Towards the engineering of trustworthy AI applications for critical systems

The Confiance.ai program

October 2022

Table of contents

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 has been trained and validated with data representing all the operational conditions that may be encountered[1].

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 known 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. ■

[1] As will be seen later in this document, the subject of Operational Design Domain (ODD) is clearly a key issue in AI trustworthiness for critical systems

The Confiance.ai program

The program Confiance.ai is the technological pillar of the Grand Défi "Securing, certifying and enhancing the reliability of systems based on artificial intelligence" launched by the Innovation Council. The two other pillars focus on standardization (norms, standards and regulation toward certification) and application evaluation.

Confiance.ai is the largest technological research program in the national AI strategy. It tackles the challenge of AI industrialization, as the very large-scale deployment of industrial systems integrating AI is a crucial stake for industrial and economic competitiveness. It has a strong ambition: breaking down the barriers associated with the industrialization of AI and equipping industrial players with methods and tools adapted to their engineering. One originality of the program lies in its integrative strategy: it globally addresses the scientific challenges related to trustworthy AI and provides tangible solutions that can be applied in the real world and are ready for deployment in operations.

As defined by the European commission and according to the new proposed regulation, the AI Act, trust is the key objective for a deployment in accordance with the European values. It can be defined through various points of views, details and encompass both engineering and usage aspects. Even if Confiance.ai has to consider all aspects, a particular effort is made to propose concrete and pragmatic answers for system and software engineering methods able to allow certification of AI based systems according to their criticality levels. The program adopts a strategy of progressive advancement: during the first year of the program, data-based AI solutions, mainly using neural networks, are the focus of research with application on image processing, time series and structured data. Then, in the following years, more complex problems and relevant industrial use cases will be looked at. Use cases using video, audio and text data will be added, as well as the introduction of other AI formalisms including knowledge-based and hybrid approaches. At the end of the program, the program will cover the whole spectrum of critical systems.

■ A research and development validation strategy based on industrial use cases

Confiance.ai is an industry-oriented project. Its outputs are expected to be usable by industrial partners within their system and software engineering process. A way to achieve this objective is to validate the produced methods and tools on industrial use cases.

Use cases are formally defined by (I) a feature implemented with AI technologies; (II) access to the data or the knowledge base used by the feature (III) a qualification/certification issue raised by relevant authorities (ranging from certification bodies to quality processes internal to the use-case provider); (IV) involvement of the feature product owner himself for the evaluation of the proposed methods and tools

To reach its goals, the program must perfectly understand the arguments that will convince authorities. That is the reason why the involvement of the product owner is crucial. Each tool provided by the project should be a step towards the demonstration of the AI-based system safety. As being developed in a research project, the AI-based features are often under development or at POC status, their integration in critical industrial systems is not expected at short term when plenty of other critical issues are to be managed today. The connection between trustworthiness system requirements and the software technical proposed solution at the component level will be the major challenge of the project. ■

1. The need for trustworthy AI on critical systems

1.1 Why can't we trust AI?

The imaginary example of dangerous, but not malicious, AI given in the introduction of the document illustrates the kind of reasons why industry, but also people, needs a trustworthy AI. If it is not considered as dependable, AI will not be accepted by people, by regulation, by safety and quality managers.

The given example points out a problem about the specification of the task to fulfill and of the environment in which this task must be executed. This problem is common with other technologies but is exacerbated by the way AI allows a very intuitive way to describe a problem to solve. What is intuitive for the human intelligence is often incomplete for AI.

But on the way to trust, AI faces two other big obstacles. The first one is mathematical fact: being mainly based on complex non-linear functions, AI is often not robust. It can give very good results on a very complex situation but a totally wrong answer for a very close situation. The second one is probably the most difficult to tackle for engineers: people are basically reluctant in front of this mysterious technology looking like a spying witchcraft[2].

● Problem description

In machine-learning based AI, everything relies on examples given during the training and validation phase. It is the main interest of AI: it is taught like humans are, with examples not with equations. Thus, if examples are not representative enough of the real operation domain (only pink and blue gloves), the AI will face unknown situations (yellow gloves) without any possibility to process them and probably no way to recognize that it should not process them. If any level of human intelligence is able to realize that a situation does not fit with any of given examples, it requires a very sophisticated AI to detect that a situation does not belong to the training set. Yellow gloves are not considered as gloves because gloves are blue or pink. If it is not blue or pink, it is not a glove, there is no hand in it so no reason to stop the robot. The AI does not have, spontaneously, this common sense that helps humans to deal with unexpected situations. During its training, a very performant autonomous car had never experimented the picture of a truck laying on the side in the middle of the highway. When this situation occurred during a journey, the damaged truck was not classified as a known obstacle, thus, for the AI, it was not an obstacle, and the vehicle went through it. We cannot trust AI if we cannot guarantee that its training set covers the operational domain of the system or that events out of it can be dealt with by a complementary component. How to guarantee this covering when the operational domain is as vast as the number of situations that an autonomous car can meet on every road of Earth?

The other major kind of AI, that is not based on data, examples, and machine learning, is the knowledge-based AI, as known as symbolic AI (expert systems, rule-based systems, constraint programming…). Beyond the fact that symbolic AI cannot solve the same kind of problem than the machine learning based AI, the knowledge representation gives access to a higher-level description of the task and its environment than examples can do. Thousands of examples can be resumed in a single rule. In any case, the question of operational domain covering is still relevant. How can we be sure that 2000 rules are enough to cover all the driving situations that an autonomous car can meet in real life? If 10000 rules are necessary, how could we demonstrate that some are not contradictory in particular situations?

● Robustness

The adversarial examples are spectacular illustrations of the lack of robustness of some AI systems. If an invisibly modified picture of a panda can be recognized as a picture of gibbon or if a tagged stop sign is recognized as a 30mph sign (these are two well-known example of adversaria attacks), it is impossible to trust AI. Of course, this kind of attacks is a scientific demonstration of this weakness of data-based AI. They are the result of malicious computations that cannot be generated by pure randomness in the real life. At any rate, it instils the doubt about the general robustness of AI. If the issue of the macroscopic coverage of the operational domain can be solved by gathering a big enough training data base, is it possible to envisage all the microscopic variations around the training examples. In engineering of automatized systems, robustness is a very concrete feature: how does the system behave when it is pulled out of its nominal state? For an AI-based perception system, is it possible to guarantee that the classification will stay the same if the lightning changes? If a panda becomes a gibbon with only a small percentage of changed pixels, will a red traffic light become green if the sky is grey?

The panda adversarial example also questioned the meaning of the confidence score that are given with the classification decision. The confidence score in the gibbon classification was much higher than the score of the original panda classification. A wrong classification with a very low score would have been acceptable. It is another highlight of the lack of robustness. It stresses the importance of the estimation of the imprecision on decisions taken. Classical engineered systems are able to evaluate the accuracy margin of their computation. The decision process takes this margin into consideration to adapt its conclusion. To become part of a trustworthy system, AI should be able to evaluate the accuracy of its conclusions.

[2] The section 3 gives more details about the sources of mistrust that the project should tackle.

● Explainability and bias

One can mathematically demonstrate that a plane can fly. One can prove statistically that flying is safer than driving a car. But it will not totally convince your best friend that they can relax during the flight. Technical arguments are necessary but not sufficient. Even when we are able to tackle the technical issues presented before, it is probable that acceptability of AI, thus trust in AI, should also consider other kinds of aspects.

AI, and particularly deep learning models, are often considered as black boxes. Reading model architectures does not give a clue, to ordinary people, on the way a classification has been performed. The magical flavor of some AI systems does not reassure about the way they work. Explanations are a very important way to give confidence in an AI system. By explaining how it has taken its decision, the AI model comes closer to our own way of reasoning. If we can understand the features that the AI has considered, we can be convinced that it has "understood" the problem it had to solve. When the AI considers the snow in the background to classify an animal in the foreground, it appears that the problem has not been clearly exposed during the training phase. Explainability is thus also a way to detect bias in data bases. If the most important feature considered by an AI used for recruitment is the gender or the place of birth, there is probably a bias in the way recruitments have been made so far. Even if it denotes a failing of the previous human behavior, not a failing of a current AI system, AI is expected to give tools to detect bias and bring transparency in a system as critical as recruitment process.

● Certifiability and risk analysis

As integrating AI based component in critical systems is the target of Confiance.ai, the question of certifying them is central. The certification processes of usual critical systems is based on sound and well-established processes, and now the question is if intrinsic properties and current development processes of AI components are compatible with them?

In practice, it appears that introducing AI without specific attention can break the basic principle of certification processes. Indeed, certification of critical systems requires clear and well formalized requirements and specifications, continuous traceability of their refinement, implementation and verification. AI development misses often (in particular for machine learning AI but also for quite all other technologies), well formalized requirement and specifications, the capability to trace their refinement and implementation and their assessment through verification and validation methods.

In addition, the notion of risk itself has also to be revisited to identify risks inherent and specific to AI based component. Defining a taxonomy of the risks and linking them with management mechanism is essential to support certification processes.

This is why Confiance.ai is devoting a major effort to define methods and processes that can be integrated in existing engineering practices. An important point there is to make explicit the links with, on one hand, proposed methods for the design and evaluation of AI components and, on the other hand, with the requirements and recommendations of the existing, or emerging, standards and regulations for AI in general and in the different application domains.

1.2 Exemplary use cases

The formal demonstration that AI can be trustworthy and dependable is still a scientific challenge in most of the cases motivation numerous research projects in France, Europe and globally in the world (see for example the dedicated European network as TAILOR and workshops associated to main international conferences as AI Safety at IJCAI, SafeAI at AAAI, WAISE at SafeComp or SAIAD at ECCV). In particular, the DEEL (Dependable and Explainable Learning) project gathers several French and Canadian scientists to work on the theoretical bases of a trustworthy ML. The aim of Confiance.ai is to provide tools and methods to guarantee a proper functioning of AI based systems. When formal methods can prove the good behavior of AI, they will be used. But in many cases, they will not be sufficient. Confiance.ai will propose pragmatic ways to convince with reasonable arguments AI stakeholders (designers, approbators, users) that it is safe to use AI-based components in a given context for critical systems.

To stay pragmatic, the Confiance.ai project decided that the work should be guided by industrial use cases. Every development, tool or method proposed within the project should answer to, at least, one industrial use case selected by the partners. An industrial use case is an application of AI, developed by a partner but that cannot be deployed in the company because approbators don't trust it. A typical example is the welding quality inspection proposed by Renault. The development team of Renault developed a model that is able to detect if a welding on the chassis is successful or not. This is a simple classification problem. The training is unbalanced because welding is generally good, but the developed system got a very good recognition score: more than 97%. However, the quality engineer refused the deployment of the solution in the factory because score was not a good enough criterion. Renault brought this use case to the Confiance.ai to get methods and tools to develop an AI that fits the expectations of the quality engineer.

For Confiance.ai a use case is made of annotated data to train and validate the system (or knowledge, for symbolic AI), a trained model (or inference engine), with a correct level of performance, and one or several trust issue (robustness, explainability, uncertainty, certification process…). All of this should be shareable, with the partners, and public enough to be shared with the future users of the Confiance.ai deliverables. Finally, the use case must be carried by a use case owner, able to describe accurately the use case constraints and to represent the expectations of the final approbator. These requirements can make difficult the proposition of use cases by industrial partners. Table 1 gives an overview of the selected use cases.

Thematic	Topic	Carrier	Kind of data
2D Vision	Road scene understanding	Valeo	2D pictures
	Aerial picture	Thales	2D pictures
	Visual similarities	Atos	2D pictures
Visual inspection	Welding inspection	Renault	2D pictures
	Industrial control	Safran	2D pictures
	Cylinder counting	Air Liquide	2D pictures
Time series	Demand forecasting	Air Liquide	Numeric values
	Plant efficiency monitoring	Air Liquide	Numeric values
	Predictive maintenance	Airbus	Numeric values
	Anomaly detection	Naval Group	Numeric values
Surrogate Model	Collision risk evaluation	Airbus	Tabular data
NLP	Opinion mining	Renault	Free text
	Ontology update	Thales	Free text
Hybrid AI	Dynamic planning with uncertainty	Safran	Constraints

Table 1: Use cases submitted to Confiance.ai

To be shared with the consortium, some of the data have been anonymized by their owner, keeping their statistical characteristics but losing their semantics. It was the case for the Air Liquide's data about demand forecasting: Air Liquide wanted to have an estimation of the uncertainty about the forecasting of certain product but neither the variable name, nor the value of the previous production was recognizable in the supplied data.

For natural language processing (NLP) use case on opinion mining, it is not possible to get rid of the semantic. Nevertheless, Renault was not ready to share the raw (good or bad) comments from its customers. Thus, Renault did not share the data gathered by its after-sales teams, but the public data found in the Google review of its branches. This already public data can be shared with the partners to validate the methods and tools dedicated to NLP.

If a critical use case can't be shared for any reason, the concerned partner can import the Confiance.ai environment and apply it on its own to its confidential use case. If some troubles are met, they will be shared with the Confiance.ai consortium to be processed.

1.3 Norms and standards

The Confiance.ai project will recommend methods and tools to develop a trustworthy AI. If industry is willing to apply best practices to ensure quality of its product, it is difficult for it to determine, by itself, what these best practices are, particularly when it does not concern its core business. By working together, the nine industrial partners of Confiance.ai, with the support of the academic partners, can give a larger base to their recommendations but it is a "local" agreement about the ways should be done. If the recommendations of Confiance.ai were part of a standard, they would be much better accepted and adopted by Industry.

As a matter of fact, the Confiance.ai project is one of the three pillars of the French "Grand Défi" (Grand Challenge) on Reliable and Certifiable AI. Confiance.ai is the technological pillar, producing software. The second pillar concerns the validation of the AI-based system. The third one is about standardization. It has the aim of bringing the output of Confiance.ai into norms.

Many initiatives are proposing to normalize AI. Each sectorial domain prepares the introduction of AI in its systems (Aeronautics SAE-AS6983, Automotive 26262, Sotif,….). Furthermore, each country has relevant propositions (Canada, China, US, Europe, UK…). For instance, the European Union is working on the AI Act that will give the main guidelines about the way an AI-based system should be used and developed. The AI Act has a transversal approach (covering all the sectorial domains) that should be compatible with the sectorial constraints. The AI Act gives high-level objectives ("the training data set must be representative of the operational domain") but does not give a standardized way to reach them.

