An end-to-end method for operationalizing trustworthiness in AI-based critical systems

Karla QUINTERO, Lucas MATTIOLI,
Henri SOHIER, Juliette MATTIOLI*

*Thales France - juliette.mattioli@thalesgroup.com

Juliette Mattioli

- As VP Thales AI fellow, Juliette Mattioli is considered a reference in AI not only within Thales but also in France. In 2017, she was one of the five representatives of France at the G7 Innovators Conference. Since 2019, she is President of the "Data Sciences & Artificial Intelligence" Hub of the Systematic Paris-Region competitiveness cluster.

- Recognized for her excellent knowledge of industrial AI issues, she advocated for Hybrid AI deployment and contributes in the field of AI engineering with a particular focus on trustworthy and responsible AI to accelerate the industrial deployment of AI-based solutions in critical systems.

Aims and contributions of our paper

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

Reminder: EU AI-system definition comprises seven main elements
(Feb. 2025)

A machine-based system

that is designed to operate with varying levels of autonomy

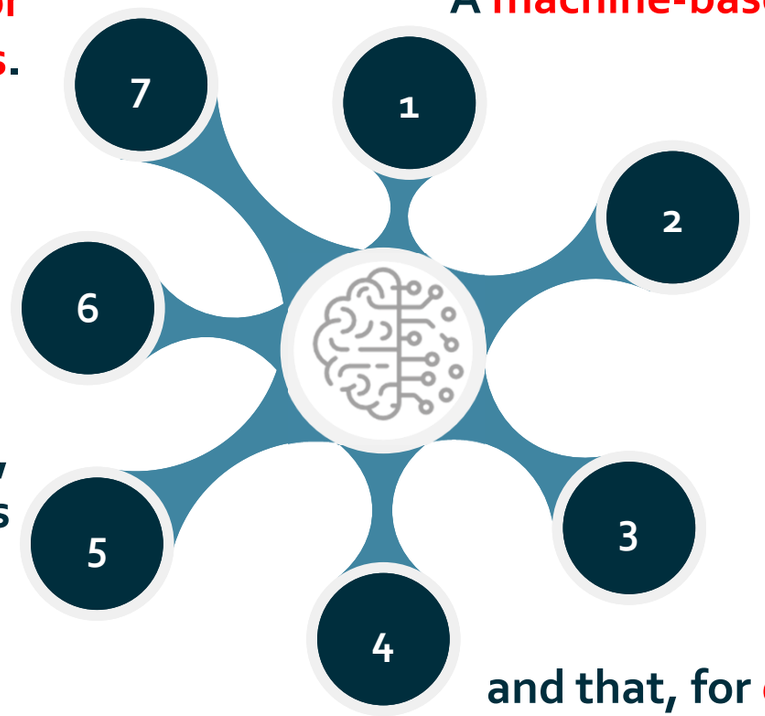
that may exhibit adaptiveness after deployment;

and that, for explicit or implicit objectives;

infers, from the input it receives, how to generate outputs

such as predictions, content, recommendations, or decisions

that can influence physical or virtual environments.



