AI engineering to deploy reliable AI in industry

Juliette Mattioli, Xavier Le Roux, Bertrand Braunschweig, Loic Cantat, Fabien Tschirhart, Boris Robert, Rodolphe Gelin, Yves Nicolas

AI engineering to deploy reliable AI in industry

J. Mattioli
X. Le Roux
Thales, France
IRT SystemX, France

B. Braunschweig
L. Cantat
F. Tschirhart
IRT SystemX, France

B. Robert
IRT Saint Exupéry, France
IRT SystemX, France

R. Gelin
Renault, France
IRT SystemX, France

Y. Nicolas
Sopra-Steria, France
IRT SystemX, France

***Abstract*—To bring competitive advantage to industry through a sound AI deployment, we need an end-to-end "AI systems engineering" process covering the overall lifecycle of an AI system, both at component level and at system level, regardless of whether the specifications come from regulation and reliability concerns.**

***Index Terms*—AI engineering, reliability, ODD**

I. AI ENGINEERING CONCERNS

Artificial Intelligence (AI) can bring competitive advantage to industry through decision support and the ability to offer higher value-added products and services. Delivering the expected service safely (*conformance to requirements*), meeting user expectations (*fitness for use*) and maintaining service continuity will determine its adoption and its use in industry. AI becomes critical for companies looking to extract value from data and knowledge by automating and optimizing processes, producing actionable insights, and making a proactive decision under risks and uncertainties. Production efficiency, product quality, and service level will be improved by AI [5] by providing typical features such as machine learning (ML), reasoning and decision support. However, concerns such as ethics, accountability, liability, security, privacy and trust are receiving increasing attention in many emerging areas such as future industry. So far, AI systems are also expected to address the risks associated with these concerns.

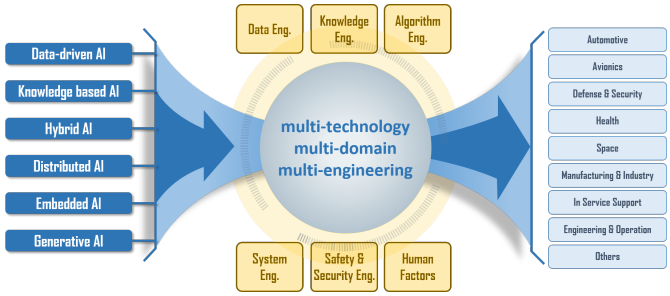

Fig. 1. AI system engineering is a multi-engineering process

A successful strategy to overcome these challenges requires collective actions around the objectives of a common industrial and reliable AI strategy to strengthen synergies and develop engineering best practices. To achieve this goal, we need an **AI engineering workbench with a sound process, methods and tools to support the overall lifecycle of an AI system** [1], both at component level and at system level, regardless of whether the specifications come from regulation, safety or security, standardization *etc.* The objective of the *Confiance.ai* program is to revisit "conventional" engineering (data and knowledge engineering, algorithm engineering, system and software engineering, safety and cyber-security engineering, and cognitive engineering) to ensure the system's compliance with requirements and constraints (fig. 1) and to guarantee RAMS (Reliability, Availability, Maintainability and Safety) properties. The challenge is to design an end-to-end "AI system engineering" process covering the entire value chain to industrialize AI.

This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute within the Confiance.ai Program (www.confiance.ai).

In the following, an AI system refers to a software-based system that contains AI-based components alongside traditional software components. It is an artificial system that acts in the physical or digital dimension through cognitive capacities by handling its environment by collecting data, interpreting the collected structured or unstructured data, inferring the knowledge or processing the information derived from this data, and deciding on the best activity(ies) to take in order to achieve the given objective. AI covers a wide range of technologies that can be divided into two broad categories: (1) data-driven AI, which includes neural networks, deep learning (DL), genetic algorithms *etc.*; and (2) knowledge-based AI, also known as symbolic AI, which includes other non-ML techniques and methods such as fuzzy logic, ontology and rule-based systems. Hybrid AI encompasses any synergistic combinations of various AI techniques such as extension or optimization of data-driven AI with expert knowledge. "*An AI system can either use symbolic rules or learn a numerical model, and it can also adapt its behaviour by analysing how the environment is affected by its previous actions*" [3].