CEN/CENELEC is one of the standardization organizations that has been designated by the European Commission to develop a standard on AI. In addition, the EC has launched the Adra-e CSA to define roadmap for AI, Data and Robotics including analysis of standards and regulation on AI in collaboration with CEN-CENELEC. CEN-CENELEC Joint Technical Committee (JTC) 21 is focused on AI. Eight Ad hoc groups (AhG) have been defined to contribute to the different aspects of the standard. Confiance.ai oversees two of them: the AhG7 on overarching unified approach on trustworthiness characteristics, the AhG8 on risks (risk catalogue and risk management).

In parallel to this initiative for aligning future AI standards on Confiance.ai propositions, the project also evaluates if these propositions are compatible with existing standards like SAE-AS6983 (Approval of Aeronautical Safety-Related Products Implementing AI) and IEEE 7000-2021 (Standard model process for addressing ethical concerns during system design). If methods and tools proposed by Confiance.ai can address some aspects of these standards, the project will validate a first step towards success.

2. Engineering Trustworthy AI Components and Systems

The objective of the Confiance.ai program is to revisit legacy engineering practices with regard to the effects induced by the use of AI in critical systems. A rigorous and interdisciplinary approach is needed to formalize the design and validation of these critical systems based on AI to ensure “safe & secure” deployment and maintenance over their lifecycle, so as to bridge the gap between Proof of Concepts and actual deployment of dependable (systems of) systems involving AI algorithms, compatible with business needs. This approach shall address collaborative engineering including disciplines such as:

- Algorithm engineering,
- Knowledge engineering,
- Software engineering,
- Hardware engineering,
- System engineering,

And also engineering specialties like, in particular:

- Safety engineering,
- Cybersecurity (including privacy) engineering,
- Human factors & cognition engineering.

2.1 Engineering Trustworthy AI Systems[3]

The operational exploitation of AI is relatively recent, determined by the spectacular improvements of algorithms and the hardware components executing those algorithms. Initially exploited for non-critical tasks showing no or a very low level of risk, building an AI system was essentially a matter of combining ad-hoc engineering practices with the objective of providing “usable” results in the most cost-effective way. When it comes to industrial critical systems, several additional constraints must be considered. First, processes must be rationalized, justified, made reproducible, optimized, etc. Second, processes must ensure that the overarching properties of the system under design are actually satisfied with the appropriate level of confidence:

- the defined intended behavior of the system is correct and complete with respect to the effective desired behavior,
- the implementation of the system is correct with respect to its defined intended behavior, under foreseeable operating conditions,
- any part of the implementation that is not required by the defined intended behavior has no unacceptable safety impact.

In the context of Confiance.ai program, we propose to reconcile these two approaches, namely learning from ad-hoc engineering practices on the different use cases on the one hand, and structuring reproducible processes for achieving appropriate level of confidence on the other hand. To achieve this reconciliation, we need an analysis framework able to capture and organize engineering contexts, constraints, activities, data, lifecycle, etc. concurrently under different viewpoints, in order to build a global and comprehensive model. Each viewpoint enriches others in an iterative and incremental, multi-viewpoint analysis. The Trustable AI analysis framework designed by Confiance.ai program allows to elaborate the strategies for System Development Activities and for IVVQ Activities and is contributing to the specification of the Trustworthy Environment.

The approach consists in:

1. Defining analysis viewpoints, formalized in a modeling tool by a meta-model containing the definition of involved concepts and semantic relationships between these concepts.
2. Consolidating the methodological outputs of the Confiance.ai projects by analyzing their various aspects: engineering context, constraints, activities, data, lifecycle, etc.
3. Formalizing the analysed methods in a modelling tool, according to the meta-models of the considered viewpoints. The modelling will help ensuring that all methods are compatible with each other, and that the Confiance.ai program will produce a consistent end-to-end process allowing the design of dependable AI-based systems.

The list and definition of viewpoints is likely to evolve during the application of this approach.

[3] based on: Morayo Adedjouma, Christophe Alix, Loic Cantat, Eric Jenn, Juliette Mattioli, et al.. Engineering Dependable AI Systems. *17th Annual System of Systems Engineering Conference 2022*, IEEE, Jun 2022, Rochester, United States. (hal-03700300)

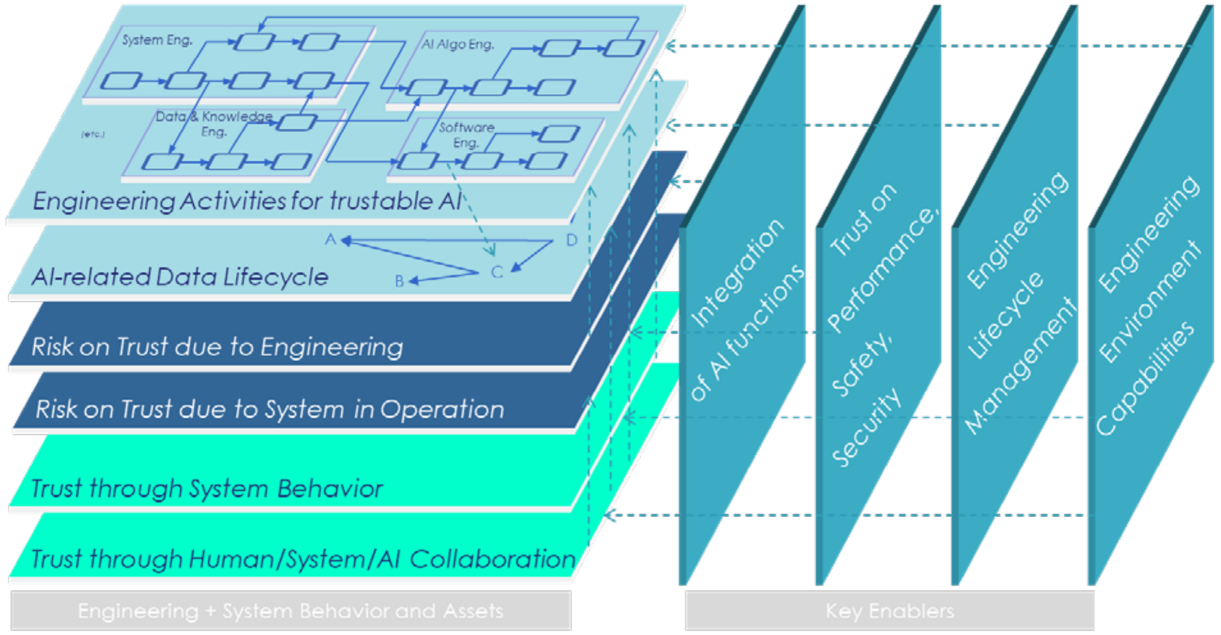
System Eng.
AI Algo Eng.
Data & Knowledge Eng.
(etc.)
Software Eng.
Engineering Activities for trustable AI
A
D
C
B
AI-related Data Lifecycle
Risk on Trust due to Engineering
Risk on Trust due to System in Operation
Trust through System Behavior
Trust through Human/System/AI Collaboration
Engineering + System Behavior and Assets
Integration of AI functions
Trust on Performance, Safety, Security
Engineering Lifecycle Management
Engineering Environment Capabilities
Key Enablers

Figure 1: Global view of the analysis framework

As of today, the viewpoints for System Development are the following:

- Two generic viewpoints:
  - Engineering activities for trustable AI: Define the tasks to perform so as to specify, design, produce, deploy and operate an appropriate and trustable solution to a well understood need, involving AI techniques,
  - AI-related data life cycle: Identify major data required/produced by AI engineering, when they are produced/used, and how they evolve with time,
- Two viewpoints dedicated to risk on trust (ie. risk that the trust on the capability of the system to deliver the expected service is reduced or lost):
  - Risk on trust due to engineering: identify major sources of bias or errors brought by other engineering activities to inputs and outputs of AI engineering and data,
  - Risk on trust due to system during operation: identify major sources of bias or corruption brought by other system components interacting with AI components in operation,
- Two viewpoints dedicated to trust development and support:
  - Trust through system behavior: define major system capabilities needed to ensure Trust in Operation,
  - Trust through Human/System/AI Collaboration: define expectations of human stakeholders, their role and workshare with System AI, in delivering the expected services in a trustable manner.
- Four transverse System viewpoints are also identified:
  - Integration of AI functions: characterize and address specific concerns related to integrating one or more AI functions together in system target context; deliver guidance on how to manage each concern,
  - Trust on Performance, Safety, Security: define main needs, contributions and obstacles regarding Trust applied to AI decision performance, safety, and security of the global solution including AI,
  - Engineering Lifecycle Management: define processes to revisit engineering choices and decisions according to evolution of context, environment and needs,
  - Engineering Environment Capabilities: define the tooling support required to make trustable AI systems engineering feasible, scalable, efficient and secure.

The analysis framework implementation is derived from Capella toolbox. The figure below shows an example of engineering activities viewpoint, limited to Machine Learning (ML) algorithm engineering.

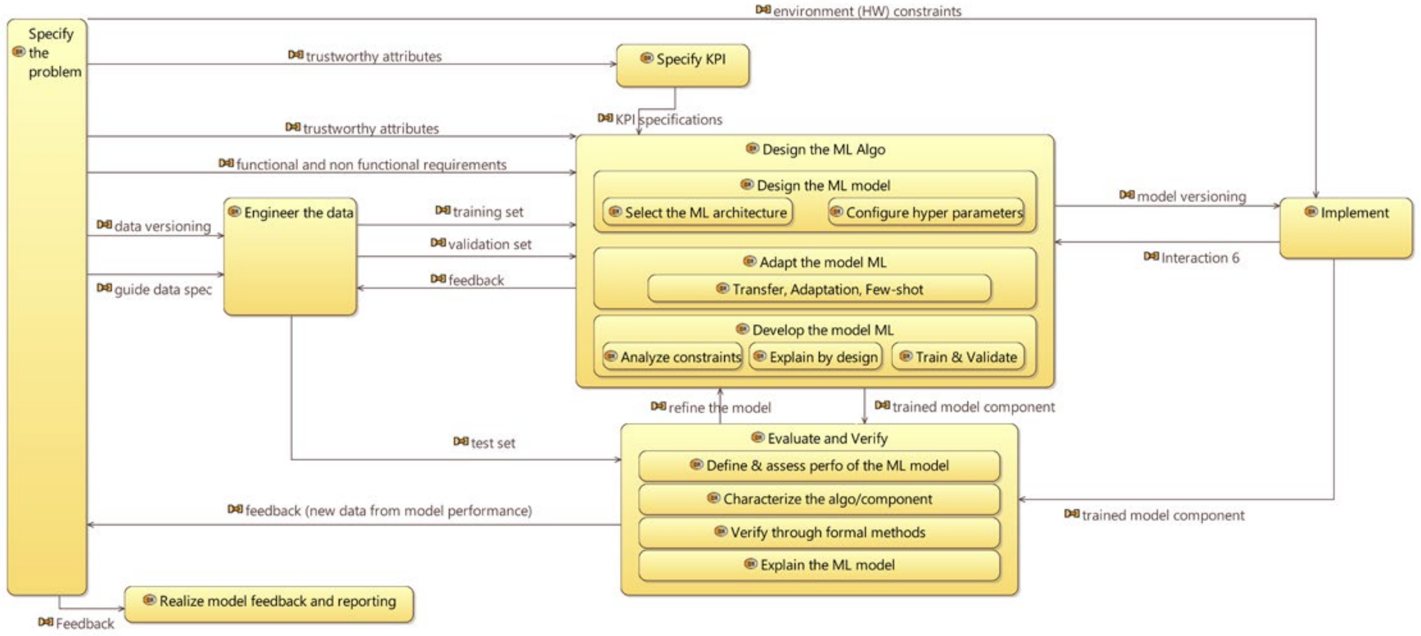

Figure 2: Illustration of a Capella model for ML Algo Engineering

As of today, the meta-model capturing and organizing the concepts supporting the different viewpoints has been developed, its implementation in the Capella environment is ready, and the work to populate the engineering model using the methods and tools developed or recommended by the different projects of the program is on-going. Next step will be to build the complete verification and validation (V&V) argumentation with clear links between the workflow activities, the engineering items, evidences, and activities providing evidences, and to bring together related methodology elements and adequate tooling to support collaboration of all engineering disciplines for trustworthy AI-based products over their life cycle.

2.2 The Confiance.ai environment

The trustworthy environment can be seen as the implementation of an end-to-end engineering method which combines:

- A top-down approach that tries to determine how “classical” Systems Engineering methods should be modified to consider the specificities of AI.
- A bottom-up approach that takes the elementary methodological bricks which can evaluate or enhance trustworthy parameters for an AI based system, makes sure that they are consistent, and assembles them.

It has been thought and built with the subject matter use cases as the main driver. The main purpose of the trustworthy environment is indeed to be able to evaluate, characterize and enhance the relevant trustworthy parameters for an AI based use case and hence be able to deploy it in a real-life industrial environment.The trustworthy environment includes the following main components:

- The companion is a methodology driven entry point into the environment: for a particular use case and its Operational Design Domain (ODD), it will help select the relevant trustworthy parameters for this particular use case.
- A set of technological bricks referred to as “components” that will enable the system developers to focus on specific trustworthy elements. There are three types of these components:
  - Python Libraries that the developer will directly include and use in their software.
  - Docker images for standalone components that can apply globally on the software being developed. These dockers images can be actioned from the MLOPs pipeline in a very simple manner.
  - Complete applications with a user interface that will help the developer to build the model. An example of these would be an intelligent labelling application that will help strengthen the dataset used to produce the model.
- An MLOps industrial chain which can be seen as the engine inside the environment. MLOps is the adaptation to AI of the Devops industrial software development techniques. This chain will be able to handle things like versioning of models or data sets as well orchestration pipelines that enable the evaluation of the different trustworthy parameters and the production deployment for a particular use case.
- The dashboard will give a visual output for the trustworthy parameters as they have been measured and computed for a given use case thru the different components of the platform. With the use case as the main entry point, it gives an end-to-end vision to assess the trustworthiness of an AI based solution aimed to be deployed in a critical environment.

To avoid the “Yet another AI platform” syndrome, the trustworthy environment has been designed so that it can easily be integrated and bring value in the environments that exist in industry today. Its key foundations are well established mainstream opensource projects, among them Kubernetes for the runtime environment and Apache Airflow for the pipeline orchestration play a major role.
This smart independent architecture enables the environment to be deployed in most of current IS environment whether one is using a cloud hyperscalers like AWS, GCP or Azure, or one relies mostly on on-premises infrastructures.
This architecture also allows the user to have a very specific customized approach where they would pick and choose the individual components that are of interests for their project and will insert them in a Lego like approach into their own preexisting MLOps environment.
To illustrate the best usage of the trustworthy environment, let’s take the example of a factory line in which one wants to introduce an image recognition system to identify defects in the products built by the line:

- The specifier of the systems will use the companion in the trustworthy environment to help them precisely define the ODD and pick up the trustworthy parameters that need to be considered for this application.
- The implementer of the solution will use different bricks in the trustworthy environment to help them make sure that the solution will meet the targeted parameters. In that case, they might for example use a synthetic data generation application to complete a dataset for poor light conditions and use a robustness brick to make the model capable of working for blurred picture.