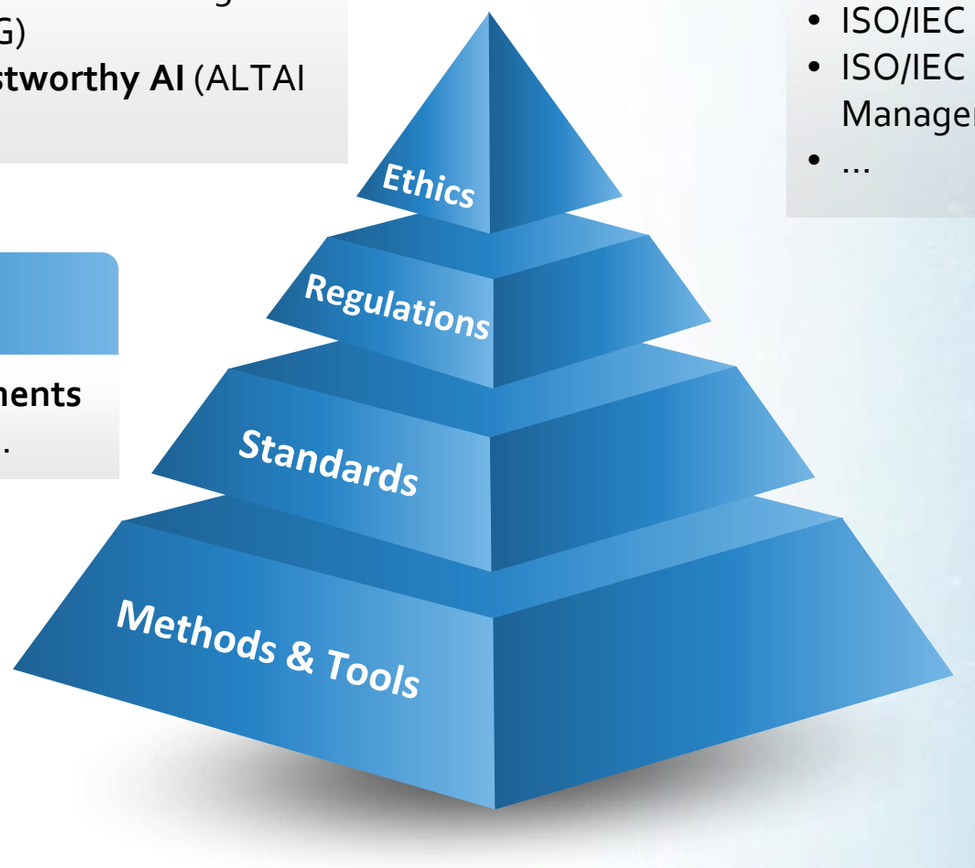
AI regulation, standards and tools

Ethics

- **Recommendations** from organizations like UNESCO and the OECD, or from EU high-level expert groups (HLEG)
- **Assessment List for Trustworthy AI** (ALTAI - 2020)

Regulations

High Level, long-term requirements
e.g. **European AI Act**, Data Act…



Standards

Glossary and technical requirements e.g.

- ISO/IEC 22989: AI concepts and terminology
- ISO 5338: the life cycle of AI systems based on ML
- ISO/IEC 23053: Framework for AI Systems Using ML
- ISO/IEC 42001: Information technology — AI — Management system
- …

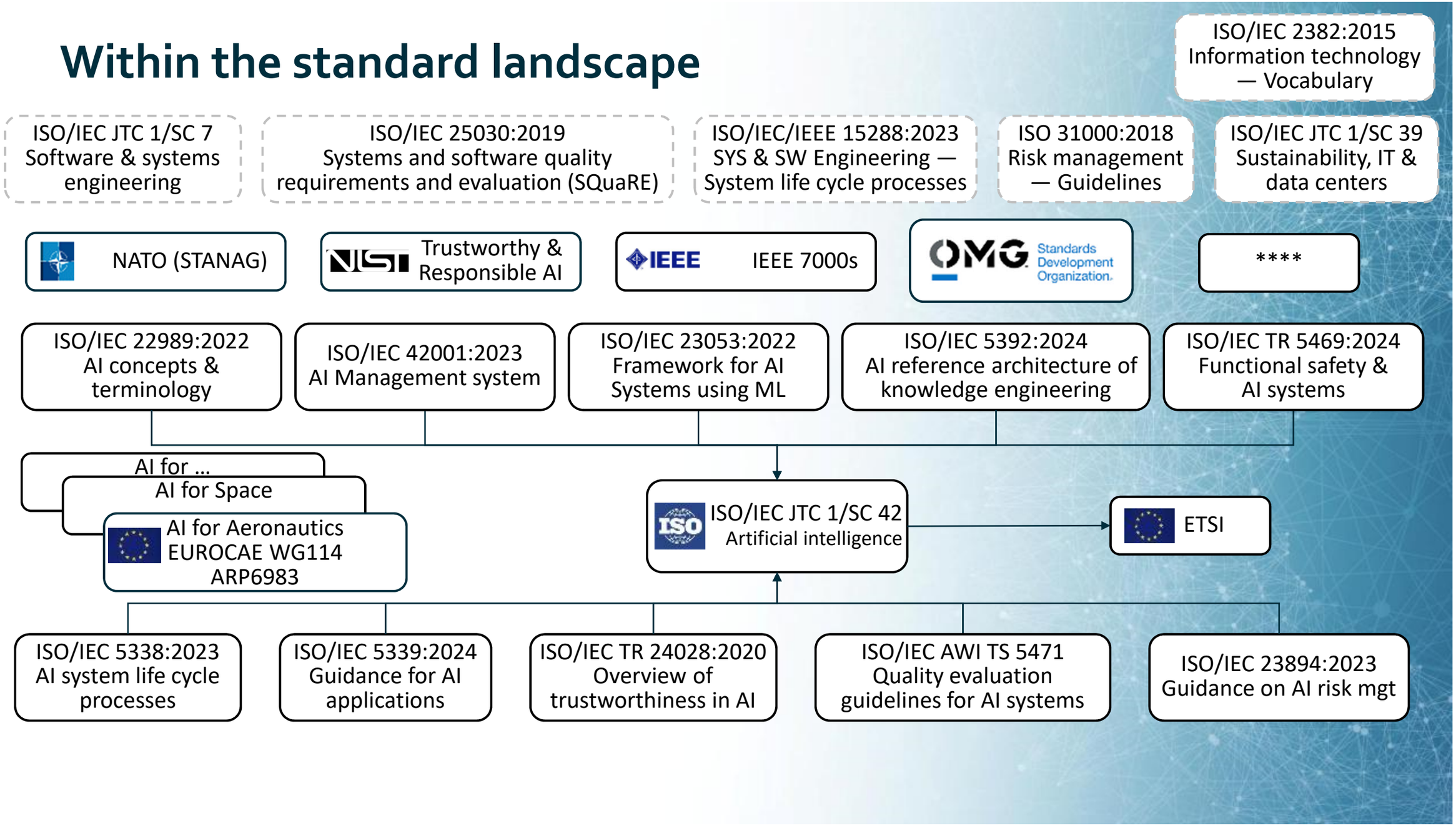
Methods & Tools

- Concepts of Design Assurance for Neural Networks – CoDANN [EASA]
- AI Pact: to support for the implementation of the AI Act.
- **Tooled Confiance.ai End-to-End methodology**
- MLOps/AIOps tool chain
- …