There are some characteristics that distinguish AI systems from classical systems, which are scoped within a given set of requirements (functional and non-functional), thus defining a design domain. For example, non-determinism is a distinctive feature of AI systems which is particularly useful during solution searching, especially for exploring complex (n-dimensional, infinity, heterogeneous) problem spaces. However, non-determinism also raises questions regarding our understanding on how and why AI algorithms find local or global solutions (their explainability).

In the case of data-driven AI systems, ML/DL subsystems or components are based upon parameters, which may require

to be set during design, implementation and validation phases. For these subsystems and components, a ML phase should be introduced whereas for other typical (non-ML-based) subsystems and components, the ML phase does not exist. At the minimum, an additional iteration phase is expected in order to conduct such training and tuning.

II. AI SPECIFICS TO BE CONSIDERED

Reliability concerns the AI system itself, but also processes (how the system was made), tools and infrastructure (what with), people (by whom) and governance (who decides). Its assessment also combines different approaches, such as risk management and quality management. Its system lifecycle processes (based on ISO/IEC/IEEE 15288) and standards should help address new challenges posed by AI systems by integrating existing AI-specific processes and methodologies. These challenges affect all AI systems or components and need to be addressed simultaneously. This end-to-end AI engineering methodology allows to elaborate strategies for development and IVVQ (Integration, Verification, Validation, Qualification) activities by:

- Defining analysis perspectives, formalized by a metamodel defining the concepts involved and their semantic relations.
- Consolidating the methodological results by analyzing their various aspects: technical context, constraints, activities, data/knowledge, lifecycle, *etc.*.
- Formalizing the analyzed methods in a modeling tool, according to the metamodels of the considered aspects. Modeling will help ensure that all methods are compatible with each other, providing a consistent end-to-end process for designing reliable AI systems.

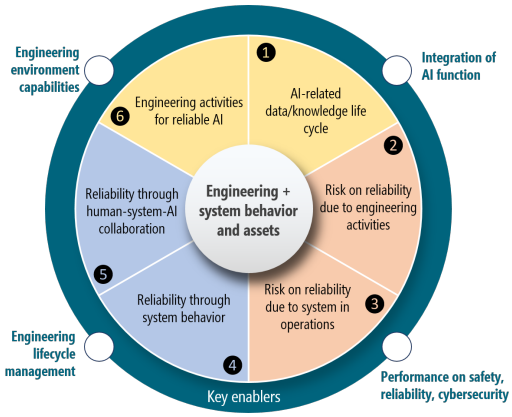

Fig. 2. Global view of the analysis framework [2]

The overall structure of the analysis framework [2] is depicted on fig.2. The viewpoints for AI system development are:

- Two generic viewpoints: (1) Engineering activities for reliable AI: Define the tasks to perform so as to specify, design, produce, deploy and operate an appropriate and reliable solution to a well understood need, involving AI techniques; and (2) AI-related data/knowledge 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 (*i.e.,* risk on the capability of the system to deliver the expected service is reduced or lost): (3) Risk on reliability due to engineering: identify major sources of bias or errors brought by other engineering activities to inputs and outputs of AI engineering and data/knowledge; (4) Risk on reliability due to system during operation: identify the main sources of bias or corruption introduced by other components of the system that interact with the AI components in operation.
- Two viewpoints dedicated to trust development and support: (5) Reliability through system behavior: specify system capabilities needed to ensure reliability in operation; and (6) Reliability through Human/System collaboration: specifying the expectations of human stakeholders, their role and work share with the system & AI, in reliably delivering the expected capacities.