The quality officer of the factory will use the dashboard to review how the model behaves and meets the targeted trustworthy parameters to audit the system and give it the go to be actually deployed on the production line.

3. Components of trustworthy AI and some tools

3.1 Introduction

The Confiance.ai program aims to cover a wide range of aspects (explainability, privacy, robustness, etc.,) that together increase society’s trust in AI-based systems.
The initial developments of the program were devoted to data-based AI systems, namely machine learning (ML) and in particular neural networks. Therefore, this section mainly addresses this kind of AI.
In practice, the ML model is only one part of the system and significant additional functionalities are required to ensure that the global system operates reliably and predictably with appropriate engineering of data workflow, monitoring and logging, etc. To capture these aspects of AI engineering, we have defined the engineering pipeline of the ML algorithm.

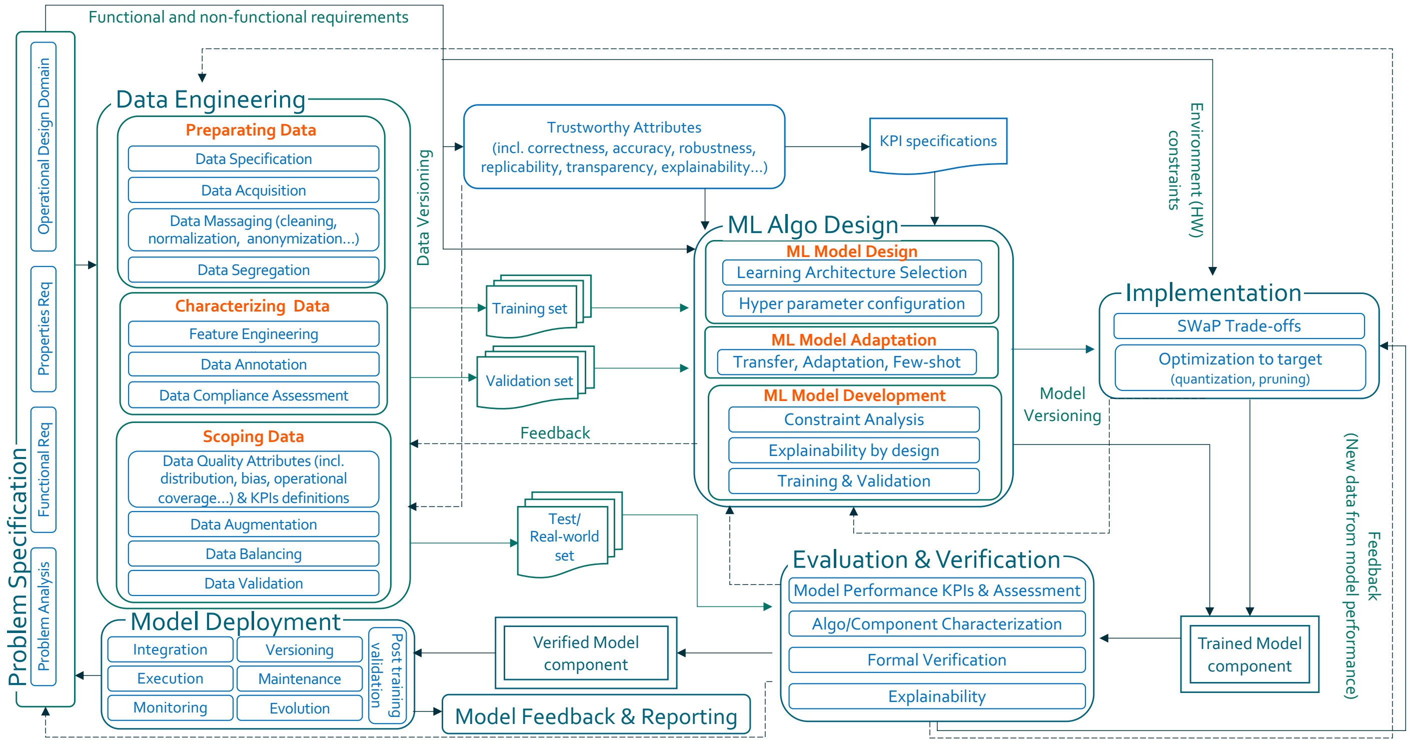

Figure 3: Machine Learning algorithm engineering pipeline

The different phases of this pipeline are:

Problem specification including the Operational Design Domain (ODD) that is the description of the specific operating condition(s) in which a safety-critical function or system is designed to properly operate, including (but not limited to) environmental conditions and other domain constraints[4].

Data engineering including data collection, preparation and data segregation. The collected data needs to be sizable, accessible, understandable, reliable, and usable. Data preparation, or data preprocessing transforms raw data into usable information.

ML algorithm design: After feeding training set to the ML algorithm, it can learn its appropriate parameters and features. Once training is complete, the model will be refined using the validation dataset. This may involve a selection of variables and includes a process of setting the model hyperparameters until an acceptable level of accuracy is achieved. **Implementation** to develop ML component, to decide on the targeted hardware platform, the IDE (Integrated Development Environment) and the language for development for which we have to consider embedded constraints.

Evaluation and verification: Finally, after finding an acceptable set of hyperparameters and optimizing the model accuracy, we can test our model. The test uses our test dataset and aims to verify that our models have accurate characteristics. Depending on the result of this verification, the model design can be reworked to improve accuracy, adjust its parameters, or instead deploy the model.

Implement, document, deploy and maintain: The final step is to implement, document, deploy and maintain the ML based system so that the stakeholder can continue to leverage and improve upon its models. This leads to consider embedded constraints for target hardware, online models monitoring to detect environmental change. This phase has strong interactions with the system activities described in part 2.

In this chapter we will present several components and tools proposed by confiance.ai in order to implement this pipeline with data specification, data workflow and active learning, robustness improvement, robustness evaluation, explainability and uncertainty evaluation and monitoring.

3.2 Data specification

Data is crucial for data driven artificial intelligence systems. Such AI models must be trained on large annotated datasets to learn how to perform a task without being explicitly programmed to do so. But the success of a data driven AI system will obviously depend heavily on the data and its relevance to the use case domain, regardless of the amount of available data or the network design. An incorrect or unbalanced dataset will negatively impact the AI component predictions (bias in predictions, missed requirements, lack of generalization in real applications etc.).

Let's take the example of an engineer designing a visual control system to detect anomalies in an industrial product. At first glance, this may be a simple classification problem, and the engineer may focus on investigating state of the art on ImageNet. But at some point, large amounts of data related to its industrial use cases will still need to be acquired. The question will be then what data should be acquired and used for training and testing. However, it will not be sufficient to setup a camera and record plenty of data. A lot of questions will then arise, for instance:

- **Which quantity of data is sufficient to train a model and reach a targeted performance?**
- **Which data should we acquire to represent the anomalies? Should we have different class of anomalies?**

…

- **Which additional data, unconnected from the data used for model design, must be acquired to ensure traceability of requirements and performance evaluation?**

These questions are common in the development of data driven AI components for the industry but to date they are still addressed based on empirical experiences of AI engineers. There is thus a strong need for methodology related to the data building process and in particular its integration into a system engineering point of view. Among other things, the concept of Operational Design Domain (ODD) originally created for autonomous driving[5], defining the operating conditions of a system, can be used to identify required parameters of a dataset, such as environmental conditions or object properties (size, color...). The "Data Specification" should take into account the "Functional and non-functional requirements" and produce output requirements for the "Data Acquisition" and the "Data Annotation" tasks. Its goal is to ensure that the output requirements cover and guarantee traceability of input requirements.

[4] Koopman and Fratrik, 2019

[5] Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems", SAE J3016, 2018

We propose the following scheme for "Data Specification" activity:

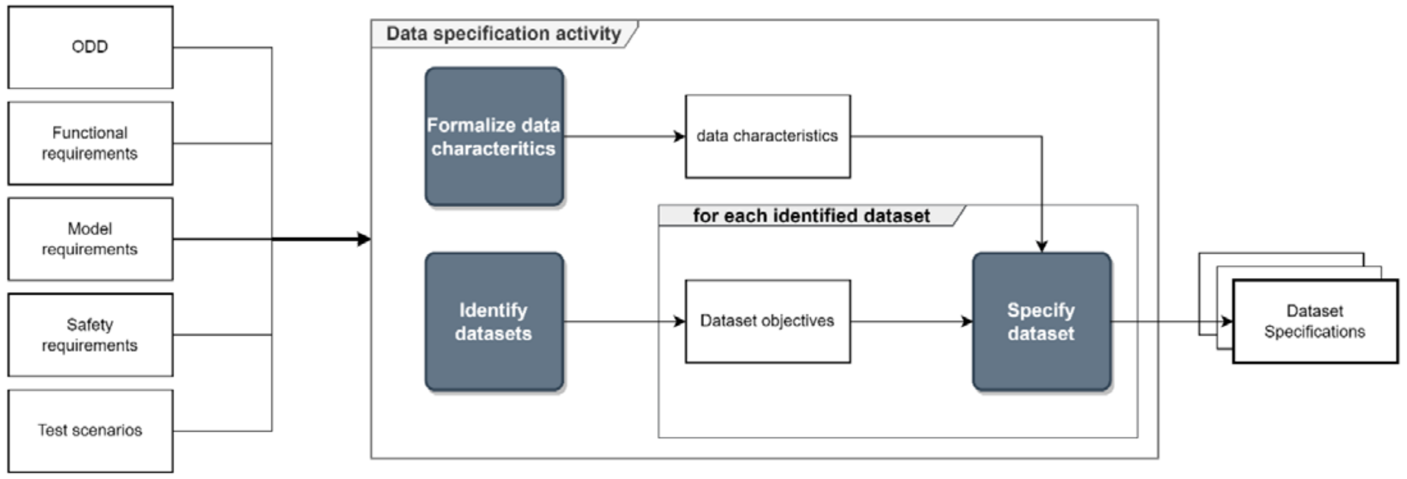

Figure 4: "Data Specification" activity

The challenges to move from a formalized ODD (a list of data characteristics, associated to a range of values) and performance targets to a data and dataset specification are mainly:
Define how to discretize data characteristics having continuous value,
Define how to combine all the data characteristics.

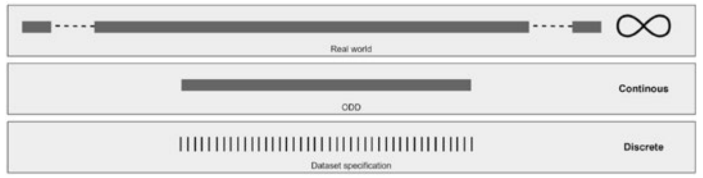

Figure 5: From Infinite Real World to Discretized ODD

One should not forget the difficulty of acquiring data in the real world and obtaining a properly discretized and balanced dataset. Indeed, the real world is often incompletely observable and this will be reflected into the datasets. Then, comes the difficulty to combine these characteristics to define the complete dataset specification. Using a naive approach of regular and uniform sampling of each data feature is not relevant for "high-dimensional" problems, i.e., problems having an ODD defined by a high number of features. Then, the dataset specification activity has to provide a way to define a combination of all the data characteristics which is compatible with coverage expectations and a realistic amount of data acquisition.
Design of a ML component might be an iterative process. Input requirements such as Operational Design Domain can be led to evolve for example to ensure "data currency" (the preservation of representativeness of data over time). Consequently, one must keep in mind that "Data Specification" is an iterative activity with a refinement procedure as the design activity progresses while feeding into all those that are taking place simultaneously. With the "Data Specification" activity a list of properties is identified (representativeness, traceability, accuracy, reliability, consistency, integrity, bias detection) and comes with associated recommendations that will be used to refer in the construction of "Data Specification" items: datasets, annotation and acquisition.

3.3 Data workflow and incremental learning

Practical AI projects follow a cyclical process rather than sequential development process, with continuous iterations, tunings and improvements. This is particularly true for data, since getting the right data at the start of a project is complicated and rarely achieved in real life development. Since data capture is long and expensive, AI components are usually built in parallel with data construction, thus in an iterative manner. Similarly, new data acquisition phases are often necessary to either improve AI model performance or to fulfill use case requirements. Rather than blindly acquiring more data, it is more efficient to select cleaner and more relevant data. We give an overview of methods and tools helpful to efficiently and iteratively refine a dataset, and thus an AI component:

- Analyzing AI model behavior with regards to ODD coverage is useful to identify dataset weaknesses (e.g., under-represented or missing data). Methods based on model error on ODD coverage can be valuable to guide dataset construction iteration (ex: acquisition guidelines, online data mining).