Reminder: Data & AI Regulation

GDPR *Applicable since May 2018*	EU DATA GOVERNANCE ACT *Applicable September 2023*	DATA ACT *Applicable September 2025*	AI ACT *Applicable August 2026 for "High risk" systems*	AI liability package *Still under discussion*
Accountability Right to access Breach notification Harmonised fines	Labelled data intermediaries EU Governance on data Data Sovereignty	Access to data generated by IoT to users Data sharing with public sector for emergencies Easier change of cloud provider Data Sovereignty	Risk based approach Robustness, accuracy and cybersecurity Transparency and information Governance and quality of datasets Human oversight Conformity assessment	Creation of an extra-contractual liability regime to repair prejudice caused by AI Revision of the product liability directive to include AI within it scope
Ensure that companies handle personal data in a responsible and accountable way, and that individuals have greater control over their data	Regulation that aims at facilitating the data sharing across EU	Regulation that aims at facilitating the data sharing across EU	Requirements and obligations tailored on a 'risk-based approach' and the role in the AI value chain	Make sure that persons harmed by AI enjoy the same level of protection as persons harmed by other technologies

Within the standard landscape



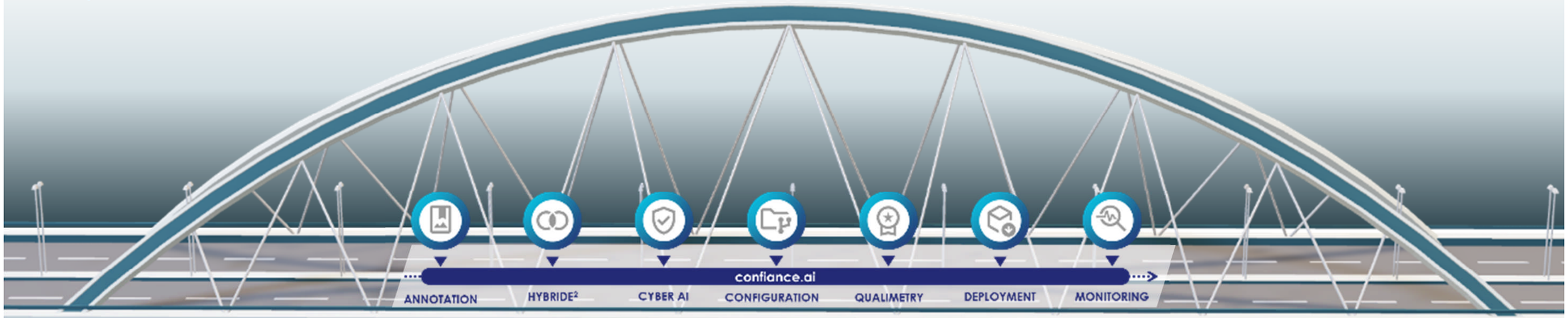
AI Main Technology Trends: Trustworthy and Responsible AI

Trustworthiness in AI-based critical system impacts the overall engineering lifecycle

From requirements & specifications
→ Stakeholder reqs.
→ Sys. & AI/ML specs.
→ Sys & AI/ML archis

MLOps/AIOps

To system deployment & maintenance
→ Monitoring
→ Toward qualification and certification



AI engineering: a necessarily condition to deploy trustworthy AI

Operational Design Domain (ODD)

Validation

Verification

Operational Analysis
regarding Intended Purpose

To ensure qualification and compliance with regulations (e.g. AI Act) **and standards** (e.g. ARP6983 in aeronautics)

