a system's resilience to specific, intelligently calculated perturbations called adversarial attacks.

4.2 Functional Set 2: "Data Lifecycle"

Another important Functional Set in the program is *Data Lifecycle*, which addresses the data lifecycle from the perspective of an end-to-end data engineering process. This is an alignment with the vision of ensuring the trustworthiness of AI systems through end-to-end engineering.
Data lifecycle management in ML is crucial for scaling the development of demanding, complex, or critical systems. It is important to formalize the data engineering process comprehensively, making it complete, repeatable, and robust.

Data Lifecycle Definition

The Data Lifecycle is a set of multiple flows, in interactions and transformed by functions of the systems.

The Data Lifecycle is a set of multiple flows, in interactions and transformed by functions of the system. In order to certify such a system, we have to respect all requirements, ensure traceability and explainability, etc. Reaching this level of expectations implies a formalization of the data lifecycle during development and deployment. All this means that there is a paradigm shift on data when developing complex and critical AI systems, which introduces uncertainty. This can be seen as moving from a code-centric development, with the associated tests, to a global and mastered data lifecycle, at development and runtime.
In the context of trustworthiness of systems built in co-engineering with AI, the objective is to introduce a breakthrough with AI with minimal changes on the traditional practices in Systems Engineering. To do that, we simply started from the traditional development lifecycle, and next customized it for data with IA/ML practices. The proposed workflow is divided into five phases, as presented in the figure 15. As an example, the first phase of *Data Orientation* identifies the business and operational goals for data, and the expectations on data consisting of requirements, ODD (Operational Design Domain) borrowed from automotive (SAE J3259, 2021), and operational scenarios. The entire workflow is detailed in (Benoit Langlois, 2024).
The process presented above contains the foundation steps. However, to guarantee the trustworthiness of a system with AI, five transversal concerns are added to the global workflow to contribute and master the lifecycle of data.

- **Data quality assessment:** trustworthiness is pursued as data quality is guaranteed during the data engineering process (i.e. data collection, filtering, processing, etc.).
- **Assurance Case**, thanks to a justified measure of confidence, ensures that a system will function as intended in the environment of use (Weinstock, 2015).
- **Automation by AI/MLops**, i.e. continuous integration applied to AI/ML models, improves productivity, and avoids manual operations, possible sources of errors.
- **(Digital) Documentation** remains an important artifact to keep trace of data development and is a mandatory activity for the certification,
- Reusability for capitalization and reuse of common data assets.

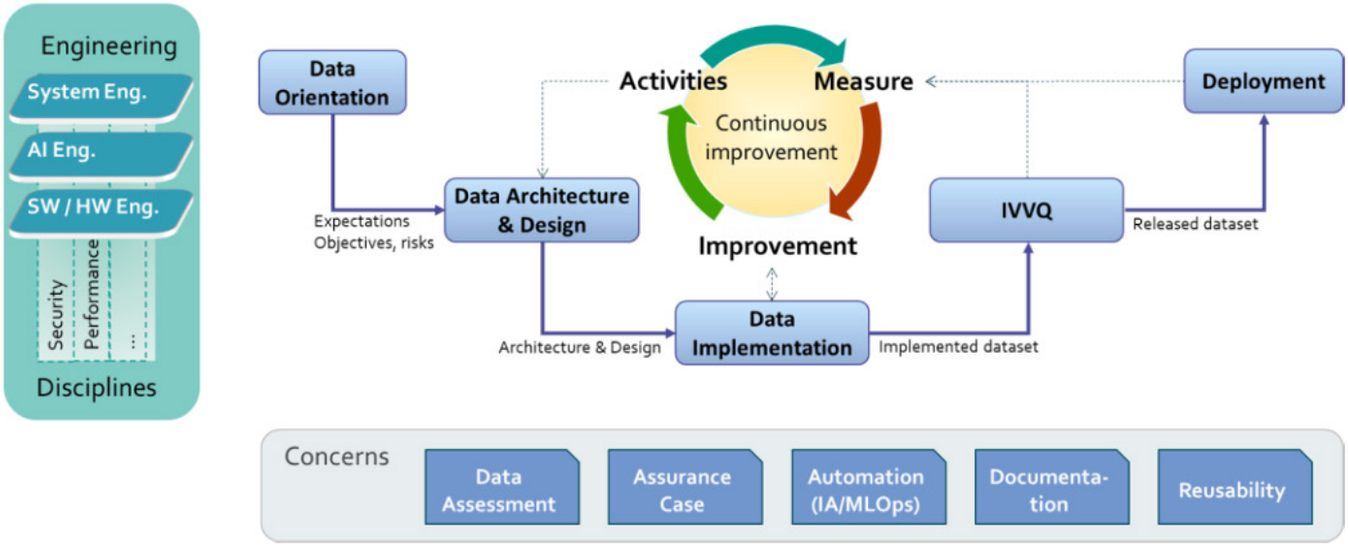


Figure 15: Overview of Data Lifecycle phases

4.3 Functional Set 3: "Explainability"

Explanation Definition

An explanation is a statement or application result that clarifies, informs, or provides reasons for a particular event, phenomenon, process, decision, or concept.

Explainability Definition

Explainability deals with the capability to provide humans with understandable and relevant information on how an AI application is coming to its result.

Explainability in industry is deemed crucial for the establishment of trust and credibility. When complex algorithms are made clear, accountability can be held by everyone. The identification and rectification of biases are facilitated, ensuring a fair and transparent decision-making process. Regulatory requirements are met, and ethical innovation is fostered, leading to sustained success in the evolving technological landscape. Definitions from (Dejean, 2023b) & (Mattioli, 2023).

Currently, the '*Explainability*' Functional Set contains six explainability libraries that have been studied in the program (DEJEAN, 2023a):

AIX360 https://github.com/Trusted-AI/AIX360
Alibi https://github.com/SeldonIO/alibi/tree/master
Captum https://captum.ai
Saliency https://github.com/PAIR-code/saliency
Shap https://github.com/shap/shap
Xplique https://github.com/deel-ai/xplique

The collection of libraries is organized within a control platform called Kaa, designed to simplify the utilization and configuration of the methods and metrics encompassed in these libraries. Accessible through a Text User Interface (TUI) or by utilizing a command file within a docker environment, Kaa currently provides access to 43 methods and 8 metrics.

To apply Explainability on a use case, it is essential to select appropriate explainability methods to apply. The selection of these methods should be guided by several key factors as outlined in (Poche, 2023). These factors include those related to use case constraints (e.g. task, data type, model architecture and access to its weights/gradient, …) and those related to use case requirements (e.g. scope of the explanation, target audience,…). The recommendation remains to use several methods whenever possible.