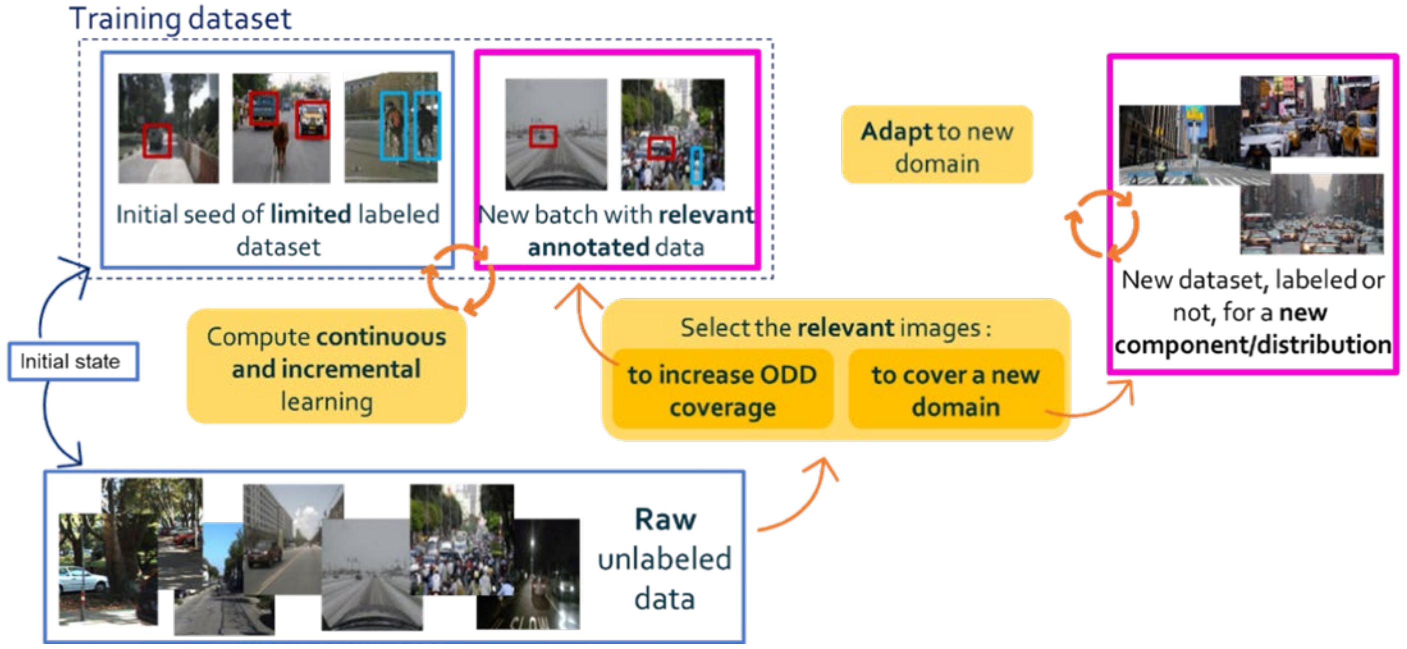
Training dataset
Initial seed of limited labeled dataset
New batch with relevant annotated data
Adapt to new domain
New dataset, labeled or not, for a new component/distribution
Initial state
Compute continuous and incremental learning
Select the relevant images :
to increase ODD coverage
to cover a new domain
Raw unlabeled data

Figure 6: Incremental dataset construction

- Training an AI model usually takes several days, while data releases can be quite regular (for ex weekly basis) in an industrial project. Updating the AI component for each data release can quickly become a repetitive and tiresome task, especially if the model is always trained from scratch. Incremental or continual learning methods, like data weighting or distributed learning, can be quite practical.
- Notably, real-world datasets are often composed of a large amount of unlabeled data that can be both acquired and modified while the system is still under development. Labelling data is expensive and time consuming, and therefore it is not a practical way to deploy an industrial system.
- Some unsupervised machine learning paradigms, namely self-supervised, and semi-supervised learning, offer a means of making use of this unlabeled data.
- In the figure bellow, we introduce a data workflow that incrementally builds the required dataset in order to increase ODD coverage and concomitantly minimize manual annotation.

3.4 Building more robust models

In this section we focus on some neural network design methods that improve the robustness of models. More specifically, the following methods will be considered:

- **Uncertainty Quantification by Design** aim at designing a deep model endowed with the ability to compute the predictive uncertainty in a supervised learning task. The total uncertainty is decomposed into the noise coming from the data (aleatoric) and the error coming from an insufficient knowledge of the underlying physical process (epistemic). Based on recent advances on both deep generative models and Bayesian deep learning, we can combine regression prediction and uncertainty quantification on an industrial use case.
- **Adversarial training methodology** aims at providing a broad practical implementation of adversarial training applied in a nearly realistic industrial context that uses machine learning to automate crucial tasks. The goal is to provide different adversarial training methods, to highlight the key methods and challenges related to the adversarial training methods and to show how to apply it to industrial use cases.
- **Randomized smoothing** method aims to build robust models for classification and regression. It also enables to empirically certify the resulting model's robustness. For classification tasks, randomized smoothing provides a certified accuracy within a certain radius where disturbances are limited. For regression tasks, it provides an interval in which the prediction is guaranteed to be settled in.
- **Built-in 1-Lipschitz neural networks**, are neural networks, where constraints are imposed during training to coerce their Lipschitz constant. Controlling the Lipschitz constant of a network has a great im-pact on adversarial robustness. Given the same outputs (logits of the neural net-work), the lower the Lipschitz constant is, the larger "attack strength" is required in order to build an adversarial example.

3.5 Local robustness evaluation

The study of reliability and robustness of a neural network consists in verifying its ability to make the same decision for all similar input data, despite the attacks they may be subject to. Therefore, a Neural Network Verification (NNV) approach is needed to provide robustness indicators for neural networks. The principle of NNV is to find, from the input data, all possible data resulting from an attack (noisy data), and check that the properties of the neural network remain valid in all cases. In this chapter we focus on method dedicated to local robustness on the evaluation (empirical) or demonstration (formal). Each of the method or tools have been tested on different use cases:

Empirical local characterization:

- **Non-overlapping corruption benchmarking tool**, provides a benchmark of synthetic corruptions on a dataset to assess the robustness of a given AI-based model.
- **AI Metamorphism Observing Software**, assesses metamorphic properties on AI models such as robustness to perturbations on the inputs but also relation between models' inputs and outputs.
- **Time-series robustness characterization,** focuses and the assessment of the robustness w.r.t. perturbations on the inputs of regression models applied to time series.
- **Adversarial attack characterization**, evaluates the impact and usability of adversarial attacks on AI models.
- **Amplification methods for robustness**, evaluates the robustness of models using amplification methods on the dataset with noise functions.

Formal methods for local characterization:

Regarding formal methods several tools are available for NN verification sur as abstract in-terpretation, SMT (satisfiability modulo theories), MILP (Mixed-integer linear program-ming) with bounds, Symbolic Intervals.

3.6 Explainability benchmark

Methods dedicated to interpretability and explainability highlighting features or feature pattern that influences a neural network's prediction can be divided in two type, white-box methods (based on computation of gradients and back-propagation), black-box explanation methods (that are based on perturbation of one or some features). Associated to these methods, we identify metrics and tools used to generate benchmarks.

White-box methods

- **DeconvNet** aims to reconstruct the input of each layer from its output, showing what input pattern caused a given activation in the feature maps. **Guided Backpropagation** is a variant to the Deconned method. Saliency method generalizes the DeconvNet method.
- Gradient x Input computes the element-wise product between the saliency maps and the input to improve the sharpness of the attribution maps.
- Integrated gradients combine the implementation of gradients along with the sensitivity (e.g., of the technique Layer-wise Relevance Propagation). Due to its implementation, the method results are better than the above methods, in most of the cases.
- Smoothgrad method consists in adding some Gaussian noise and averaging saliency maps. Adding some noise to the samples reduces the noise to the explanation. With variant SquareGrad and VarGrad which differ from SmoothGrad is the way of averaging the noisy samples.
- **Grad-CAM** is a generalization of Class Activation Maps (CAM), a technique for identifying discriminative regions, by using the gradient information flowing into a model layer. This method increases the accuracy and the complexity w.r.t. the CAM approach. Grad-CAM++ is a generalization of Grad-CAM and increases the precision and complexity w.r.t.

Black-box methods

- **LIME** method interprets individual model predictions (of any classifier or regressor) in a faithful way, by approximating it locally with an interpretable model around a given prediction.
- **KernelSHAP** method is a combination of linear LIME method and Shapley values (feature importance for linear models). This method increases the accuracy and decreases the simulation time w.r.t. Lime and SHAP.
- **Occlusion** The occlusion sensitivity is a method for understanding which features are the most important for a deep network's classification. When occluding an important feature, a strong drop-in activity can be seen in the feature map.
- **RISE** method generates an importance map indicating how salient each feature is for the model's prediction. This method estimates importance empirically by probing the model with randomly masked versions of the input image and obtaining the corresponding outputs.

Several explanation metrics have been used

- **MuFidelity** metric represents the faithfulness for explainers. This metric corresponds to the Pearson's correlation coefficient between the predicted logits of each modified test point and the average explanation for only the subset of features. Faithfulness increases with the increasing of subset size.
- **Deletion** metric is a fidelity metric, it measures the decrease of the prediction score when removing progressively the most important features, which are given by the explainer.
- **Insertion** is a fidelity metric, it measures the increase in the prediction score when adding progressively the most important features, which are given by the explainer.
- **Average Stability** metric ensure that close inputs with similar predictions yields similar explanations.
- **MeGe** (Mean Generalizability) is a metric for representativity of explanations. MeGe gives an overview of the generalization of your explanations: the explanations provided by datasets, and the others.
- **ReCo** (Relative Consistency) is a metric for consistency of explanations. ReCo is the consistency score of the explanations by ensuring that contradictory predictions lead to different explanations.

Several toolboxes were evaluated in order to provide benchmark on

- **GemsAI :** operates on a full validation set and associated observations to provide insight, not on single observations (like local gradient-based methods) and does not require access to the model's internal engine / layers
- **Xplique** library is composed of several modules for multiple Attributions Methods module, Metrics, and Feature Visualization, and the Concepts module (concept extraction from a model).

Four bullets are used…

… for efficiency:	…for usability:
🟢 The method gives good results	🟢 The method is easy to set up
🟠 The method gives mixed results	🟠 The method is relatively easy to set up
🔴 The method gives bad results	🔴 The method is difficult to set up
⚫ The method is not suitable	⚫ The method is hard to set up

Use case	Library	Method	Eff.	Usa.	Comment
Welding Inspection (Renault)	Xplique (ANITI)	**Lime**	🔴	🔴	Blackbox
		KernelShap	🟠	🔴	Blackbox
		Occlusion	🟢	🟠	Blackbox
		Rise	🟢	🟠	Blackbox
	Gems.AI (ANITI)	**Image Classification**	⚫	//	Blackbox
	Shap (Microsoft)	**PartitionShap**	🟠	🟠	Blackbox
	AIX 360 (IBM)	**Contrastive Explanations Method**	//	⚫	Blackbox

Table 2: Extract of the benchmark results

- **AIX360** library contains many tools that are applicable to different places in an AI project. These tools are applicable to both tabular data, images and texts. Some tools are adapted to regression models and others to classification models.
- **SHAP** is based on an API provided by Microsoft named SHAP. Shapley values are a widely used approach from cooperative game theory that come with desirable properties.

We evaluate on confiance.ai use cases these tools, you can find bellow a summarize of the benchmark conclusions of the benchmark. For each set use case vs method, the efficiency (Eff. column) and usability (Usa. column) are evaluated and give a short conclusion.

3.7 Multi-timescale Online Monitoring

Objective and scope

The main objective of the online monitoring of AI models is to detect any deviation of the AI component deployed in operation from the specified expected behavior or from a predefined set of safety operational properties. A product has been developed using AI technologies, and it should demonstrate that the AI model can perform its prediction over its entire Operation Design Domain (ODD) with an accuracy of 99.9 % and that this accuracy is maintained over time in operation. Let's assume that after a full training phase, the model's performance does not exceed 99 % of correct predictions, it implies that 10 failures may statistically occur over the reference period (1000 hours) when only one failure would have been tolerated. This situation is unacceptable from a product safety point of view. The deployment of a monitoring component operating in parallel with the AI model (online monitoring as depicted in Figure 1) is a concrete way of managing this type of residual risk induced by a model for which it is not possible or feasible to formally demonstrate the achievement of the performance/safety objectives resulting from the system analyses. Online monitoring is a safety architectural pattern that is well known to operational safety engineers, but it had to be adapted to AI technologies.

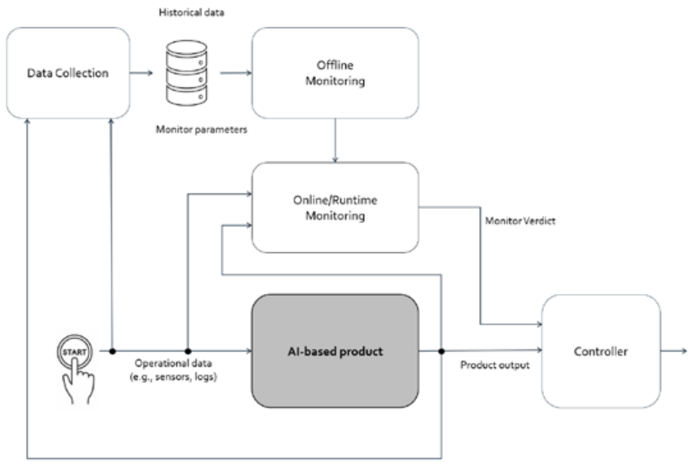

Figure 7: Monitoring Architectural Pattern

The work performed in confiance.ai defines an innovative engineering method to develop and verify online monitoring components that combine several monitoring timescales: a monitoring of the AI-based product at present time, a monitoring on a configurable time window of the near past, a monitoring on a configurable time window of the near future. These three (3) monitoring scales complement each other to ensure a high rate of detection of failures that could occur in operational conditions when the AI model is in production.

Monitoring functions

Monitoring relies on monitoring function that will detect inconsistence in the inputs and/or product output such as, in the case of images:

- **Standard Defocus (Out of focus) Blur Detection:** In optics, defocus is the aberration in which an image is simply out of focus.
- **Standard Motion Blur Detection:** Motion blur is the apparent streaking of moving objects in a photograph.
- **Standard Brightness Detection:** This function aims at extracting the degree of brightness from an image and raising an alarm if this degree of brightness can impact the prediction of the model on this image.
- …

For the evaluation of the monitor function, we need to assess the ability of the black box monitoring function to detect Out of Distribution images. These images correspond to camera problems encountered by the use case provider.

3.8 Combination of methods as a conclusion

All the method and tools briefly introduced in this section allow us to improve the trust properties of our system including AI component, but there is no silver bullet, we need to apply most of them and make trade-offs during the design in order to adapt the ODD, improve our model, improve the dataset, …

Back to the product discussed in 3.7, let's consider we have a model with the following properties:

- Overall performance of 99% on the whole dataset
- Performance increased to 99.5% when the degree of brightness is above a threshold
- The local robustness evaluation indicate that model is not affected by a "small" noise
- Explainability tools allow us to identify which pixels contribute to decision

What can we do?

As long as a model is not sensible to small noise, a monitoring function shall be added to exclude image with "more" noise. As our performance target is 99.9%, investigation with explainability tools the 0.5% failure might enable for example identification of numerous samples with invalid color mapping. We have the choice of improving the model sensibility to color mapping, or adding a new monitoring function and limit the ODD concerning camera color mapping….

And now?

We do not include here all the methods and tools integrated in confiance.ai environment. All the presented tools and method works only to on subset of data and AI model type AI type. There is still scientific roadblock (dataset ODD representativity demonstration, …) for which few solutions exist when models and or data dimensionality increase.