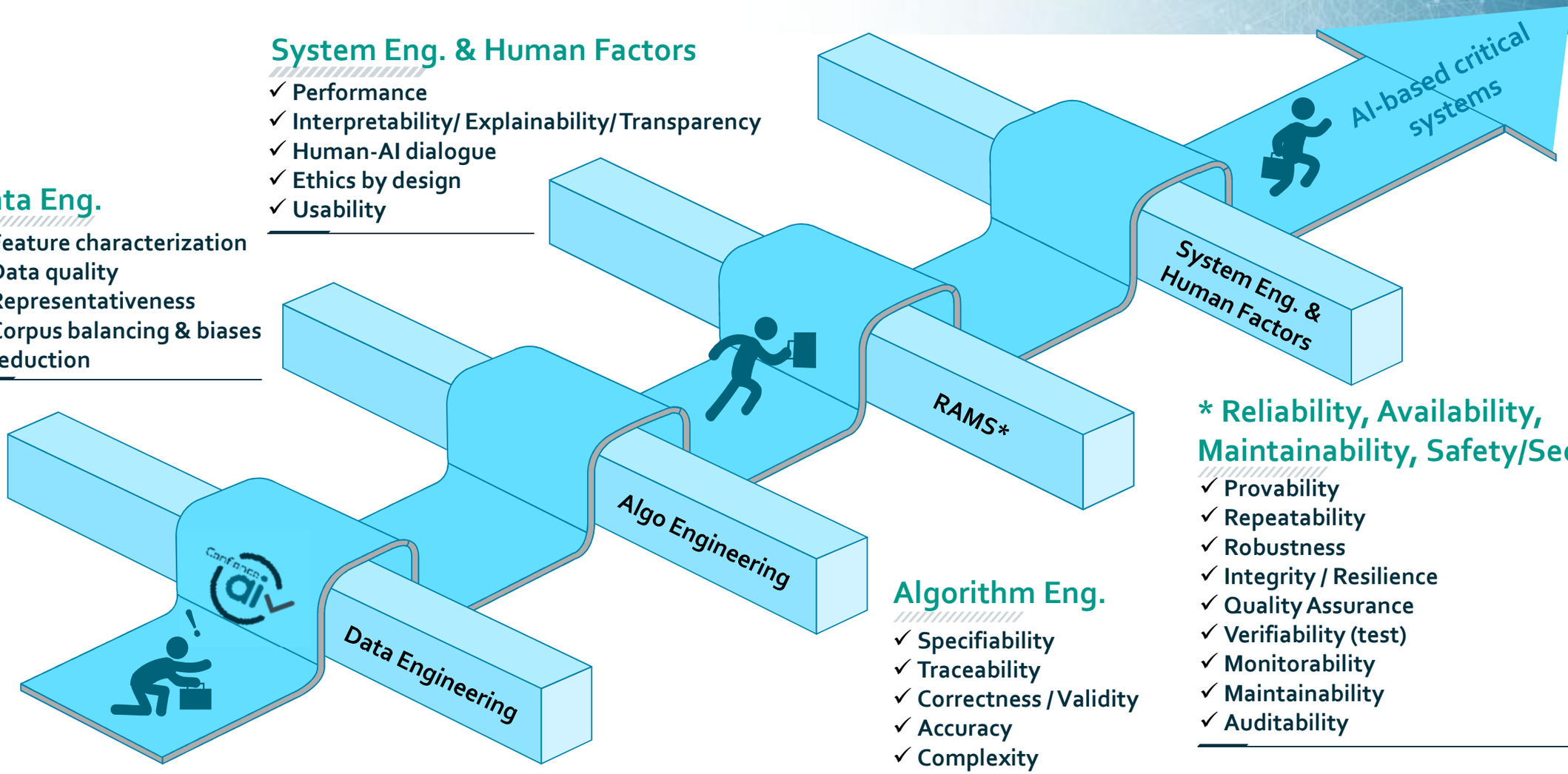
Why: ML deployment induces some (engineering) challenges…

System Eng. & Human Factors

- ✓ Performance
- ✓ Interpretability/ Explainability/ Transparency
- ✓ Human-AI dialogue
- ✓ Ethics by design
- ✓ Usability

Data Eng.

- ✓ Feature characterization
- ✓ Data quality
- ✓ Representativeness
- ✓ Corpus balancing & biases reduction