Interpretability Definition

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.

In its current version, the limitations of the Explainability Functional Set are mainly constraints of the functionalities that stem from both the inherent limitations of the underlying explainability libraries and the restrictions outlined in the explainability literature; These include: lack of diversity in explainability formats, superficial understanding (i.e. humans are only able to understand a proxy of the model's behavior, therefore a trade-off arises on between explanation comprehensibility and explanation faithfulness to the model behavior), interpretability (still an open field), and visualization (of the explanation which is impactful on humans).

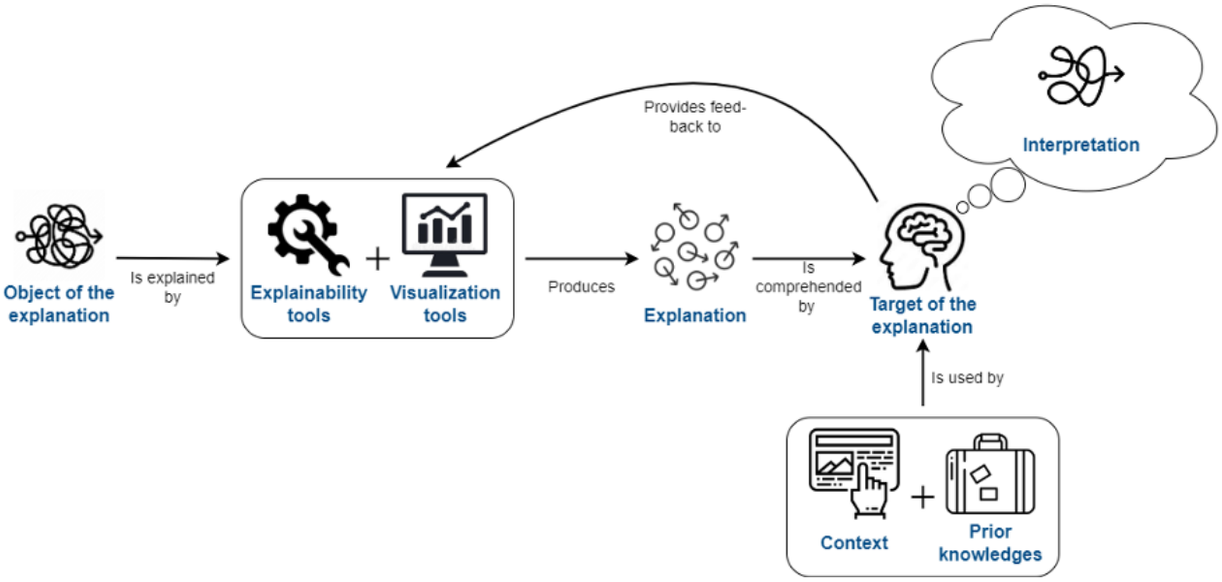


Figure 16: Schema of explainability elements

4.4 RUM Methodology, a Combination of Functional Sets

The Need for Combining Specific Functional Sets

Robustness must be ensured! However, how sure are we on the measured attributes for Robustness? Can we quantify this uncertainty? In order to achieve this, the ML-models must constantly be monitored.

An important fact about robustness techniques is that, in order to successfully address robustness attributes, they usually need to work alongside Uncertainty Quantification and Monitoring Techniques. Robustness, Uncertainty quantification, and Monitoring are crucial aspects in ensuring the reliability and effectiveness of ML models. Most importantly their associated employed methods need to work together to successfully address the challenges that are presented.

- First, Monitoring is a crucial aspect of ensuring the robustness and reliability of systems, especially in the context of complex and dynamic environments. Its absence hinders the ability to adapt to changing conditions, detect anomalies, maintain system health, implement effective fault tolerance, calibrate models accurately, and gather data for uncertainty quantification. These factors collectively contribute to reduced robustness and increased uncertainty in the performance of a system.
- Second, Uncertainty quantification is essential for understanding the limitations and potential variations in system behavior, and its absence can lead to overconfident decision-making, inaccurate risk assessment, ineffective adaptation to changing conditions, misleading monitoring indicators, limited sensitivity analysis, underestimation of errors, reduced confidence in decision support systems, and inadequate resource allocation. These factors collectively contribute to a decrease in the robustness and reliability of systems.
- Third, machine learning robustness addresses the question of how well a machine learning model can maintain its performance and make accurate predictions in the face of various challenges, perturbations, or uncertainties. Its absence can lead to uncertain system responses, inadequate model calibration, unreliable monitoring indicators, increased false alarms, difficulty in identifying root causes, challenges in adaptive control, compromised fault tolerance, and limited resilience to environmental changes. These factors collectively undermine the effectiveness of uncertainty quantification and monitoring efforts in maintaining a reliable and well-performing system.

In summary, these three concepts are interrelated and play complementary roles in ensuring the reliability, adaptability, and performance of AI models in real-world settings. Regular monitoring, uncertainty quantification, and robustness considerations collectively contribute to the development of trustworthy and effective AI systems. As such, 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.

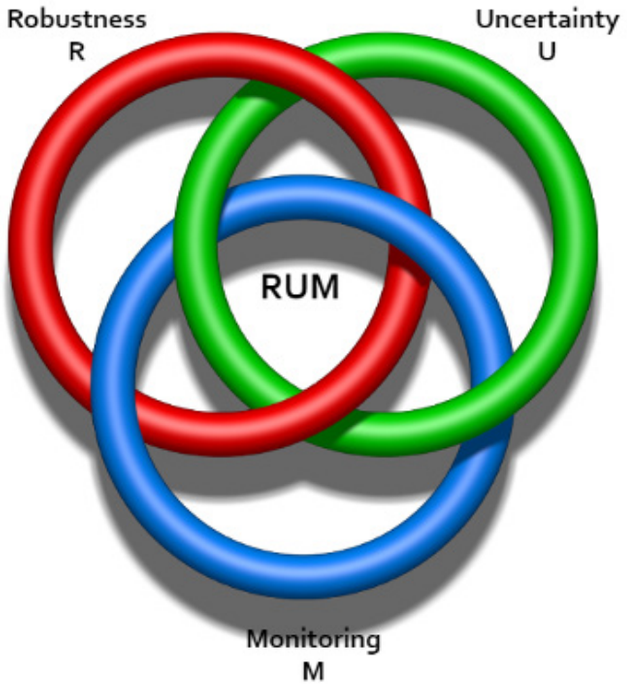


Figure 17: RUM methodology as three 3D loops topologically linked, i.e. any two such loops are only linked by the third one

Takeaways

- Confiance.ai has produced the "Trustworthy Environment" a framework providing modular components, methodological guidelines and proper documentation allowing to build interoperable tool chains to ensure trustworthy AI-based systems from design, all the way up to maintenance.
- Confiance.ai has produced nice Functional Sets available in their catalogue, they include components and methods allowing to tackle a specific topic of trustworthy AI.
- Functional Sets can and should be combined for specific needs where several properties are interdependent. The RUM methodology is an example on this regard proposed by Confiance.ai.