3.9 Embedded AI components

One objective for Confiance.ai is to provide guidelines, methodology and tools to support the implementation of AI Components on an embedded hardware, which has limited resources. Most of the time, the implementation takes as input data or specification a model of the AI component (usually a neural network) that has to be implemented on a specific hardware target. The source code corresponding to this implementation is most often specific to the target. During this implementation phase, the main challenge is to preserve the trust properties described previously.

This implementation step is most often optional for AI components deployed at the cloud level. Indeed, frameworks such as pytorch, which allow designing and generating models of AI components, provide all the components necessary for the automatic deployment of these models.

Thus, from a practical point of view, the main activity related to embedded systems consists to move from a model running in a cloud environment to a model running in an embedded environment with therefore additional constraints related to embedded systems.

This transition from the world of the cloud to that of the embedded systems relies on the following topics:

- Methodology and guidelines: to define the workflows and processes to take into account the constraints related to embedded systems
- Software to transform the AI model to an implementation compliant to the target hardware (compilation…)
- Tools to deploy the model. In a similar way to the cloud, the notion of pipeline can be applied for example for a chain of AI components including both algorithms and components related to trust (monitoring, explainability…)

The main challenges associated with this topic are the following:

● Resources estimation

The objective of resource estimation is to characterize the resources needed to run an AI component before its implementation. This phase upstream of the development cycle makes it possible to add constraints upstream of the design of an algorithm (for example the number of quantization bits) and can also support a hardware sizing phase.

● Interface/ Interoperability and Semantic preservation

The description of the ML Model is defined as the interface between the design and the implementation processes. For safety-related component, one challenge of the AI embedded systems is to demonstrate that the implementation process does not alter the safety/functional/operational properties of the ML model obtained by the design process.

To this purpose, a ML model has to be defined with the following requirements:

- Explicitly and fully described, with no possible interpretation,
- Exactly replicable on a software/hardware target, with no possible approximation,
- In order to:
  - Ensure the preservation of the semantics
  - Be fully verifiable (for conformity to regulation requirements)

● **Compilation toolchain and benchmarking environment:**

In the domain of Deep Learning, compilers are designed to optimize the execution time of the inference phase, they don't take into account trust properties such as repeatability or WCET (Worst Case Execution Time). The challenge for Confiance.AI is to adapt the compilers in order to give better guarantees related to the trust properties expressed by the trustworthy components.

Indeed, in the frame of embedded real time system, a service shall generally be delivered in some bounded time, i.e., before some deadline.

● **Model optimization**

AI components corresponding to neural networks generally require a huge number of MAC operations, that are not compliant with the available resources of the Hardware target. So, compression and quantization operations are necessary to fit the constraints of embedded hardware (limited resources) and with trust guarantees (ie : error control of accuracy).

● **Certification and safety**

The current certification standards for the implementation of safety-critical systems have to be adapted to fill the gaps and provide a new certifiable approach to tackle the new requirements related to AI components (Machine Learning components).

Another challenge is to analyse the impact of hardware faults on neural networks and to propose new, low-cost approaches for attenuating their effect.

4. Assessing Trustworthiness

4.1 Rationale

The design of AI-based safety-critical systems[6] (and more generally critical systems) such as in avionics, mobility, defense, and healthcare, requires proving their trustworthiness. Trustworthiness assessment begins from the early stages of development, including the definition of the specification requirements for the system, the analysis, the design, etc. Trustworthiness assessment must generally be considered at every phase of the system lifecycle, including sale and deployment, updates, maintenance or disposal.

Due to the multi-dimensional nature of trustworthiness, the main issue we faced is to establish objective trustworthiness attributes that are clearly identified and mapped onto the AI processes and its lifecycle. Some are identified, such as accountability, accuracy, availability, controllability, correctness, integrity, privacy, quality, reliability, resilience, robustness, safety, security, transparency, usability. Thus, the notion of trustworthiness shines a light on quality requirements ("-ilities", or non-functional requirements) which appear particularly challenging in an AI system, although many of them can be considered in any system. Furthermore, beyond quality requirements, the notion of trustworthiness can also encompass risk and process considerations. The expected attributes and the expected values for these attributes depend on contextual elements such as the level of safety criticality of the application, the application domain of the AI-based system, the expected use, the nature of the stakeholders involved, etc. This means that in some contexts, certain attributes will prevail, and other attributes may be added to the list. In addition to that, one can easily understand from the list of attributes presented above, that they do not have the same level of granularity: some may be directly assessable through a single quantitative or qualitative metric (or a set of metrics), while in most cases the attribute requires decomposition into several sub-concepts, each requiring their own metrics. Following the decompositions, trustworthiness is thus composed of a list of so-called "atomic attributes", meaning that they can be directly linked to metrics. Finally, trustworthiness is composed of attributes whose levels are not directly comparable to each other: These attributes can have different scales and dimensions, and their comparison is in any case context dependent. Their aggregation to a single trust score must be based on a consistent and realistic methodological approach.

[6] We call safety-critical system, any system where failures can lead to catastrophic consequences in term of victims or material damage.

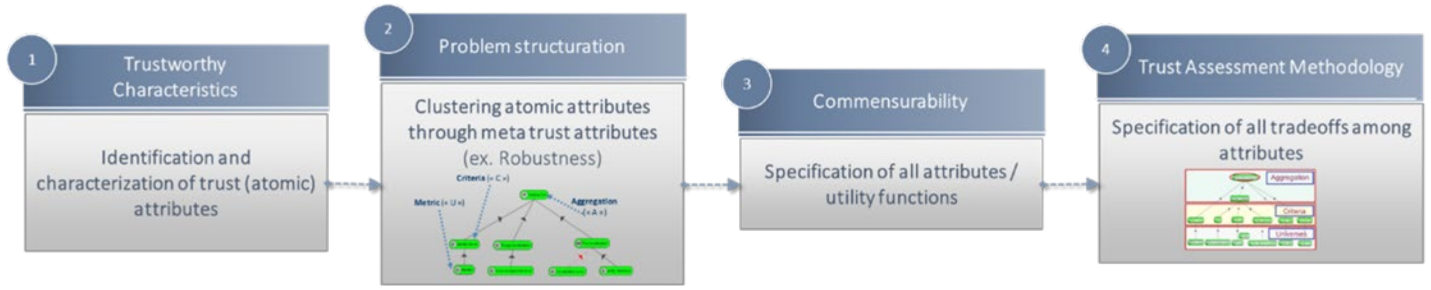
1
Trustworthy Characteristics
Identification and characterization of trust (atomic) attributes
2
Problem structuration
Clustering atomic attributes through meta trust attributes (ex. Robustness)
Criteria (« C »)
Metric (« U »)
Aggregation (« A »)
3
Commensurability
Specification of all attributes / utility functions
4
Trust Assessment Methodology
Specification of all tradeoffs among attributes
Aggregation
Criteria
Universes

Figure 8: Main steps of the MCDA approach

4.2 Unified approach on trustworthiness attributes based on Multi-Criteria Decision Aiding

Multi-Criteria Decision Aiding (MCDA) is a generic term for a collection of systematic approaches developed specifically to help one or several Decision Makers (DM) to assess or compare some alternatives on the basis of several criteria.

The difficulty of this problem is that the decision criteria are usually numerous and conflicting. One may indeed have performance criteria versus cost criteria which cannot be met both at the same time. The viewpoints are quantified through attributes.

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

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

Thus, our approach is based on the following steps:

Step 1: Definition of the different attributes that constitute trust[7].

Step 2: Structuring of the attributes in a semantic tree allowing a first hierarchy ….

Step 3: Identification of metrics, assessment methods or control points for each atomic attribute;

Step 4: Definition of an aggregation methodology to capture operational trade-offs and evaluate higher-level attributes.

4.3 Trust Characteristics

Based on different sources (norms, standards, scientific communications, industrial and institutional reports, Confiance.ai reports...), the characterization and evaluation of trust attributes focus on the definition, structuring and metrics of the attributes that constitute trust in the context of AI-based critical systems.

These attributes are currently grouped according to the capabilities they characterize: technical, ethical, interaction and trust intermediaries (such as certification). "Technical" is system-centric, it refers to the ability to verify that the AI-based component has valid and robust intrinsic properties. This includes attributes such as reliability, dependability, accuracy, reproducibility and maintainability. "Ethical", strongly linked in Europe to the notion of fundamental rights, is notably put forward in the work of the AI HLEG (High-Level Expert Group on Artificial Intelligence) of the European Commission. A system must, for example, offer a guarantee of fairness or privacy. The interaction with stakeholders (end-users, designers and auditors) is linked to notions such as transparency, explicability or usability. Finally, the attributes for trusted ecosystem intermediaries focus on the relationships to third-parties, in particular quality assurance, audit and certification activities. All these proprieties apply to the AI-based component, but they also apply to the quality of the data used for training connectionist AI and/or to the quality of knowledge modeling used in symbolic AI.

These four groups of attributes form, along with trustworthiness governance, different perspectives to trustworthiness. An overall description of these perspectives is summarized in the following table and captured in the Mind-Map presented below.

[7] The program Confiance.ai has already identified ~120 trustworthiness atomic attributes.

Perspective	Principles
Trustworthiness governance	Trustworthiness, Trust
Intermediaries in trust	Certification, Labels, Homologation...
Technical	Accuracy, Efficiency, Effectiveness, Functional Suitability, Security, Data Quality, Reliability, Robustness Freedom from risk, Verifiability, Portability, Maintainability, Process quality...
Interaction	Usability, Explainability...
Ethical	Benevolence, Social well-being, Environmental well-being, Diversity, Fairness, Nondiscrimination, Respect for the rule of law, Privacy, Bias reduction, Respect for fundamental rights, Integrity

Table 3: Some AI trustworthiness attributes

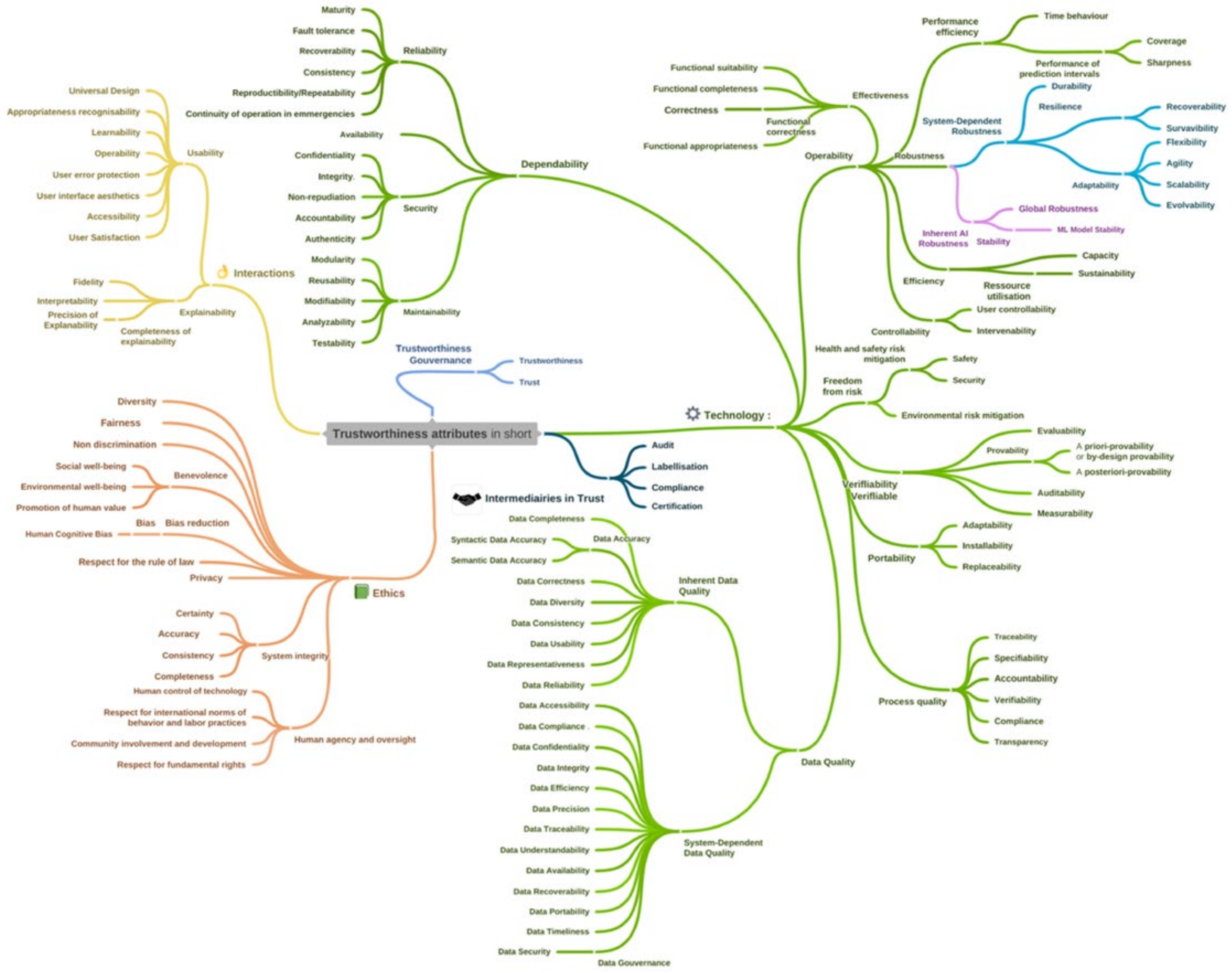

Figure 9: The AI-based system trustworthiness attribute mindmap

From a technical perspective, as the program confiance.ai aims to develop methods and tools to support critical AI-based systems during their life cycle (from specification to in service support through design and deployment), we have to demonstrate the properties of accuracy, robustness, safety and security. Thus, AI-based systems should generate accurate output consistent with the ground truth as much as possible. This is the most basic motivation for defining trustworthiness attributes. Additionally, AI systems should be robust to changes, even if real environments are usually very complex, dynamic and uncertain. Moreover, no AI programs or systems are expected to harm any people under any conditions, which put the safety of users as the first priority. In addition, the autonomy of trustworthy AI should always be under user's control. In other words, it is always a human right to grant an AI system any decision-making power or to withdraw that power at any time.

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

From ethical perspective, trustworthy AI should be law-abiding, fair, accountable, environmentally friendly and compliant with the user privacy. Specifically, AI systems should operate in full compliance with all relevant laws and regulations and comply with the ethical principles of human society.