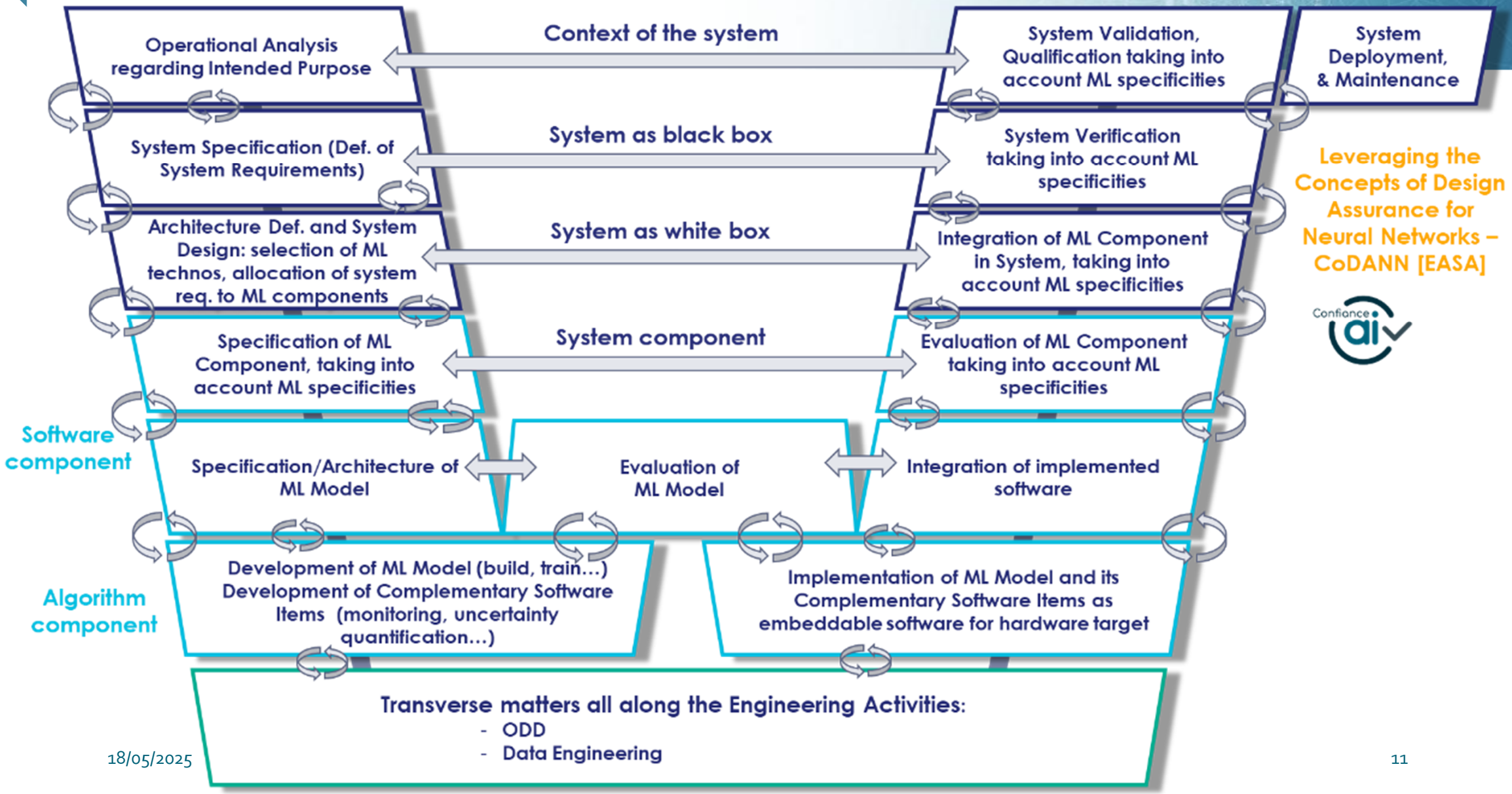
* Reliability, Availability, Maintainability, Safety/Security

- ✓ Provability
- ✓ Repeatability
- ✓ Robustness
- ✓ Integrity / Resilience
- ✓ Quality Assurance
- ✓ Verifiability (test)
- ✓ Monitorability
- ✓ Maintainability
- ✓ Auditability

Algorithm Eng.

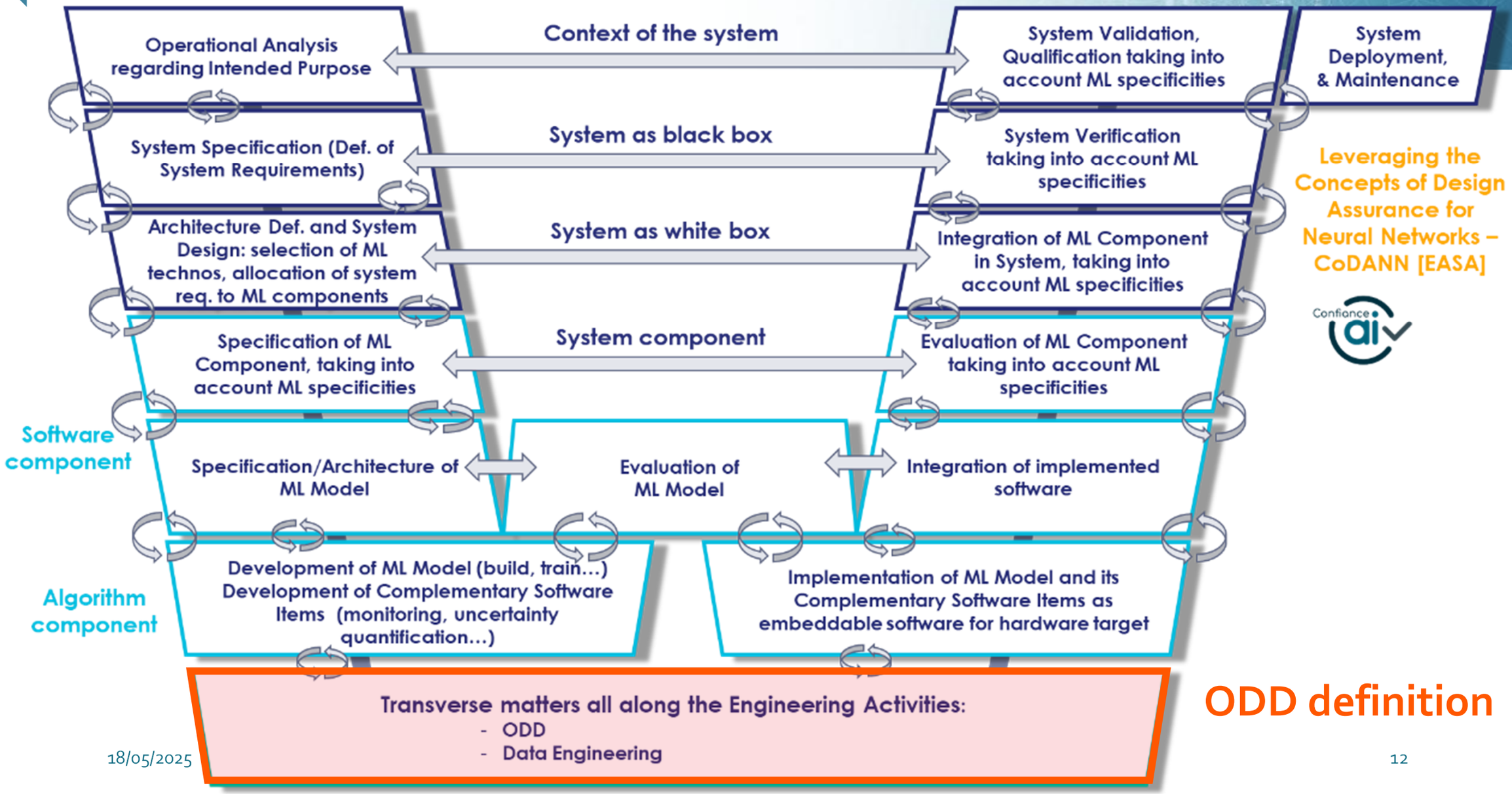
- ✓ Specifiability
- ✓ Traceability
- ✓ Correctness / Validity
- ✓ Accuracy
- ✓ Complexity
- ✓ Transparency
- ✓ Vulnerability mitigation (Robustness by design)