This translates exactly the fact that Robustness, Uncertainty and Monitoring methods can only successfully address their different challenges by working together. At a technical level, the RUM methodology helps to articulate and characterize different ODD zones to better detect possible failure modes, assess possible trade-offs or overall system-level compensations to be considered, which would have not been possible by the independent consideration of robustness UQ or monitoring apart from each other. When constituted with the RUM methodology, these zones are constructible in the context of the FS Data Lifecycle that will be presented in the following section. ■

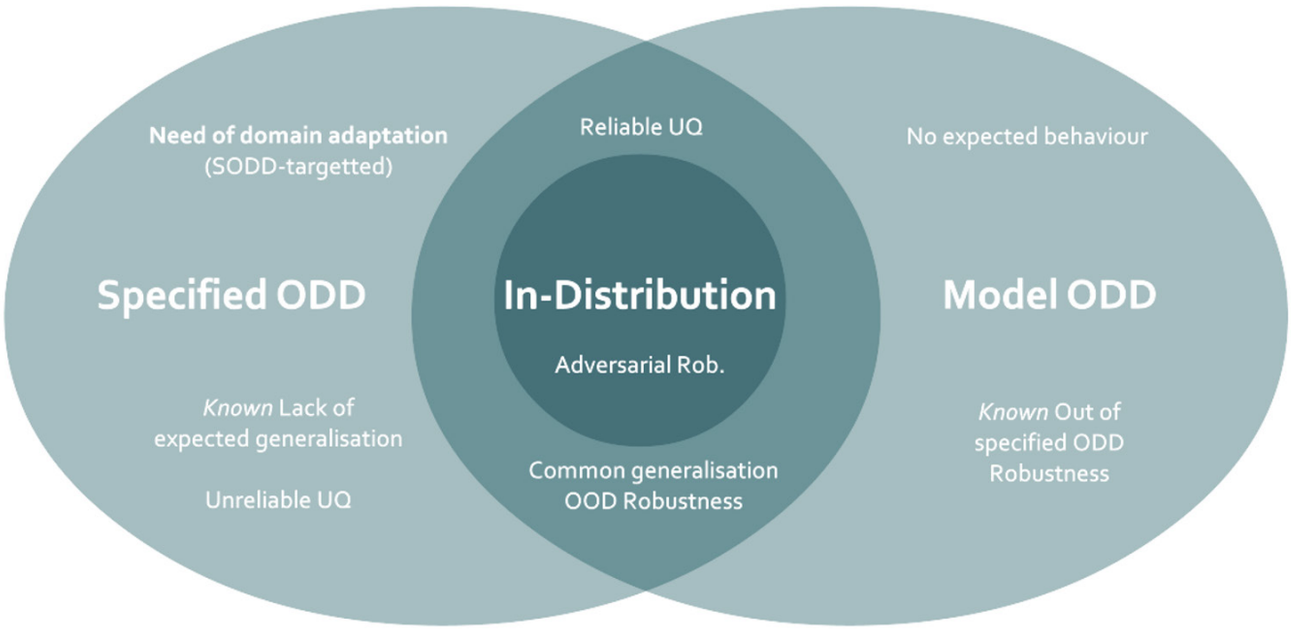


Figure 18: ODD Zones ought to be articulated with the RUM methodology

5. Deploying the End-to-End Approach

In this chapter we present two concrete output of the End-to-End approach from two uses cases in order provide an understanding of the kind of artifacts produced by Confiance.ai's tooled methodology