4.4 Trust Scoring

While the necessity and usefulness of reasoning about trust assessment is obvious, obtaining trustworthiness scores remains a challenging task. As stated previously, some aspects linked to trustworthiness are highly subjective or context dependent. For example, the notion of "data quality" (resp. "robustness") require having a knowledge of all induced attributes including those that are system dependent such as data availability, data portability, data precision... (resp. adaptability, durability, resilience...). The subjectivity or vagueness of the attribute definitions does not always represent a major hindrance to use them in operational settings, because skills and knowledge of AI and safety engineers may be enough to determine what may be appropriate thresholds and scores. There are 4 types of assessment: Nominal - the variable can only be categorized; Ordinal - the variable can be categorized and ranked; Interval - the variable can be categorized, ranked, and evenly spaced; Ratio - the variable can be categorized, ranked, evenly spaced, and has a natural zero.

At the **nominal** level, the numbers represent mutually exclusive categories (e.g. availability of a quality process: yes=1, no=0). Attributes that can be measured on a nominal scale have the following properties: they have no natural order and categories are mutually exclusive. At the **ordinal** level, the numbers only indicate order (i.e., class rank). For example, the usability attribute "appropriateness recognisability" is the degree to which users can recognize whether a product or system is appropriate for their needs. This attribute can be assessed on an ordinal scale: "Very inappropriate", "inappropriate", "neutral", "appropriate", "very appropriate". Ordinal variables have no exact difference between variables – we do not know if the difference between "very satisfied" and "satisfied" is the same as the difference between "satisfied" and "neutral". For the **interval** level, the distances between numbers have meaning. You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point. Attributes on an interval scale have the following property: they have a natural order. We can compute their mean, median, mode, and standard deviation. They have an exact difference between values. The last type is a ratio scale used to label attributes that have a natural order, a quantifiable difference between values, and a "true zero" value. Examples of ratio scales usually include length, weight, duration, and more.

Assessment Type	Mathematical operations	Assessment of central tendency	Assessment of variability
Nominal	Equality (=, ≠)	Mode	None
Ordinal	Equality (=, ≠) Comparison (>, <)	Mode Median	Range Interquartile range
Interval	Equality (=, ≠) Comparison (>, <) Addition, subtraction (+,–)	Mode Median Arithmetic mean	Range Interquartile range Standard deviation Variance
Ratio	Equality (=, ≠) Comparison (>, <) Addition, subtraction (+,–) Multiplication, division (×, ÷)	Mode Median Arithmetic mean Geometric mean	Range Interquartile range Standard deviation Variance

Table 4: Mathematical operations for trustworthiness assessment

4.5 Focus on Data Quality

Data quality is a problem that has been studied for several decades now. However, primarily the focus has been on the data in operational databases and data warehouses. Data-driven AI is generating renewed interest, especially to characterize the training and validation data sets in machine learning. The ISO 25012 standard on general data quality distinguishes "Inherent Data Quality" and "System-Dependent Data Quality" (a point of view which will be completed by the ISO 5259-3 standard on data quality for machine learning):

Data Quality	Trustworthy attributes
Inherent Data Quality	Completeness, Accuracy, Correctness, Diversity, Consistency, Usability, Representativeness, Reliability…
System-Dependent Data Quality	Accessibility, Confidentiality, Integrity, Traceability, Portability, Timeliness…

Table 5: Some data quality attributes

As an example, we could consider the following score for the assessment of the timeliness attribute:

$$\mathit{Timeliness} := \max\left(1\text{-}\ \frac{\text{age of the data value}}{\mathit{shelf\ life}},0\right)^{s}$$

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

$$\mathit{Correctness} := \frac{1}{1 + d(\omega, \omega_m)}$$

where ω is the data value to be assessed, ω_m is the corresponding real-world value and d is a domain-specific distance measure such as the Euclidean distance or the Hamming distance. A larger difference between ω and ω_m is represented by a larger value of the distance function, which in turn leads to a larger denominator and thus a smaller metric value.

4.6 Focus on Robustness

Robustness is part of the ML certification process; it deals with resilience of the model against potential unexpected or corner cases examples, which represent potential shortcomings a system needs to deal with in a safety-critical scenario. Thus, robustness is the property of a system to remain effective even outside its usual conditions of operation. Thus, AI robust models are required to be resilient to unknown inputs.

Robustness	Trustworthy attributes
Inherent Robustness	Global Robustness, Stability
System-Dependent Robustness	Durability, Resilience, Adaptability

Table 6: Some robustness attributes

In most ML approaches, robustness issues arise because of the distributional shift problem, i.e., when the training distribution the model was trained on is different from the deployment distribution. Examples of such phenomena are the adversarial attacks, where carefully crafted perturbations can deceive a ML model. Here again, we have decomposed robustness into two categories: "Inherent Robustness" and "System-Dependent Robustness". However, the AI component robustness assessment depends in general on the used technology.

4.7 Trustworthiness assessment take-away

This section presents the method used in the program Confiance.ai to tackle the issue of trustworthiness assessment, in the context of safety-critical AI-based systems. Trustworthiness is a complex notion, combining subjective aspects, heterogeneity of granularity in the attributes composing it, and non-commensurability of the different attributes. The approach consists in identifying the different attributes constituting the notion of trustworthiness, exploring each attribute to determine related metrics, assessment methods or control points, and defining an aggregation methodology based on a Multi-Criteria Decision Aiding approach. The work envisions the creation of a framework for the assessment of trustworthiness that leverages expert knowledge (for example in the definition of thresholds), a modelling of the environment of the application (e.g. influence of the Operational Design Domain on the selection of attributes), and usability in an engineering process (each atomic attribute is linked to a method or metric).

5. Other aspects

● Ethical aspects

Although Confiance.ai focuses on a technical approach to trustworthy AI, the program acknowledges the importance of ethical aspects in the line of the writings of the European Commission. Among the seven key requirements identified in the guidelines of the High-Level Expert Group:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing, and
7. Accountability,

Confiance.ai directly addresses requirement 2, and contributes to requirements 1, 4 and 5, not ignoring the other requirements which are indirectly impacted by our developments.

No need to restate why Confiance.ai addresses req. 2, this is obvious from the previous sections. Work on explainability contributes to human agency and oversight and on transparency. Work on bias identification and elimination contributes to non-discrimination and fairness. Some of our work on data management in EC5 is relevant for questions of privacy and governance. Our developments on assurance cases for certification in EC6 relate to accountability; societal and environmental wellbeing can be improved as a consequence of all these contributions.

● VV & Assurance Case[8]

Verification and validation activities are essential contributors to trust. Indeed, those means are essentially aimed at providing evidence that the system will realize the intended function. Towards that goal, we propose to establish a clear, traceable, auditable, and as formal as possible relationship between the engineering items produced by an AI system engineering activity, the properties that those items must satisfy, and the activities providing evidence that those properties are actually satisfied. In the Confiance.ai program, this relationship is captured by means of assurance cases.

An assurance case provides a structured argument to justify certain properties (sometimes called "claims") about the system, based on evidence concerning both the system and the environment in which it operates. The objective is to demonstrate as rigorously as possible that if some evidence is provided then some claim is justified. This argument cannot be as rigorous as a mathematical demonstration, simply because it does not refer to mathematical entities and mathematical properties. Nevertheless, the objective is to make it as close as possible to what a mathematical demonstration would be. In particular, terms and properties must be defined as precisely as possible, hypothesis and assumptions must be clearly stated, etc.

A claim concerns the satisfaction of some property by the system (e.g., system-level properties such as safety, security, or item-level properties such as "completeness", "consistency", etc.). Building an argument consist of decomposing the initial claim into sub-claims deemed easier to justify in a divide- and-conquer approach, down to the point where claims can be directly justified by showing evidences. During the construction of the argument, claims and properties may concern the whole system or some engineering items produced and used to engineer the system. As for the design of the system itself, building the assurance case of the system is not an ideal and strict top-down process that would start from some top-level property (e.g., "the system performs the intended function") and would be progressively decomposed into more and more primitive properties applicable to more and more primitive items. It is rather a combination of top-down and bottom-up approaches.

The introduction of AI in industrial practices strongly changes well-established practices, in particular those related to certification, and many initiatives are currently working on updating or defining new practices addressing the specificity of AI (e.g., EUROCAE WG114, ISO/TC204 WG14). It is worth noting that, in the context of Confiance.ai, assurance cases are strictly considered as a way to formalize the argumentation and build trust. It is a reference model from which other, specific representations, can be extracted. This includes, in particular, representations that will eventually be required by certification authorities. Our assurance cases are

- Not a verification and validation plan, but it may be used to build it.
- Not a Certification Standard, but it may be used to build one and, at least, it may be used to determine the activities to be carried out to comply with the existing ones. Traceability between the objectives / recommendations of standards and our assurance case is not yet addressed.

The methods identified in the assurance case to provide evidence are normally captured by verification and validation activities of the engineering process. Therefore, the overall process is the following:

- The design / development / deployment / etc. workflow is established and the artefacts involved in this workflow are described,
- Claims concerning the properties that those artefacts must possess are expressed,
- Assurance cases are built to show how those properties will be assessed,
- At the "bottom" of the argumentation (the leaves of the tree), one finds evidence,
- Evidence are brought thanks to some dedicated verification and validation activity,
- Those activities are inserted in the workflow to give the complete picture.

The assurance cases developed in Confiance.ai are fairly generic for they need to be applicable in different contexts (embedded system,

[8] based on: Morayo Adedjouma, Christophe Alix, Loic Cantat, Eric Jenn, Juliette Mattioli, et al.. Engineering Dependable AI Systems. *17th Annual System of Systems Engineering Conference 2022*, IEEE, Jun 2022, Rochester, United States. (hal-03700300)

production lines, etc.), industrial domains (aeronautics, space, automotive, etc.), and for different types of applications involving different types of sensors and algorithms, with different levels of criticality.
At the end of the day, the objective is to use the model to help making the optimal choice considering:

- the additional confidence brought by the method on the capability of the system to perform its intended function, and
- the cost of implementing the method.

To reach this objective, we are currently developing a tool that allows navigating between the engineering workflow and the assurance case. Thanks to this tool, the user will eventually be able to

- build a V&V strategy considering the risk level associated with errors affecting engineering item, cost and trust indicators associated with the production of evidence, and
- display the resulting engineering workflow including V&V activities.

● Standards

The current lack of AI standards is identified as a serious barrier to its development and wider usage in industry and society. Standards and norms are major components of trust for AI as well as for other technological domains. Several national and international bodies invest in AI standards development, including the International Standards Organisation (ISO) in its JTC1/SC42, the European Commission through CEN/CENELEC and ETSI in their Joint Technical Committee 21 (JTC 21). In France, AFNOR is in charge of preparing the national contribution, in the framework of Pillar 3 (norms and standards) of the Grand Challenge on Safe and Certified AI.
Confiance.ai contributes to the European initiative by participating in some ad hoc groups (AhG) of CEN/CENELEC that prepare the future European harmonized standards on AI. Confiance.ai experts have been chosen to coordinate some of these ad hoc groups, namely AhG7 “Overarching unified approach on trustworthiness characteristics”, and AhG8 “AI risks catalogue and risks management”, where relevant Confiance.ai deliverables (e.g., on robustness, bias, explainability, and the Confiance.ai taxonomy) are proposed as constitutive elements. Standards documents developed by these groups should be published in 2024 in support of the first release of the AI Act by the European Commission.

● Non-critical applications

Confiance.ai targets critical applications in industry, mobility, energy and environment, defense and security. All these sectors belong to what the European Commission identifies as “high-risk” applications for which specific demands are put for AI systems in the future AI Act. In the initial version of the AI Act, eight application domains were listed as high-risk, of which seven are not addressed in our program (e.g., biometric identification, employment, justice, education …). This does not mean that the methods and tools delivered by Confiance.ai are irrelevant for these other domains. On the contrary, all these domains should benefit from most methods and tools produced by Confiance.ai because of their genericity.
Moreover, this is true as well for other application sectors considered to be less risky (ie. characterized as “transparency risk” or “no risk” in the AI Act). In absence of obligations regarding robustness, explainability, fairness etc. for such applications, their promoters will still have the possibility to use Confiance.ai methods and tools if they want to for internal or for communication reasons.

Annex 1.

1. First version of Confiance.ai Taxonomy

1.1 Introduction

The confiance.ai taxonomy is today a list of keywords fixing concepts dealt with in the Confiance.ai program and characterizing the core domains of the trustworthy AI engineering including the following fields: AI Engineering, Data Engineering, Knowledge Engineering, Algorithm Engineering, Software and System Engineering, Safety Engineering, Human factor and Cognitive Engineering.

The various definitions have been collected through all the different states of the art realized in 2021 within the Confiance.ai program. In most cases, the definitions were taken from external literature by the working group in charge of the state of the art; in some few cases, literature does not offer a definition adapted to the scope of the program, and the working group coined a new one. This collection has been enriched in this present document with definitions coming from European and worldwide standardization (e.g. CEN, CENELEC, IEC, ISO), European projects (the Franco-Canadian DEEL project, JRC Flagship on AI), scientific publications or working groups such as the HLEG (High-Level Expert Group on Artificial Intelligence) or the AI Safety Landscape initiative (\url{https://www.ai-safety.org/}).

Moreover, in line with the two other pillars of the Grand National Challenge (and more specifically Pillar 3 dedicated to the French AI standardization strategy), this taxonomy will be updated during the overall duration of Confiance.ai, mainly to take into account the outcomes of all batches. In addition, this taxonomy will be consolidated to define the trusworthy AI engineering ontology.