Systems/Software/Algorithm/Data Engineering lifecycle to design a ML-based System



Leveraging the Concepts of Design Assurance for Neural Networks – CoDANN [EASA]

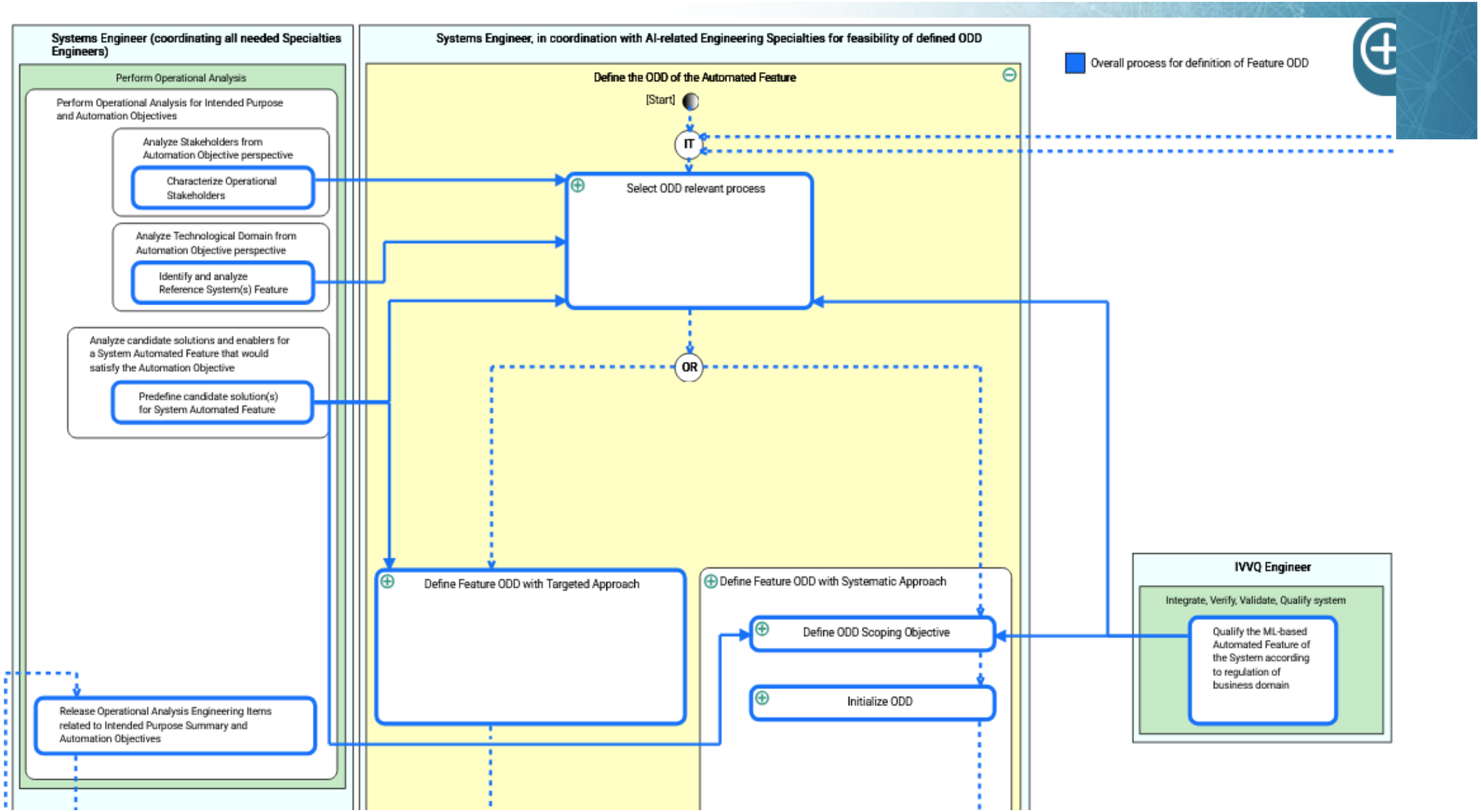
Systems/Software/Algorithm/Data Engineering lifecycle to design a ML-based System