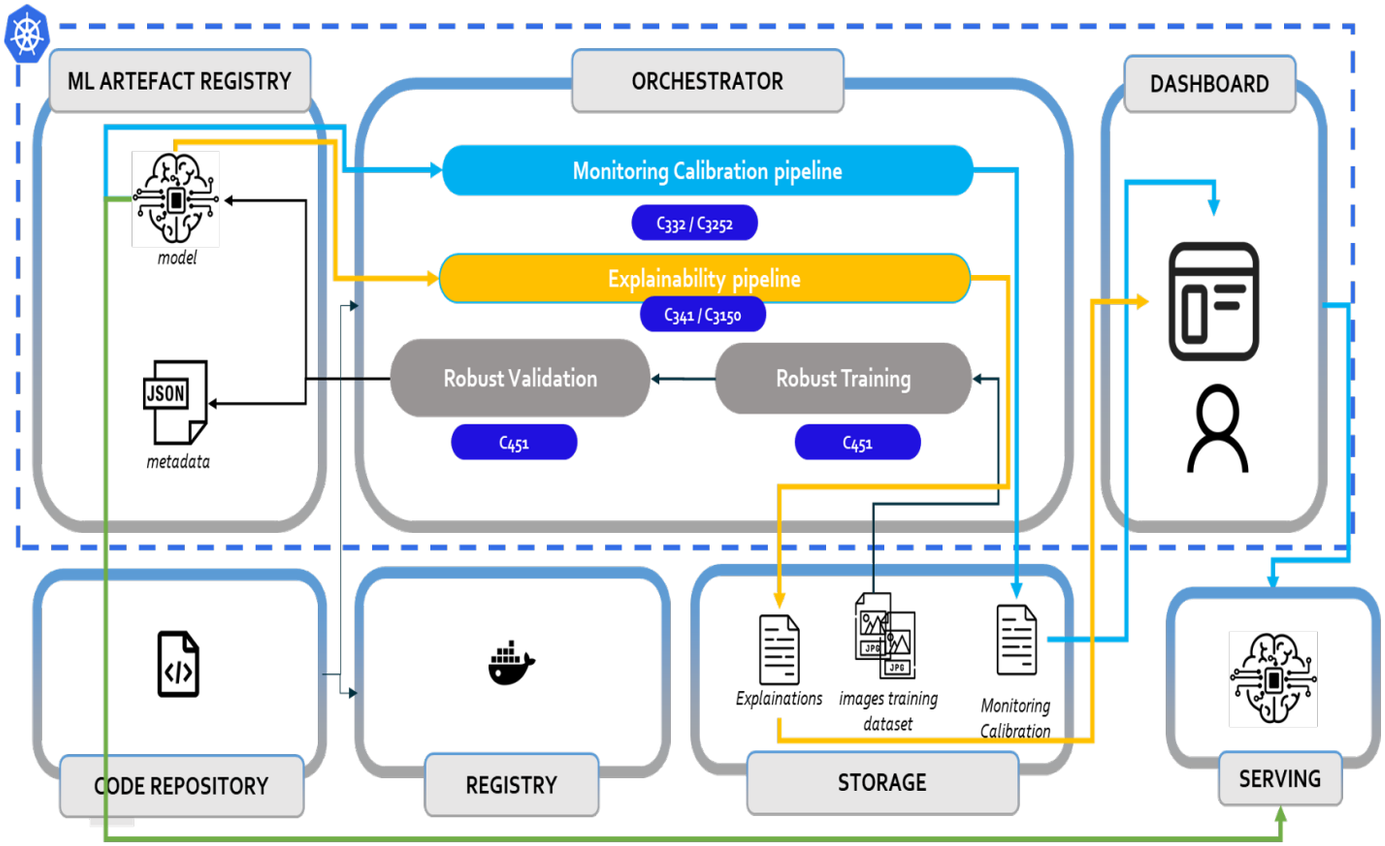


Figure 19: Tools integration inside a MLops pipeline

As the largest technology research program in the national AI strategy, Confiance.ai has evolved since its inception (2021), starting with a first year dedicated to covering the state of the art and pre-existing tools related to the integration and evaluation of data-driven AI. The following years (2022-2023) were devoted to the proper 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 with reference to pre-existing processes. The fourth and final year covers the evaluation of this End-to-End method, the dissemination and adoption by industrial of key results.

To facilitate the adoption of the tool-based methodology by industry, several implementations were carried out on use cases. These experiments illustrated the importance of combining several tools and methods to meet expectations in terms of trust properties. Here are two examples:

- For an autonomous driving use case, the diversity analysis of a dataset shows a low night-time image rate, which triggers the generation of synthetic night-time data. These data show a "domain deviation" and are subjected to "domain adaptation" before being integrated into the model's training data. These tools, implemented in the dataset construction method, will also be reused in the use case supervision stage.

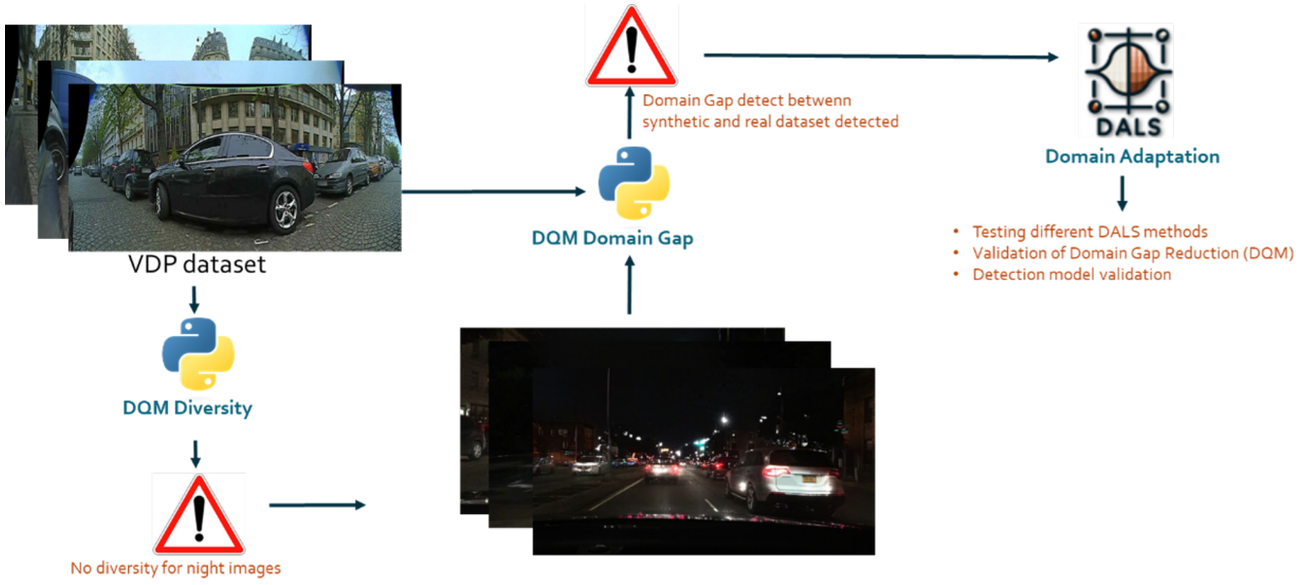


Figure 20: Dataset analysis and improvement

- In an aeronautical use case involving the detection of a runway, to consolidate the confidence score of an ML model, a data quality supervision module is added (see illustration). In this example, local image quality estimators (blur level, brightness, etc.) are considered in the detection zone where the runway is detected. These indicators are combined with model intrinsic indicators and used to build a confidence level for the AI component. In addition to providing a numerical value, this implementation is a tool to help interpret model errors and data when projected in the image.