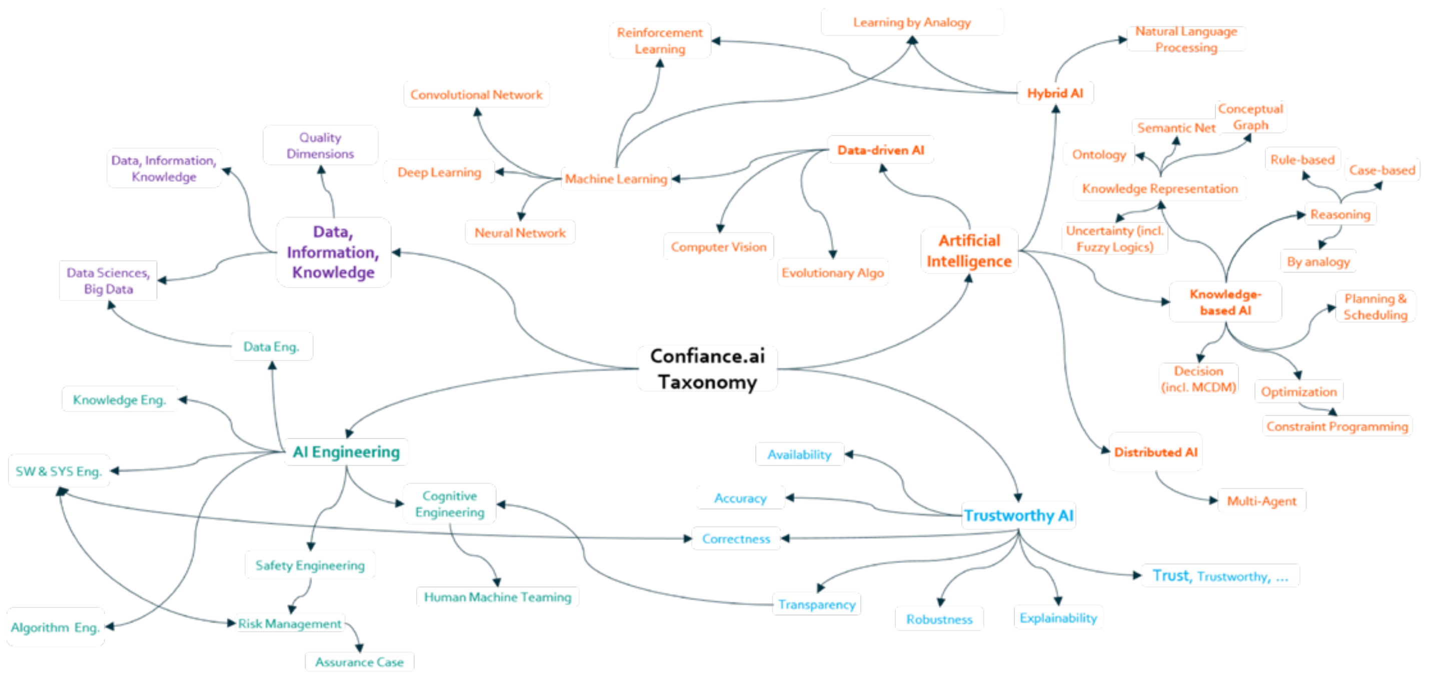

Figure 1: First levels of the Confiance.ai Taxonomy

At the end of Sept. 2022, Confiance.ai identified ~350 concepts, but for consistency, we will only present in this appendix those that are related to trust, and more specifically those that define trustworthiness attributes presented in this white paper.

These four groups of attributes form, along with trustworthiness governance, different perspectives to trustworthiness. An overall description of these perspectives is summarized in the mind map shown in fig. 9 of the paper, and reproduced below for convenience.

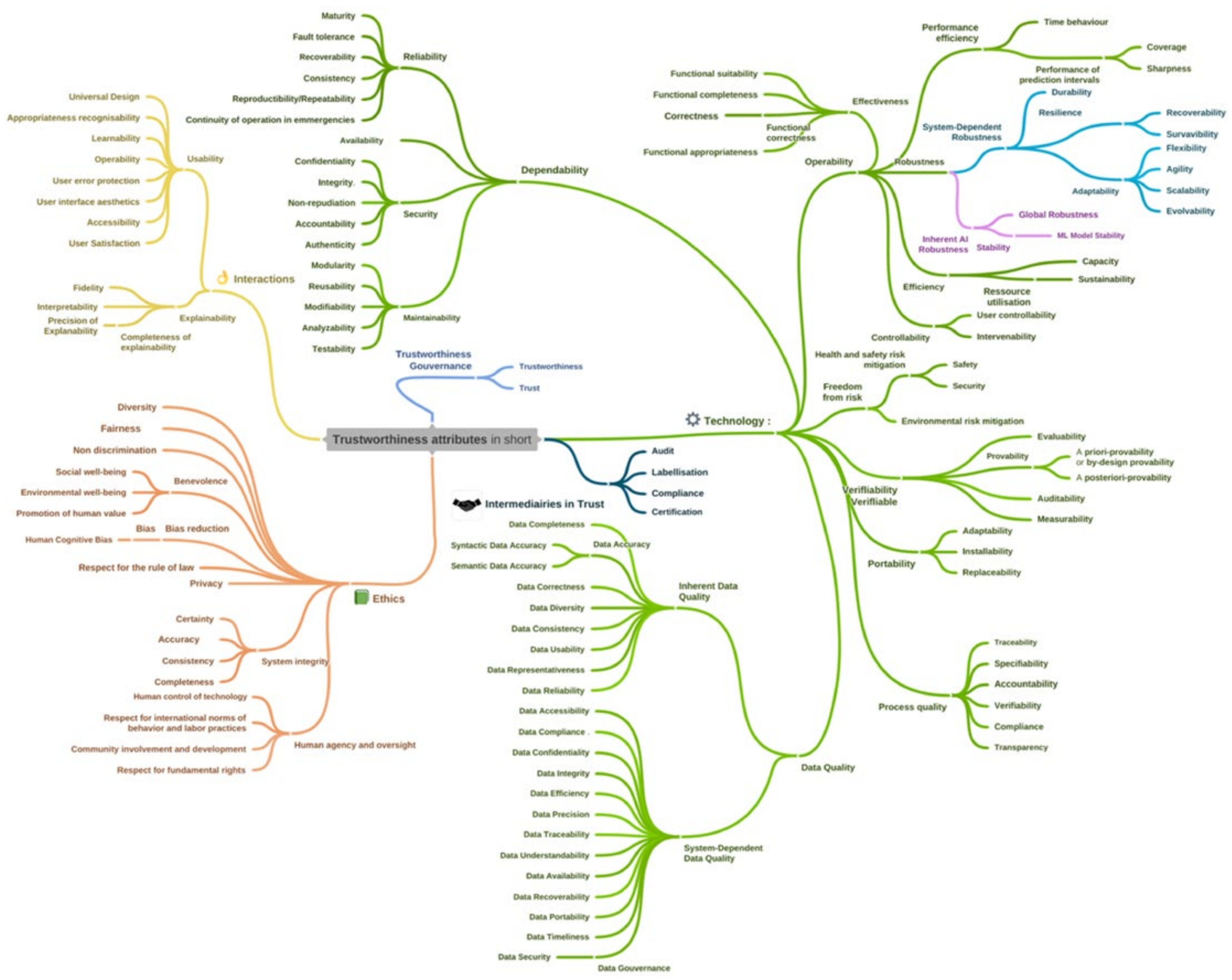

Figure 2: The AI-based system trustworthiness attribute mindmap

1.2 Trustworthiness attributes: Governance perspective

Trust [ISO/IEC 25010]:
degree to which a user or other stakeholder has confidence that a product or system will behave as intended.

Trustworthiness [ISO/IEC TR 24028]:
Ability to meet stakeholders' expectations in a verifiable way

Note 1: Depending on the context or sector and on the specific product or service, data and technology used, different characteristics apply and need verification to ensure stakeholder's expectations are met.

Note 2: Characteristics of trustworthiness include, for instance, reliability availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, usability.

Note 3: Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations.

[ISO/TS 21089:2018] In a governance context, **accountability** is the obligation of an individual or organization to account for its activities, for completion of a deliverable or task, accept the responsibility for those activities, deliverables or tasks, and to disclose the results in a transparent manner.

1.3 Trustworthiness attributes: Technical perspective

From a technical perspective, as the program confiance.ai aims to develop methods and tools to support critical AI-based systems during their life cycle (from specification to in service support through design and deployment), we have to demonstrate the properties of accuracy, robustness, safety and security. Thus, AI-based systems should generate accurate output consistent with the ground truth as much as possible. This is the most basic motivation for defining trustworthiness attributes. Additionally, AI systems should be robust to changes, even if real environments are usually very complex, dynamic and uncertain. Moreover, no AI programs or systems are expected to harm any people under any conditions, which put the safety of users as the first priority. In addition, the autonomy of trustworthy AI should always be under user's control. In other words, it is always a human right to grant an AI system any decision-making power or to withdraw that power at any time.

A. DATA QUALITY

Data Quality [ISO/IEC 25012]:
The grounds where the system for assessing the quality of data products is built on.
[ISO/IEC 25024:2015] Degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions
[DEEL Project]: The extent to which data are free of defects and possess desired features.

Inherent Data Quality [ISO/IEC 25012]:
Degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions.

- **Data Accuracy** [ISO/IEC 25012]:
Degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.
[DEEL Project] Accuracy depends on data gathering/generation and measures the faithfulness to the real value. It also measures the degree of ambiguity of the representation of the information.
  - **Syntactic Data Accuracy** [ISO/IEC 25012]: Closeness of the data values to a set of values defined in a domain considered syntactically correct.
  - **Semantic Data Accuracy** [ISO/IEC 25012]: Closeness of the data values to a set of values defined in a domain considered semantically correct.

- **Data Completeness** [ISO/IEC 25012] Degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
[EASA WG114] A dataset is considered to be complete if it is comprehensive in the sense that it has been sampled properly to cover the specified space of the ODD (Operational Design Domain) of the intended application.
[ED-76A] The degree of confidence that all of the data needed to support the intended use is provided

- **Data Correctness** [ED-76A]:
Data meeting stated quality requirements.

- **Data Diversity** is achieved by using different data sets, since those sets should capture the essence of requirements and any drawback impacts their fulfilment by the respective ML modules. The requirements can be achieved globally, i.e. relying upon different data sources, or locally, i.e., relying upon different subsets of the same source.

- **Data Consistency** [ISO/IEC 25012]:
Degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.
[DEEL Project] The deviation of values, domains, and formats between the original dataset and a pre-processed dataset

- **Data Usability** [DEEL Project]:
Quality bound to the credibility of data, i.e. if their correctness is regularly evaluated, and if data exist in the range of known or acceptable values.

- **Data Representativeness** [DEEL Project] Refers in statistics to the notion of sample and population. Transposed to AI, the sample corresponds to the dataset available for the development of the model (training, validation, testing), and the population corresponds to all possible observations in the field of application.
[EASA WG114]: A dataset is representative when it is complete, and the distribution of its key characteristics is similar to the intended space of the ODD of the targeted application.

- **Data Reliability** [EASA WG114]: Confidence level in the goodness of the data (e.g. because it's provided by a trusted source, or by a high-fidelity model).

System-Dependent Data Quality [ISO/IEC 25012]:
Degree to which data quality is reached and preserved within a computer system when data is used under specified conditions.

- **Data Accessibility** [ISO/IEC 25012]:
  Degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.
  [DEEL Project] The effort required to access data

- **Data Compliance** [ISO/IEC 25012]:
  Degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.

- **Data Confidentiality** [ISO/IEC 25012]:
  Degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.

- **Data Integrity** [ED-76A]:
  A degree of assurance that aeronautical data and its value has not been lost or altered (since the data origination or authorized amendment)

- **Data Efficiency** [ISO/IEC 25012]:
  Degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.

- **Data Precision** [ISO/IEC 25012]:
  Degree to which data has attributes that are exact or that provide discrimination in a specific context of use.

- **Data Traceability** [ISO/IEC 25012]:
  Degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.
  [DEEL Project]: Reflects how much both the data source and the data pipeline are available. Activities to identify all the data pipeline components have to be considered in order to guarantee such quality.

- **Data Understandability** [ISO/IEC 25012]:
  Degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.

- **Data Availability** [ISO/IEC 25012]:
  Degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use.

- **Data Portability** [ISO/IEC 25012]:
  Degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.

- **Data Recoverability** [ISO/IEC 25012]:
  Degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.

- **Data Timeliness** [DEEL Project]:
  The "time delay from data generation and acquisition to utilization". If required data cannot be collected in real time or if the data need to be accessible over a very long time and are not regularly updated, then information can be outdated or invalid.
  [EASA WG114] this property ensures that data are up-to-date and not obsolete.

- **Data Governance** [EASA WG114] the capability of an organization to ensure that high data quality exists throughout the complete life cycle of the data, and data controls are implemented that support business objectives. The key focus areas of data governance include data availability, usability, consistency, integrity, and sharing.

- **Data Security** [ISO 27000:2018] Preservation of confidentiality, integrity and availability of data. In addition, other properties, such as authenticity, accountability, non-repudiation, and reliability can also be involved.

- **Data integrity** [ISO/IEC 29167-19:2019] Property whereby data have not been altered in an unauthorized manner since they were created, transmitted, or stored

B. OPERABILITY

Effectiveness [ISO-9241]:
accuracy and completeness with which users achieve specified goals

- **Functional suitability** [ISO/IEC 25010]:
  Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions

- **Functional completeness** [ISO/IEC 25010]:
  Degree to which the set of functions covers all the specified tasks and user objectives

- **Functional correctness** [ISO/IEC 25010]:
  Degree to which a product or system provides the correct results with the needed degree of precision
  - **Correctness** [ISO-24765:2017]:
    Degree to which a system or component is free from faults in its specification, design, and implementation.
  - **Functional appropriateness** [ISO/IEC 25010]:
    Degree to which the functions facilitate the accomplishment of specified tasks and objectives

Robustness [ISO/IEC TR 24029-1]:
Ability of an AI system to maintain its level of performance under any circumstances

- **System-Dependent Robustness**
  - **Resilience** [DEEL Project] Ability for a system to continue to operate while an error or a fault has occurred
    [ISO/IEC TS 5723:2022] Capability of a system to maintain its functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary
    [HLEG2019ALTAI] Robustness when facing changes.
  - **Adaptability** [ISO/IEC 25010] Degree to which a product or system can effectively and efficiently be adapted for different or evolving hardware, software or other operational or usage environments.

- **Inherent AI Robustness**
  - **Global Robustness** [DEEL Project]:
    Ability of the system to perform the intended function in the presence of abnormal or unknown inputs
  - **Local Robustness** [DEEL Project]: The extent to which the system provides equivalent responses for similar inputs.