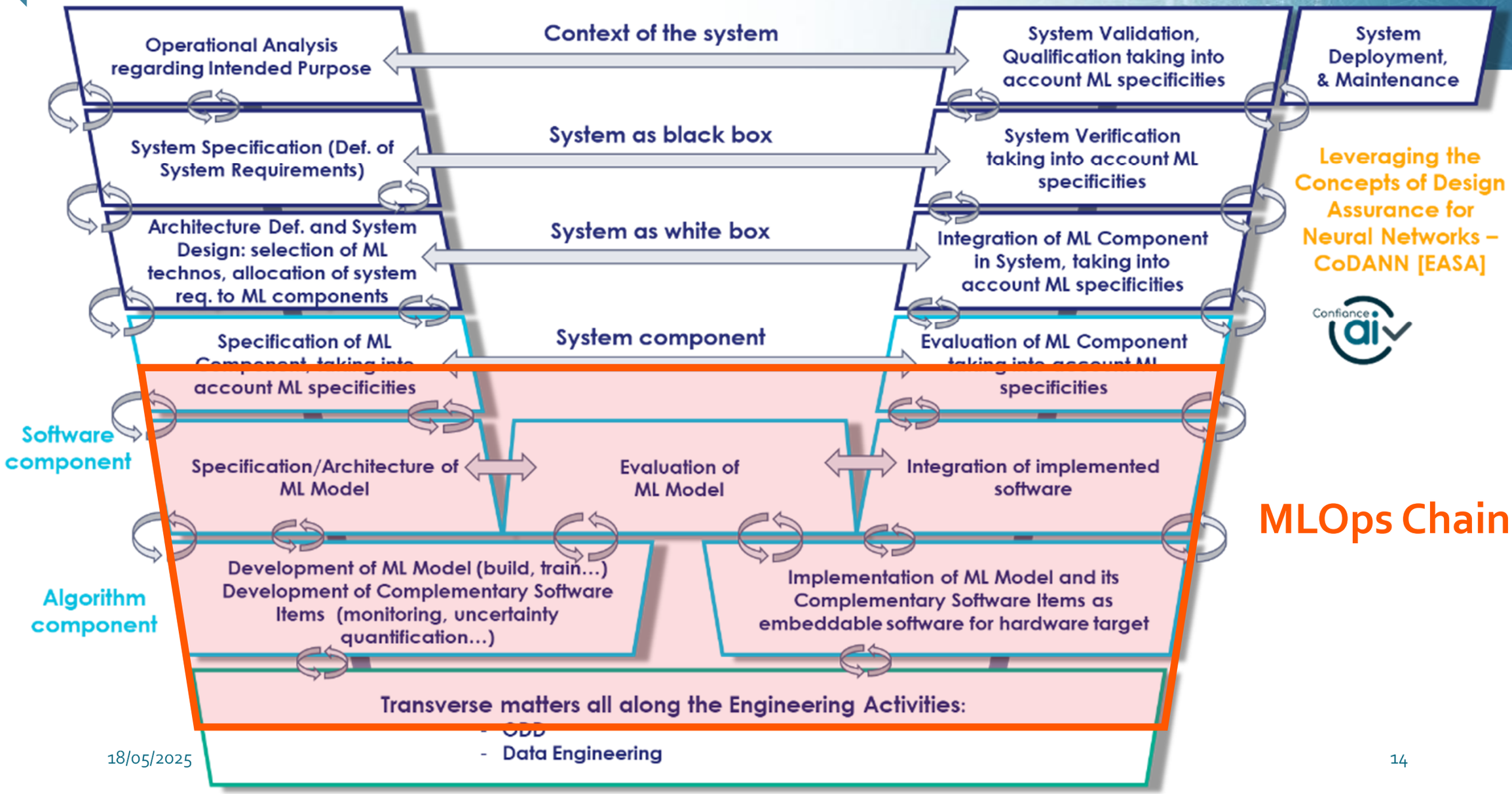
Leveraging the Concepts of Design Assurance for Neural Networks – CoDANN [EASA]