Four transverse system viewpoints are identified (ring of fig.2):

- Integration of AI functions: characterize address specific issues related to the integration of one or more AI functions together in the target system context; provide guidance on how to address each issue,
- Performance on reliability, safety, security: define main needs, contributions and obstacles regarding reliability applied to AI decision RAMS performance 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 reliable AI systems engineering feasible, scalable, efficient and secure.

Consistency of the content in all those viewpoints shall be checked.

III. A NEW AI RELIABLE METAMODEL

To capture information and different interrelations needed to assess AI reliability, *Confiance.ai* proposes a metamodel with concepts at different levels of abstraction. (see fig. 3).

The red part describes the way the tree of attributes is built. It highlights the abstract concepts central to reliability assessment. An attribute which aggregates other attributes is called a macro-attribute (e.g. robustness, explainability *etc.*). It is assessed with an aggregation method. An atomic attribute is assessed with a clear and actionable observable which can take different forms (metric, ”expected proof”). The green part of fig. 3 is the metamodel fragment with concrete concepts. These concepts represent the different possible subjects and relations between them. For example, the product is developed following processes as technical processes (through which the product must go: design definition, implementation, operation, *etc.*), agreement processes (with external organizations: acquisition, supply), and management processes (supporting

the development of the product: quality management, risk management *etc.*). Risk and quality management ensure the compliance with the specification which includes the different expected reliability attributes. Processes are applied with tools by people respecting a certain governance. The blue part summarizes key systems engineering concepts, part of the non-functional specification: defining not what the system "does" or how it works, but what it "is". Because they often have this suffix, attributes are often referred to as "*-ilities*". They can also be called quality requirements. They are influenced by stakeholders such as the user, the developer *etc.*, whether a specification is functional or non-functional.

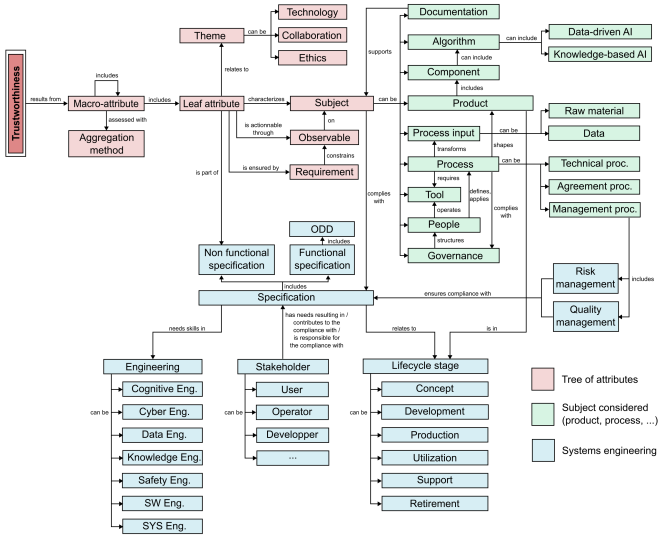

Fig. 3. A new AI reliable metamodel [6]

In contrast to non-functional requirements, which define what the system is, functional requirements define what the system does: does it move? does it roll? does it roll fast? under what conditions? From this point of view, the Operational Design Domain (ODD), which characterizes the operating conditions of the system/feature of interest, can be considered as part of the functional specification in relation to the reliability attributes in a number of ways: 1) the transparency of the ODD makes it possible to understand the limitations of the system (a requirement of the AI Act); 2) the ODD is the domain to be considered for the different operational reliability attributes; 3) the ODD has its own attributes (it should be complete, free of inconsistencies, human readable, *etc.*).

At every stage of the system lifecycle (see fig 4), from engineering and design to operation, RAMS relationships must be established and maintained. According to the seven pillars of reliability [3], *Confiance.ai* specifies AI reliability [6] by six macro-attributes: data/information/knowledge quality, dependability, operability, robustness, explainability/interpretability, and human control.