Efficiency: Relationship between the results achieved and the resources used. Resources expended in relation to the accuracy and completeness with which users achieve goals

- **Capacity** [ISO/IEC 25010]: Degree to which the maximum limits of the product or system, parameter meet requirements.
  - **Resource utilization** [ISO/IEC 25010]: Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

- **Performance efficiency** [ISO/IEC 25010]:
  Performance relative to the amount of resources used under stated conditions
  - Time behavior [ISO/IEC 25010]: Degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements

- **Controllability** [ISO 26262-1:2018]: Ability to avoid a specified harm or damage through the timely reactions of the persons involved, possibly with support from external measures
  - User controllability : Involved individual's possibility of avoiding harm in the situation that is putting him/her at risk

Dependability [avizienis2004basic]:
The ability to deliver service that can justifiably be trusted. It entails Availability: readiness for correct service; Reliability: continuity of correct service; Safety: absence of catastrophic consequences on the user(s) and the environment; Confidentiality: absence of unauthorized disclosure of information; Integrity: absence of improper system alterations; Maintainability: ability to undergo modifications, and repairs. Security: the concurrent existence of availability for authorized users only, confidentiality, and integrity (with 'improper' meaning 'unauthorized' here)
[HLEG2019ALTAI] Ability to deliver services that can justifiably be trusted.
[ISO/IEC/IEEE 15026-1:2019] Ability to perform as and when required

- **Reliability** [ISO/IEC 25010]:
  Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.
  [ISO/IEC 27000] Property of consistent intended behavior and results.
  [ISO/IEC TS 5723:2022] Ability of an item to perform as required, without failure, for a given time interval, under given conditions
  [ARP4761] The probability that an item will perform a required function under specified conditions, without failure, for a specified period
  - **Maturity** [ISO/IEC 25010]:
    Degree to which a system, product or component meets needs for reliability under normal operation.
  - **Fault tolerance** [ISO/IEC 25010]:
    Degree to which a system, product or component operates as intended despite the presence of hardware or software faults
  - **Recoverability** [ISO/IEC 25010]:
    Degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.
  - **Consistency** [ISO/IEC 21827]:
    Degree of uniformity, standardization and freedom from contradiction among the documents or parts of a system or component

- **Availability** [ISO/IEC 25010]:
  Degree to which a system, product or component is operational and accessible when required for use
  [EN50129] The ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval assuming that the required external resources are provided.

- **Security** [ISO/IEC 25010]:
  Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.
  - **Confidentiality** [ISO/IEC 25010]:
    Degree to which the prototype ensures that data are accessible only to those authorized to have access.
  - **Integrity** [ISO/IEC 27000:2018]:
    Property of accuracy and completeness.
    [ISO/IEC 27000] Property of protecting the accuracy and completeness of assets.
    ISO/IEC 25010] Degree to which a system, product or component prevents unauthorized access to, or modification of, computer programs or data.
  - **Non-repudiation** [ISO/IEC 25010]:
    Degree to which actions or events can be proven to have taken place, so that the events or actions cannot be repudiated later.

- **Accountability** [ISO/IEC 25010]:
  Degree to which the actions of an entity can be traced uniquely to the entity.
  [ISO 7498-2:1989] For systems, accountability is a property that ensures that actions of an entity can be traced uniquely to the entity.
- **Authenticity** [ISO/IEC 25010]:
  Degree to which the identity of a subject or resource can be proved to be the one claimed

• **Maintainability** [ISO/IEC 25010]:
Degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements
[DEEL Project]: Ability of extending/improving a given system while maintaining its compliance with the unchanged requirements.
[mamalet:hal-03176080]: Ability of extending/improving a given system while maintaining its compliance with the unchanged requirements

- **Modularity** [ISO/IEC 25010]:
  Degree to which a system or computer program is composed of discrete components such that a change to one component has minimal impact on other components.
- **Reusability** [ISO/IEC 25010]:
  Degree to which an asset can be used in more than one system, or in building other assets
- **Modifiability** [ISO/IEC 25010]:
  Degree to which a product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.
- **Analyzability** [ISO/IEC 25010]:
  Degree of effectiveness and efficiency with which it is possible to assess the impact on a product or system of an intended change to one or more of its parts, or to diagnose a product for deficiencies or causes of failures, or to identify parts to be modified.
- **Testability** [ISO/IEC 25010]:
  Degree of effectiveness and efficiency with which test criteria can be established for a system, product or component and tests can be performed to determine whether those criteria have been met.

C. FREEDOM FROM RISK

Health and safety risk mitigation

• **Safety** [ISO/IEC/IEEE 12207:2017]:
Property of a system such that it does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered
[ISO-12207:2017]: Expectation that a system does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered.
[ISO25010]: The ability to have acceptable risk levels in relation with damage to people, companies, software, property, or environment.
[EN50129]: Freedom from unacceptable risk of harm.
[ISO-51:2014]: Freedom from risk which is not tolerable.

• **Security** [ISO/IEC 23643:2020]:
Resistance to intentional, unauthorized act(s) designed to cause harm or damage to a system

D. VERIFLIABILITY AND VERIFLIABLE

Verifliability [mamalet:hal-03176080]:
Ability to evaluate an implementation of requirements to determine that they have been met (adapted from ARP4754A)

Verifliable [ISO/IEC/IEEE 15289]:
Can be checked for correctness by a person or tool

• **Provability** [DEEL Project]:
The extent to which a set of properties on this algorithm can be guaranteed mathematically.

- A **priori-provability** or **by-design provability** [DEEL Project]:
  The desired property is mathematically “transferable” as a design constraint to the ML algorithm. Then, to prove the property, it is necessary to demonstrate the validity of this transfer (i.e., if the design constraint is satisfied then the property holds on the model) and to demonstrate compliance with the design constraint.
- A **posteriori-provability** [DEEL Project]:
  The desired property is verified on the model after training. This approach may also rely on some assumptions on the ML algorithm (e.g. the architecture, the size of the network, the activation function type for a NN...), but these assumptions depends on the problem

• **Auditability** [DEEL project]:
The extent to which an independent examination of the development and verification process of the system can be performed

• **Measurability** [ISO/IEC TS 5723:2022]:
Ability to assess an attribute of an entity against a metric Note 1 to entry: The word “measurable” is the adjective form of measurability.

E. PORTABILITY

Portability [ISO/IEC 25010]:
Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another

• **Adaptability** [ISO/IEC 25010]:
Degree to which a product or system can effectively and efficiently be adapted for different or evolving hardware, software or other operational or usage environments.

• **Installability** [ISO/IEC 25010]:
Degree of effectiveness and efficiency in which a product or system can be successfully installed and/or uninstalled in a specified environment.

• **Replaceability** [ISO/IEC 25010]:
Degree to which a product can replace another specified software product for the same purpose in the same environment.

F. PROCESS QUALITY

Traceability [EASA WG114]:
An association between artifacts, such as between process outputs or between an output and its originating process
[IEEE-610.12-1990:2002]: (1) The degree to which a relationship can be established between, two or more products of the development process, especially products having a predecessor, successor, or master-subordinate relationship to one another; for example, the degree to which the requirements and design of a given software component match. (2) The degree to which each element in a software development product establishes its reason for existing; for example, the degree to which each element in a bubble chart references the requirement that it satisfies.

Specifiability [DEEL Project]:
The extent to which the system can be correctly and completely described through a list of requirements.

Accountability [ISO-24028:2020]:
Property that ensures that the actions of an entity may be traced uniquely to that entity

Verifiability [DEEL Project]:
Ability to evaluate an implementation of requirements to determine that they have been met

Compliance [CENELEC-EN50126]:
A demonstration that a characteristic or property of a product satisfies the stated requirements.

Transparency[ISO/IEC 27036-3:2013]:
Property of a system or process to imply openness and accountability ISO-22989 Property of a system that appropriate information about the system is communicated to relevant stakeholders.
[arrieta2020explainable]: A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models are divided into three categories: simulable models, decomposable models and algorithmically transparent models.
[brundage2020toward]: Making information about the characteristics of an AI developer’s operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.

1.4 Trustworthiness attributes: Interaction perspective

From interaction’s perspective, trustworthy AI should possess the properties of usability, and explainability. Specifically, AI-based systems should not cease operation at inappropriate times (e.g. at times when the lack of output could lead to safety risks), and these programs or systems should be easy to use for people with different backgrounds. Last, but not least, trustworthy AI must allow for explanation and analysis by humans, so that potential risks and harm can be minimized, and human users can remain empowered. In addition, trustworthy AI should be transparent so people can better understand its mechanism.

A. USABILITY

Usability only exists with regard to functionality and refers to the ease of use for a given function. The ability to learn how to use a system (learnability) is also a major sub characteristic of usability.
[ISO/IEC 25010] 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.
[ISO 9241-11:2018] Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use:

- The “specified” users, goals and context of use refer to the particular combination of users, goals and context of use for which usability is being considered.
- The word “usability” is also used as a qualifier to refer to the design knowledge, competencies, activities and design attributes that contribute to usability, such as usability expertise, usability professional, usability engineering, usability method, usability evaluation, usability heuristic

• **Universal Design** [Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS 2021]:
An AI system should address the widest possible range of users and, in particular, be accessible to them. Therefore, universal design principles should be considered during the planning and development of an AI system, especially taking into account and, if possible, involving, potential end-users with special needs. It should be ensured that no group of people is disproportionately affected by the results of the AI system.

• **Appropriateness recognisability** [ISO/IEC 25010]:
Degree to which users can recognize whether a product or system is appropriate for their needs.

• **Learnability** [ISO/IEC 25010]:
Degree to which a product or system enables the user to learn how to use it with effectiveness, efficiency in emergency situations.

• **Operability** [ISO/IEC 25010]:
Degree to which a product or system is easy to operate, control and appropriate to use

- **User error protection** [ISO/IEC 25010]:
Degree to which a product or system protects users against making errors

- **User interface aesthetics** [ISO/IEC 25010]:
Degree to which a user interface enables pleasing and satisfying interaction for the user.

- **Accessibility** [ISO/IEC 25010] Degree to which a product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use.

B. EXPLAINABILITY

Explainability [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans [ISO/IEC DIS 22989] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.

- **Explainability benchmark**:
Given a set of libraries and methods for explainability, construct a decision tree that allows users to select the most relevant library and method for their context.

- **Explainability evaluation**:
The evaluation of the appropriateness of explainability libraries and methods is performed per use case.

- **Interpretability** [DEEL Project] 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.

- **Completeness of explainability** [DEEL Project] relates to the capability to describe a phenomenon in such a way that this description can be used to reach a given goal.
  - **Precision of Explainability** [DEEL Project] indicates how much details must be provided to the human to let her/him execute mentally the inference in a right way with respect to her/his goal. For instance, there is no need to know the laws of general relativity or quantum mechanics to predict the trajectory of a ball.

1.5 Trustworthiness attributes: Ethical perspective

From ethical perspective, trustworthy AI should be law-abiding, fair, accountable, environmentally friendly and compliant with the user privacy. Specifically, AI systems should operate in full compliance with all relevant laws and regulations and comply with the ethical principles of human society.

- **Fairness** [Stevenson, 2015]:
Impartial and just treatment or behavior without favoritism or discrimination
[High-Level Expert Group on Artificial Intelligence, Assessment List for Trustworthy AI (ALTAI), 2019] Fairness refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally referred to as 'substantive' fairness. But fairness also encompasses a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated

- **Benevolence** [Mayer 1995]:
The extent to which a trustee is believed to want to do good the trustor. Benevolence suggest that the trustee has some specific attachment to the trustor
[ISO 24028] The extent to which the AI system is believed to do good or in other terms, to what extent "DO NO Harm" principle is respected

- **Environmental well-being** [Fraunhofer IAIS]:
The environmental impact of an AI system should be assessed throughout its life cycle and its entire supply chain. Measures should be taken to reduce this impact

- **Bias reduction**:
  - **Bias** [ISO-29119-11:2020]:
  Measure of the distance between the predicted value provided by the ML model and a desired fair prediction.
  [ISO-24028:2020]: Favoritism towards some things, people or groups over others.
  - **Human Cognitive Bias** [ISO/IEC TR 24027:2021:] Bias that occurs when humans are processing and interpreting information. Note 1 to entry: human cognitive bias influences judgement and decision-making.

- **Privacy** [ISO/IEC 2382]:
Freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual

- **Human agency and oversight** [High-Level Expert Group on Artificial Intelligence, 2019]:
AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight.
  - **Respect for fundamental rights** [High-Level Expert Group on Artificial Intelligence, 2019] A fundamental rights impact assessment should be undertaken. This should be done prior to the system's development and include an evaluation of whether those risks can be reduced or justified as necessary in a democratic society in order to respect the rights and freedoms of others. Moreover, mechanisms should be put into place to receive external feedback regarding AI systems that potentially infringe on fundamental rights.
  - Human control of technology "**human oversight**" [High-Level Expert Group on Artificial Intelligence, 2019]: Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach

- **System integrity** [ISO/IEC 27000:2018]:
Property of protecting the accuracy and completeness of assets.
[ISO/IEC 25010] Degree to which a system, product or component prevents unauthorized access to, or modification of, computer programs or data.
[ISO 24028] An AI system's respect of sound moral and ethical principles or the assurance that information will not be manipulated in a malicious way by the AI system.
  - **Accuracy** [ED-76A]: Degree of conformance between the estimated or measured value and its true value
  ISO-24765:2017]: A qualitative assessment of correctness, or freedom from error.
  [ISO-24765:2017]: A quantitative measure of the magnitude of error.
  [ISO-24765:2017]: Within the quality management system, accuracy is an assessment of correctness. [ISO 17572-1:2015]: Measure of closeness of results of observations, computations, or estimates to the true values or the values accepted as being true

Annex 2

list of contributors

By alphabetical order

Jean-Luc Adam, *Renault*
Morayo Adedjouma, *CEA*
Patrice Aknin, *IRT SystemX*
Christophe Alix, *Thales*
Xavier Baril, *Airbus*
Guillaume Bernard, *LNE*
Yannick Bonhomme, *IRT SystemX*
Bertrand Braunschweig, *IRT SystemX*
Loïc Cantat, *IRT SystemX*
Guillermo Chale-Gongora, *Thales*
Zakaria Chihani, *CEA*
Philippe Dejean, *IRT Saint-Exupéry*
Agnès Delaborde, *LNE*
Geoffrey Delhomme, *Airbus*
Flora Dellinger, *Valéo*
Rodolphe Gélin, *Renault*
Hatem Hajri, *IRT SystemX*
Eric Jenn, *IRT Saint-Exupéry*
Fateh Kaakai, *Thales*
Angélique Loesch, *CEA*
Juliette Mattioli, *Thales*
Yves Nicolas, *Sopra Steria*
Paul-Marie Raffi, *IRT SystemX*
Boris Robert, *IRT Saint-Exupéry*
Henri Sohier, *IRT SystemX*
François Terrier, *CEA*
Fabien Tschirhart, *IRT SystemX*
Jean-Luc Voirin, *Thales*
Thomas Wouters, *IRT SystemX*
Jacques Yelloz, *Safran*

Notes

To learn more about
the Confiance.ai program:

www.confiance.ai

Confiance.ai

Confianceai

Director of publication:
Michel MORVAN
Editorial director:
Aurélie BOURRAT
Chief Editor:
Samanta DUGUAY-FANTI
Pictures:
Shutterstock
Graphic design:
www.maiffret.net
Printing:
Burlet Graphics
Contact:
contact@irt-systemx.fr

Founding members

www.confiance.ai

Confiance.ai

Confianceai