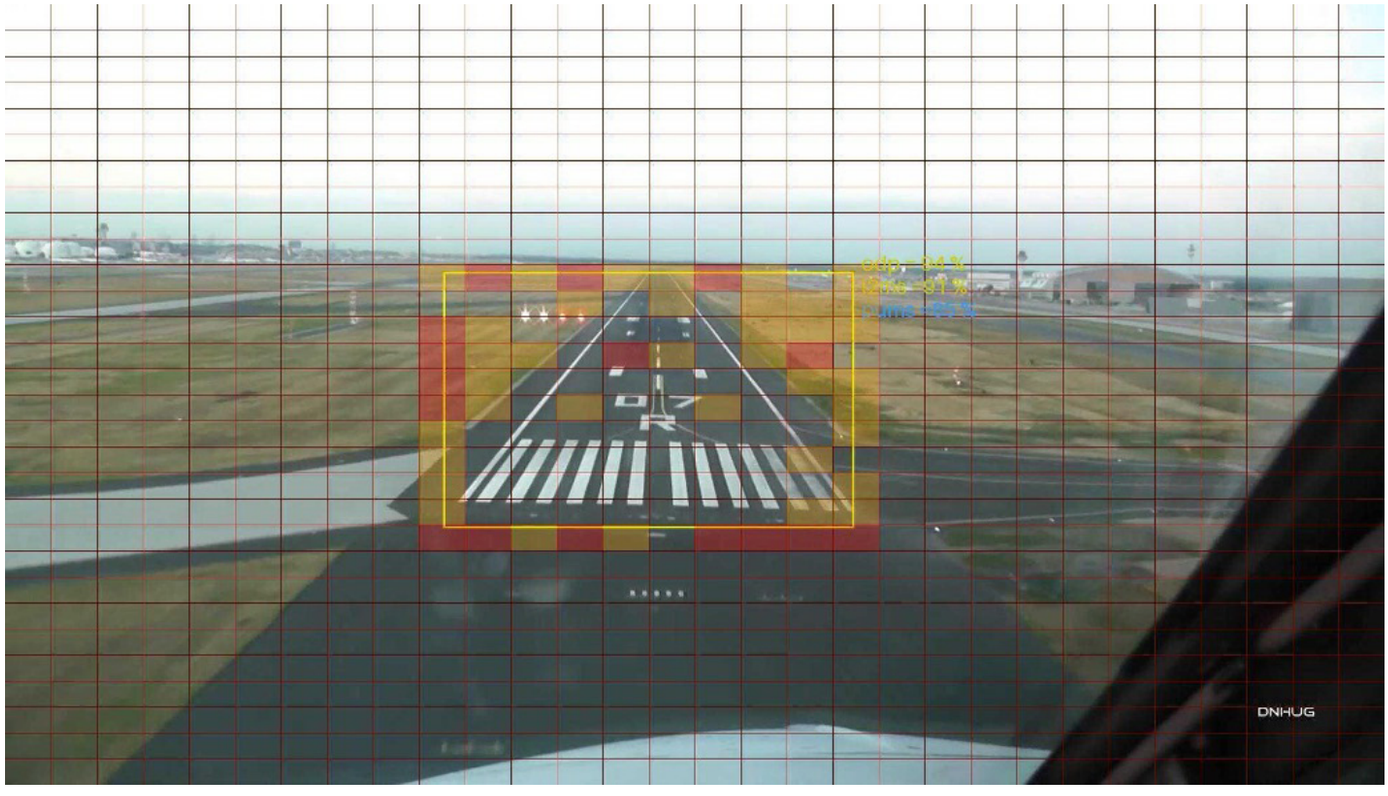
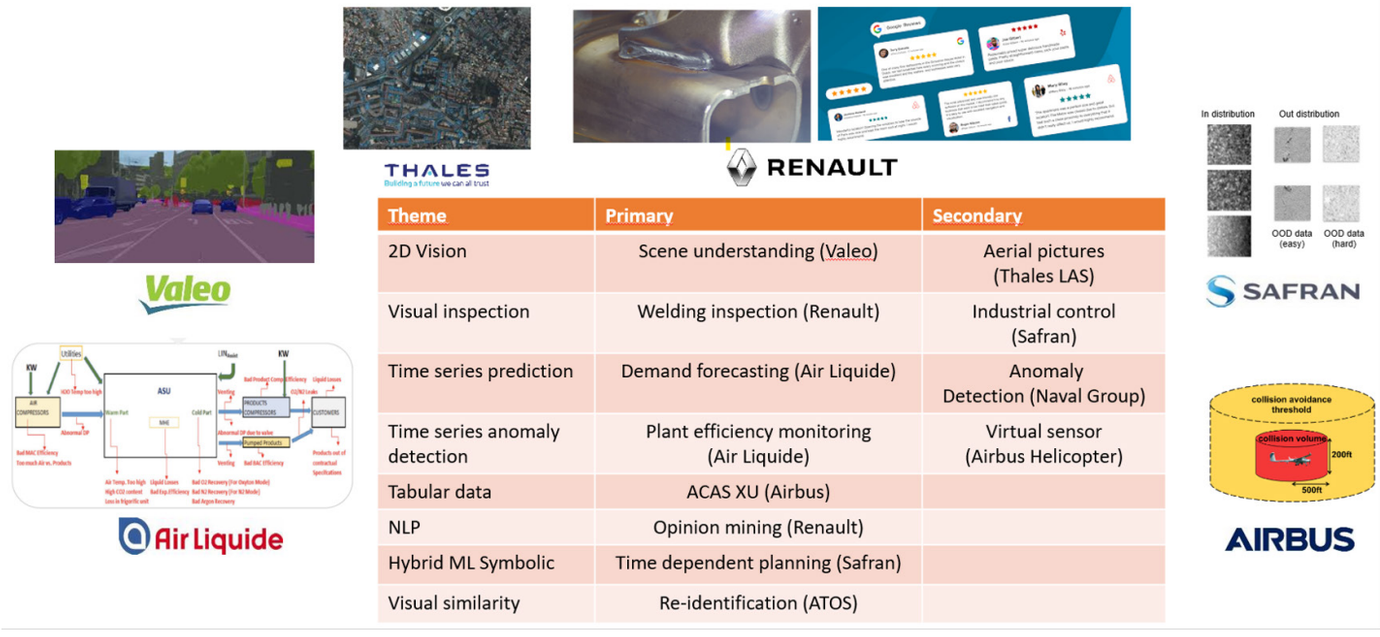


Figure 21: Monitoring indicator build to evaluate track detection quality

The end-to-end process evaluation for AI is a holistic approach that involves multiple stages (16 in our methodology), each aimed at ensuring the AI system's effectiveness, fairness, and alignment with both technical and ethical standards. By following this comprehensive framework involving different competencies (System engineer, Data engineer, Software engineer, …) in your organizations, you can build robust AI systems that deliver value while minimizing risks. ■

Takeaways

- There is no silver bullet tool that provides trust, we must combine several methods, and tools to build demonstration of achieved level of trust.
- Integrating tools and methods inside a MLops pipeline imply new technological challenge, but shall not be avoided to provide traceability and accountability on AI.



Theme	Primary	Secondary
2D Vision	Scene understanding (Valeo)	Aerial pictures (Thales LAS)
Visual inspection	Welding inspection (Renault)	Industrial control (Safran)
Time series prediction	Demand forecasting (Air Liquide)	Anomaly Detection (Naval Group)
Time series anomaly detection	Plant efficiency monitoring (Air Liquide)	Virtual sensor (Airbus Helicopter)
Tabular data	ACAS XU (Airbus)	
NLP	Opinion mining (Renault)	
Hybrid ML Symbolic	Time dependent planning (Safran)	
Visual similarity	Re-identification (ATOS)	

Figure 22: Use cases in the Confiance.ai program

6. The Context of the Trustworthy Environment

The execution environment designates an engineering workbench conceived as an MLops toolchain agnostic of technological adherence or constraints. First used to design the libraries and software component of Confiance.ai, this environment is now dedicated to evaluating the end-to-end design process of an AI-based component in accordance with rules, processes, methods and results produced and defined in the program Confiance.ai, through their implementation over industrial use-cases.