A. AI features specifications through the ODD analysis

All along the methodological steps, a risks/opportunities-driven approach has to be applied to ensure that the engineering orientations, decisions and technological choices, specific to the involvement of AI, are identified, analyzed and mitigated as early as possible (e.g. datasets evolution, occurrence of unwanted emerging behaviors ...). Potential opportunities, typically enabled by some technologies, must be considered, analyzed and valorized. This risks/opportunities-driven approach may require additional iterations of the qualification and certifications phases. This process aims at defining the subset of a target operational domain where systems/features of interest can be automated with an acceptable level of confidence. These systems/features have to be considered as functional chains, potentially involving several AI-based and non-AI-based components. This process relies on the reliability assessment process that aims at identifying and characterizing expectations on trust. To achieve this objective, the ODD analysis process (see fig. 4) hearafter from an initial legacy system/feature, to be identified from business domains and/or technological domains. This legacy system/feature is considered as a reference.

From this reference, system/feature automation objectives, expected level of automation and design intentions can be defined (for instance, human activities or behaviors that could be automated at a particular level, with regard to their operational environments, conditions and trust expectations). Based on the trust characteristics analyzed by the reliability Assessment Process, the ODD analysis process defines the observable/measurable conditions and properties that need to be supervised and monitored. It also defines nominal and edge/corner cases scenarios. All this characterizes and describes the ODD of the AI systems/features, related to the automation objectives, their associated level and design intentions and the target operational environment.

B. The reliability assessment process

This process aims at analyzing and characterizing the trust expectations related to the targeted automation objectives [6]. It contributes, along with the ODD analysis process, to define the observable/measurable conditions and properties. Based on a set of trust assessment categories, this process contributes to analyze the automation objectives, their operational environment and to define the characteristics and properties to be observed, monitored and measured to guarantee the trustful operation of the target system/feature. It defines a multi-viewpoint value analysis approach, enabling to consider the concerns and expectations of involved stakeholders, and a trade-off approach to find the best compromise among trust attributes and expectations. The ODD analysis process and the reliability assessment process contribute, with other traditional System Engineering processes, to state automation requirements and tests definitions.

C. The Algorithm Engineering Process

As already mentioned, beyond the impact on Systems Engineering and consequently on the DDQ (Design, Develop and Quality) processes, AI engineering strongly impacts the algorithm and software engineering approach. Indeed, AI