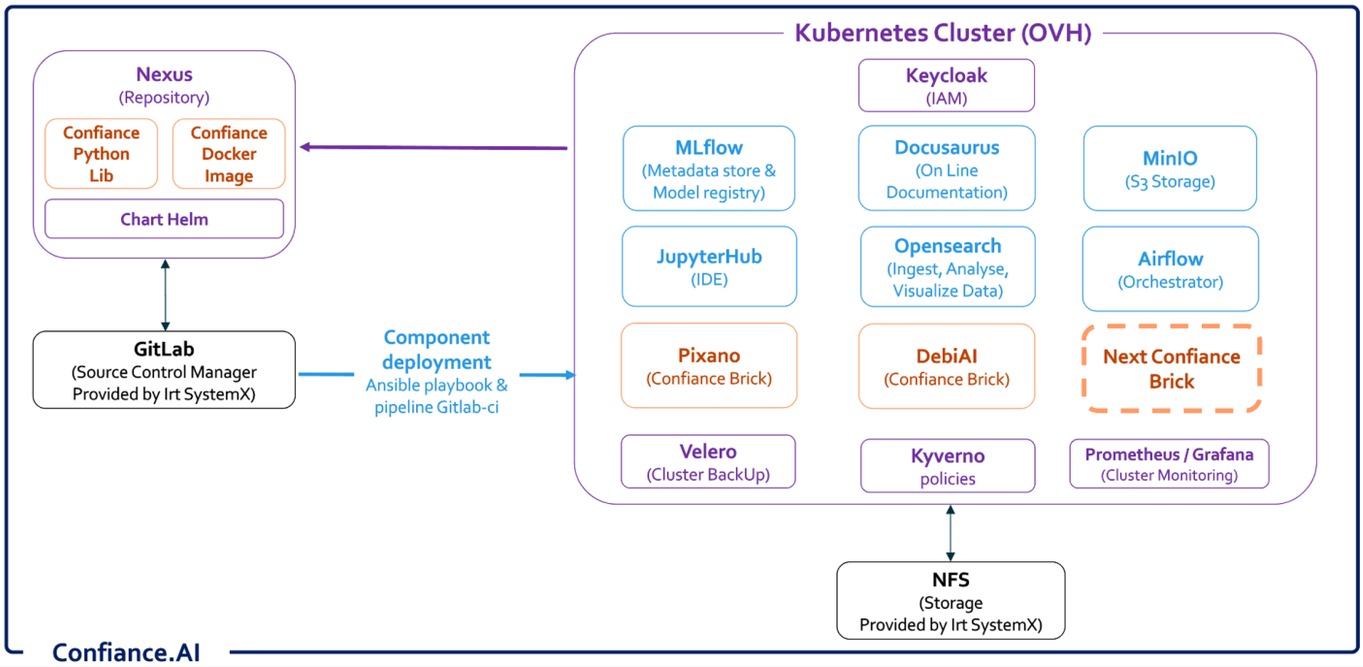


Figure 23: The execution environment architecture

This environment is designed in such as to allow:

- the manipulation of use-case data and model,
- collaborative working between the multiple contributors to the design process,
- the integration and rise in maturity of AI libraries and application,
- the exposition and sharing of theses libraries and application between different users,
- an acceleration of the industrial implementation of AI components in critical systems by partners,
- to be iteratively and consistently updated, integrating changes and evolution.

While it is made up exclusively of open-source components, which means that it can be redeployed on any custom infrastructure, using the documentation and scripts made available, it is also available as part of the Confiance.ai foundation to carry out component evaluation and maturation activities. ■

7. Reflective Summary and the Way Forward

Confiance.ai has made significant strides in addressing the scientific challenges associated with trustworthy AI in critical systems. By leveraging a robust scientific methodology and fostering collaboration with academic and industrial partners, the program has developed key components and platforms that lay the groundwork for future advancements. Moreover, Confiance.ai has cultivated a strong international community through various global events, further enhancing its influence. However, the journey is far from complete. The future of Confiance.ai involves continuing to tackle unresolved scientific challenges, ensuring the widespread adoption of its innovations across industries, and broadening its international influence.

7.1 Ensure Development of Industrial and Responsible AI

Confiance.ai is a cornerstone programme of the French national strategy for artificial intelligence, and a worldwide pioneer. The programme has helped position France as one of the global leaders in industrial and responsible AI by developing a sovereign methodological and technological environment which is open, interoperable and durable. It furthers integration of industrial (explicable, robust, etc.) and responsible (trustworthy, ethical, etc.) AI in strategic industries.

7.2 The Scientific Challenges that Remain Unresolved

The rise of generative AI presents both opportunities and challenges. Although generative AI was not the primary focus of the program, the methodologies and components developed within Confiance.ai have promising applications in this rapidly growing field. For example, generative AI is already being utilized to generate specific data or add auxiliary models within AI components. Future initiatives of Confiance.ai should consider integrating generative AI into the broader methodology, particularly exploring how foundation models can be effectively incorporated within AI components. Additionally, future work will need to extend efforts in areas such as:

- Cybersecurity for AI Components: developing robust strategies to protect AI systems from emerging cybersecurity threats.
- Bridging the gap between system-level activities and AI component design: ensuring a seamless integration of AI components within broader system architectures, with a focus on maintaining trustworthiness and performance.

7.3 Wide Adoption Across Industries

To maximize the impact of Confiance.ai, it is crucial to ensure that the developed components and platforms are widely adopted across various industries. This involves:

- **Scalability and Customization:** tailoring solutions to meet the specific needs of different industrial sectors, ensuring that they can be easily integrated and scaled.
- **User-Friendly Tools:** providing accessible documentation, training, and support to facilitate the adoption of these technologies by industry professionals, including those without specialized AI expertise.
- **Demonstrating Value in Real-World Applications:** conducting pilot projects and case studies that showcase the practical benefits of Confiance.ai's innovations in diverse industrial contexts.

7.4 Broaden its International Influence

Confiance.ai has already established a strong presence in France, but its potential extends far beyond national borders. To expand its impact a number of actions have been identified:

- **Strengthening International Collaborations:** building on existing partnerships and forging new ones with global academic institutions, industry leaders, and regulatory bodies to align with international standards and best practices.
- **Participation in Global AI Discourse:** continuing to engage in international events, conferences, and workshops to share Confiance.ai's findings, learn from global peers, and influence the global conversation on trustworthy AI.
- **Contributing to Global Standards and Regulations:** actively participating in the development of international AI standards and contributing insights from Confiance.ai's research and experience to shape future regulations. ■

Takeaways

The future of Confiance.ai community present opportunities to further its mission of developing trustworthy AI systems that are robust, scalable, and internationally recognized. By focusing on the unresolved scientific challenges, ensuring widespread industrial adoption, and expanding its global impact, the objective of the resulting initiatives of Confiance.ai is to address trustworthiness and operationalize it, leveraging the backbone that has been built to this day.