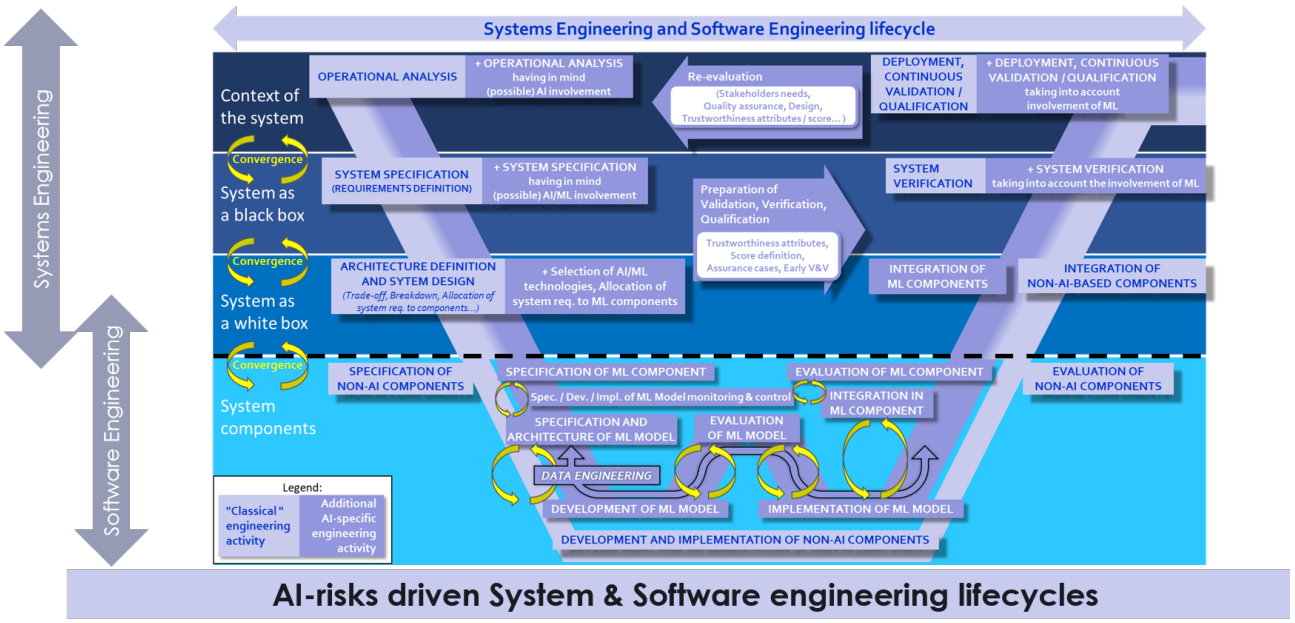

Fig. 4. The "W" AI/ML Engineering approach

system engineering has to be considered as a true software engineering discipline that stage at algorithm level. Algorithmic engineering [7] has often been overlooked in the last decades. Today, engineers cannot avoid this discipline when considering AI applications (especially for ML) since it has many impacts at software level in terms of accuracy, performance, robustness *etc.* Fig. 4 summarizes, in a W cycle, a proposed AI software Engineering approach. This approach focuses on the AI/ML technological paradigm, because this kind of AI paradigm is explored by *Confiance.ai*.

In order to better understand the gap between "conventional" Software Engineering and AI/ML Engineering, fig. 4 emphasizes additional activities: most importantly, this W cycle that shows that the ML model has to be evaluated (to provide some levels of reliability at algorithm level) before its implementation (at software level).

D. The Assurance Case Process

This process aims at supporting the AI-based systems/features V&V (validation & verification) strategy by characterizing and defining evidence-based justifications of the reliability [4]. It provides a top-down approach that should be applied to decompose high-level properties, and a bottom-up approach that can be used to exploit the available evidences. Both approaches have to be combined to reach the best possible provable level of reliability.

IV. Conclusion

Adopting AI in our industry poses many technical and non-technical challenges, and significant efforts are currently underway to address these challenges and facilitate early industrial adoption of AI cost effectively and safely. In the context of the Confiance.ai program, the most critical of these challenges have been identified. They cover two main aspects: *Realizability* to provide the ability to have justified confidence in the system's ability to deliver the expected service, and *Industrial Efficiency* to ensure that dependability is achieved in a cost-effective manner.

Confiance.ai addresses most dimensions of this problem, from providing new AI/ML algorithms and techniques that address different dimensions of trust, including reliability, accountability, fairness, time determinism, and so on. But the engineering practices themselves must also be updated to account for the specificity of AI. Towards that goal, we propose an "**AI/ML Engineering Framework**" to build a system development and V&V workflow which explicitly integrates the different dimensions of reliability. The framework relies on a model-based approach involving a series of ten viewpoints capturing the many aspects of system development including those related to data engineering, risk analysis, *etc.*

References

[1] J. Adam et al. Towards the engineering of trustworthy AI applications for critical systems - The Confiance. ai program, 2022.

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

[3] High-Level Expert Group on Artificial Intelligence. Assessment list for trustworthy artificial intelligence (altai). Technical report, European Commission, 2019.

[4] F. Kaakai et al. Toward a machine learning development lifecycle for product certification and approval in aviation. *SAE International Journal of Aerospace*, 15(01-15-02-0009):127–143, 2022.

[5] B. Li et al. Applications of artificial intelligence in intelligent manufacturing: a review. *Frontiers of Information Technology & Electronic Engineering*, 18(1):86–96, 2017.

[6] J. Mattioli et al. An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering. In *AI Trustworthiness Assessment (AITA) @ AAAI Spring Symposium*, 2023.

[7] Peter Sanders. Algorithm engineering–an attempt at a definition. In *Efficient algorithms*, pages 321–340. Springer, 2009.