8. Annex

8.1 Release Notes

Down below the total deliveries of the Confiance.ai program is presented and the release notes including two main categories:

- **Components**: engineering tools, python libraries, web applications, demonstrators or experiments.
- **Documents**: taxonomy, methodological guidelines, the state of the art, benchmark, scientific contributions, user manual, conformity to standard, application of Confiance.ai components/methods in Use Cases and specification design document.

Number of integrated components	Number of delivered components (not integrated)	Number of delivered methodological guidelines	Number of delivered benchmarks	Number of SotA (in progress)
46	30	34	62	34

Table 1: Confiance.ai deliveries

Type	Mean	Description
Engineering tools	Software component	Tools needed to manipulate AI components and therefore to the generic operations associated with their realization. These tools exist outside the Confiance.ai program but have been selected and integrated because of the generalization of their use in the industrial domain, on the one hand, but also because of their compliance with the requirements of the program, especially in terms of intellectual property
Library	Software component	Python Library.
Web application	Software component	A Web Application (front + back or just back behind an API)
Demonstrator / Experimentation	Software component	This result is a demonstrator, it implements a method on a use case in order to evaluate its interest.
State of the art	Documentation	Provides a review of the current knowledge about the studied topic, through the analysis of the similar or related published works.
Benchmark	Documentation	Technical report that provides information on how several tools and/or methods compare to each other.
Methodological guideline	Documentation	Describes a clear and precise method allowing users to reach one or several stated objectives.
Scientific contribution	Documentation	Aims to deepen a specific question relating to an already existing theme.
Application of Confiance.ai components/ Methods in use cases	Documentation	The purpose of this document type is to test a product from Confiance.ai, whether a component or a method, in the context of a specific use case.
User manual	Documentation	User guides or manuals are the documents produced for the delivered software components/products.
Normative contribution	Documentation	In progress.
Specification/ design document	Documentation	Details the requirements, the expectations and the limits of a product or system.
Conformity to standard	Documentation	Establishes the adequacy between Confiance.ai processes and the concerns of an identified standard.
Taxonomy	Documentation	Proposes definitions for terms used within the Confiance.ai program, in relation to trustable AI-based systems. There is only one taxonomy.

Table 2: Release note content typology

8.2 Results on Functional Sets

Topic	Number of associated documents	Number of associated components
End-to-End	35	36
Data Lifecycle	30	46
Model Component	31	41
Deployment	27	29
Operation	34	48
Evaluation	28	42
Robustness	12	37
Uncertainty	4	9
Explainability	8	14

Table 3: Results on Functional Sets within Confiance.ai (see definition on section 4.2 "Functional Sets")

8.3 Robustness Components of Confiance.ai

The Robustness Functional Set offers three functionalities for users which can be used independently or together. The three use cases of the robustness platform are as follows:

- A user seeking to conduct a formal evaluation of their AI model;
- A user desiring to assess the robustness of their model against various types of perturbation;
- A user seeking to retrain a model that should be more robust than an old training against input perturbation.

Here we take the case of formal evaluation of robustness and we present the list of studied components in table below.

ID	Name	Type	Covered problems	Input format	UCs tested	Technical Maturity	Functional Maturity
321	Saimple	Python Library	Classification	ONNX Keras/TF Images	Welding	1	TBD
322	nnenum	Python Library	Classification	ONNX nnet Images/Tabular	Acas-Xu	2	TBD
323	α-β-crown	Python Library	Classification	ONNX PyTorch Images/Tabular	Acas-Xu	2	TBD
3171	PyRAT	Python Library	Classification	ONNX Keras/TF PyTorch nnet Images/Tabular	Welding Acas-Xu	2	TBD
391	MIP Solver	Python Library	Classification	ONNX Keras/TF Images/Tabular	MNIST CIFAR	TBD	TBD

Table 4: Components list of formal evaluation of Robustness

Next, we present the compatibility of the components with different types of data (inputs) in a first table, the compatibility of the components with different formats of neural networks in a second table, the applicability of the components on the use cases of the program in the last table.

In tables related to identify the applicability of components on use cases of the Confiance.ai program, three checkmarks are used:

✗ The component is tested on the Use Case.
✗ The component is untested with the Use Case but is compatible and testable.
✗ The component is incompatible with the Use Case.

In the same manner the assessment of components for Empirical Robustness Evaluation and improving robustness are provided in (Khedher, 2024).

Use Case	Saimple	nnenum	alpha-beta-crown	PyRAT	MIP Solver
Images	✗	✗	✗	✗	✗
Tabular	✗	✗	✗	✗	✗
Time-Series	✗	✗	✗	✗	✗
NLP	✗	✗	✗	✗	✗

Table 5: Components vs supported data type

Use Case	Saimple	nnenum	alpha-beta-crown	PyRAT	MIP Solver
Tensorflow	✗	✗	✗	✗	✗
PyTorch	✗	✗	✗	✗	✗
ONNX	✗	✗	✗	✗	✗
NNET	✗	✗	✗	✗	✗

Table 6: Components vs supported model type

Use Case	Saimple	nnenum	alpha-beta-crown	PyRAT	MIP Solver
Tensorflow	✗	✗	✗	✗	✗
PyTorch	✗	✗	✗	✗	✗
ONNX	✗	✗	✗	✗	✗
NNET	✗	✗	✗	✗	✗

Table 7: Applicability of components to use cases

9. References

Adedjouma, M. (2023). *Methodological Guideline for Operational Design Domain.*
Available on : https://catalog.confiance.ai/records/p4vvf-0h737.

Austin P.Wright, Z. J. (2020). *A Comparative Analysis of Industry Human-AI Interaction Guidelines.*
Retrieved from https://arxiv.org/abs/2010.11761

Benoit Langlois, R. B. (2024). *Methodology for Dataset Development.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/t88d9-a9114

Blanc, B. L. (2024). *Expérimentation de la confiance d'un utilisateur de système d'IA.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/prsn1-02m59

Bohn, K. M. (2024). *Methodological Guideline for Operational Approach.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/05zsa-rx870

Confiance.ai. (2024a). *Body of Knowledge,* Confiance.ai.
Available: https://bok.confiance.ai/. [Accessed April 2024].

Confiance.ai. (2024b). Retrieved 07 2024, from Confiance.ai, Our contribution in support of future European AI regulation: https://www.confiance.ai/en/our-contribution-in-support-of-future-european-ai-regulation/

Dejean, P. (2023a). *Benchmark on methods and metrics for Explainability.*
Access link: https://catalog.confiance.ai/records/n45k0-f0t97.

Dejean, P. (2023b). *Description and analysis of the components in relation to the taxonomy and standards.*
Retrieved from https://catalog.confiance.ai/records/rysat-3h377

Dejean, P., Arlotti, C., & Heulot, N. (2024). *Analysis and Capture protocol of use case holders needs and requirements.*
Access link: https://catalog.confiance.ai/records/vny3p-v9s52.

Jenn, E. (2023). *Methodological Guideline for Assurance Case.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/pbqq6-e4a11

Jenn, E. (2024). *Methodological Guideline for Assurance Cases Evaluation*. Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/6a7n0-ca528

Khedher, M. I. (2024). Methodological Guideline for Robustness Functional Set. Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/km6fw-qsq36

Mantissa, K., & Bohn, C. (2024). *Methodological Guideline for System Approach.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/8y81y-szd28

Mattioli, H. S. (2023). *Methodological Guideline for Trustworthy AI Assessment.*
Access link: https://catalog.confiance.ai/records/z1f55-7t378.

Poche, A. (2023). *Methodological Guideline for Explainability.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/jjbzz-h0z58

Liano, D. G. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (pp. 1-15).
Retrieved from https://dl.acm.org/doi/10.1145/3313831.3376590

Liano, H. S. (2023). Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. *In proceeding of the 2023 CHI conference on human factors in computing systems*, (pp. 1-21).

Robert, B. (2024). *End-to-end approach for engineering trusted AI-based systems.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/rnqvz-atg47

SAE J3259. (2021). *Taxonomy & Definitions for Operational Design Domain (ODD) for Driving Automation Systems*.
Retrieved from https://www.sae.org/standards/content/j3259/

Amershi, D. W. (2019). *Guideline for Human-AI Interaction*. In proceddings of the 2019 Chi conference on human factors in computing systems.
Retrieved from https://dl.acm.org/doi/10.1145/3290605.3300233

Shneiderman, B. (2021). *Responsible AI: Bridging from ethics to practice* (8 ed., Vol. 64). Communications of the ACM.
Retrieved from https://doi.org/10.1145/3445973

Shneiderman, B. (2022). *Human-centered AI*. Oxford University press.

Sohier, H. (2024). *Position de Confiance.ai par rapport à l'AI Act.* Confiance.ai.
Retrieved from https://catalog.confiance.ai/records/bz98p-9wz71

Weinstock, C. B. (2015). *Assurance Cases.* Software Engineering Institute.

Notes

To learn more about
the Confiance.ai program:

www.confiance.ai

ConfianceAI

Confianceai

Director of publication:
Michel MORVAN

Editorial director:
Aurélie BOURRAT

Chief Editor:
Samanta DUGUAY-FANTI

Pictures:
Confiance.ai, Shutterstock

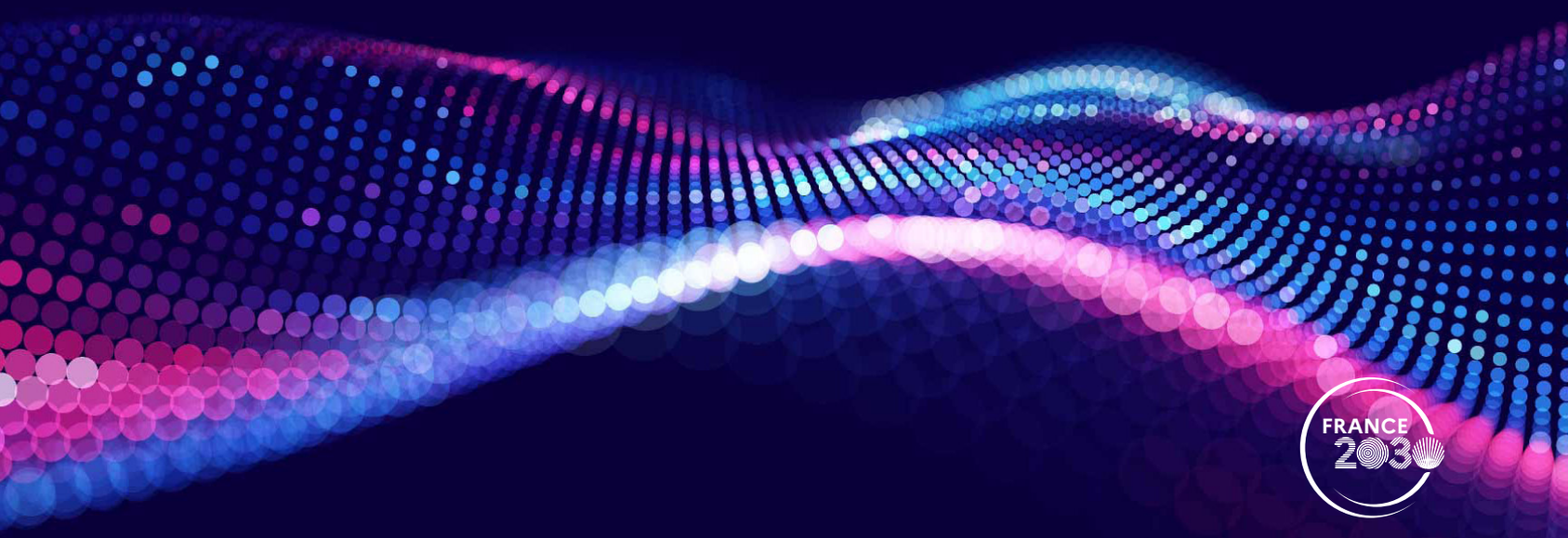
Graphic design:
www.maiffret.net

Contact:
contact@confiance.ai

Founding members

Confiance.ai

Driven by a group of 13 major French companies and research organisations, Confiance.ai is a cornerstone programme of the French national strategy for artificial intelligence. Launched in January 2021 and financed through France 2030, the ambition of this 4-year project is to design a platform of sovereign, open, interoperable and sustainable methods and tools that will enable trustworthy AI to be integrated into critical products and services. It brings together some fifty industrial and academic partners in Saclay and Toulouse around seven R&D projects. Confiance.ai contributes to the implementation of the AI Act led by the European Commission.



FRANCE 2030

www.confiance.ai

ConfianceAI

Confianceai