Confiance ai

ODD definition


Systems Engineer (coordinating all needed Specialties Engineers)
Perform Operational Analysis
Perform Operational Analysis for Intended Purpose and Automation Objectives
Analyze Stakeholders from Automation Objective perspective
Characterize Operational Stakeholders
Analyze Technological Domain from Automation Objective perspective
Identify and analyze Reference System(s) Feature
Analyze candidate solutions and enablers for a System Automated Feature that would satisfy the Automation Objective
Predefine candidate solution(s) for System Automated Feature
Release Operational Analysis Engineering Items related to Intended Purpose Summary and Automation Objectives
Systems Engineer, in coordination with AI-related Engineering Specialties for feasibility of defined ODD
Define the ODD of the Automated Feature
[Start]
IT
Select ODD relevant process
OR
Define Feature ODD with Targeted Approach
Define Feature ODD with Systematic Approach
Define ODD Scoping Objective
Initialize ODD
Overall process for definition of Feature ODD
IVVQ Engineer
Integrate, Verify, Validate, Qualify system
Qualify the ML-based Automated Feature of the System according to regulation of business domain

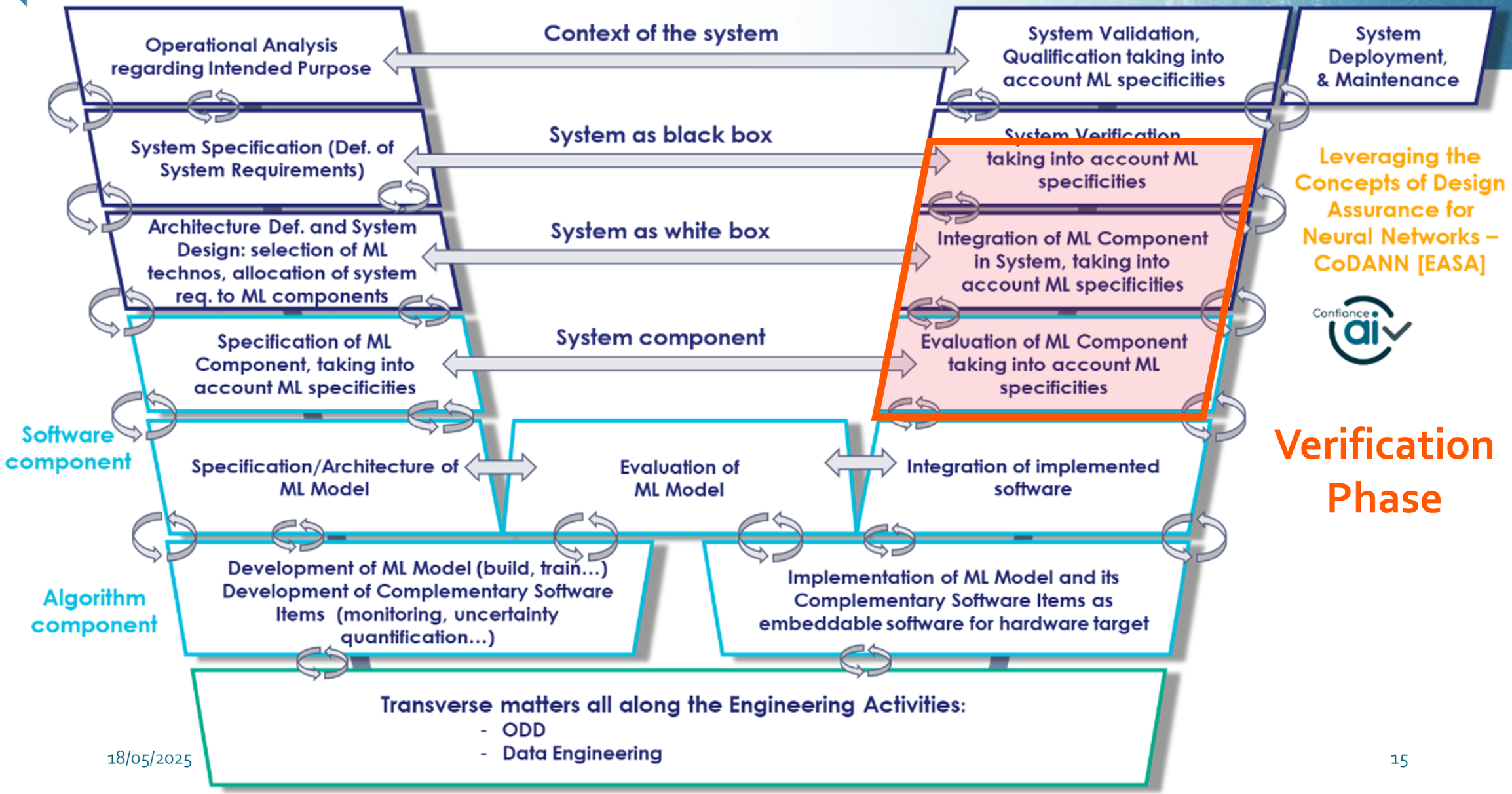
Systems/Software/Algorithm/Data Engineering lifecycle to design a ML-based System



Leveraging the Concepts of Design Assurance for Neural Networks – CoDANN [EASA]

MLOps Chain

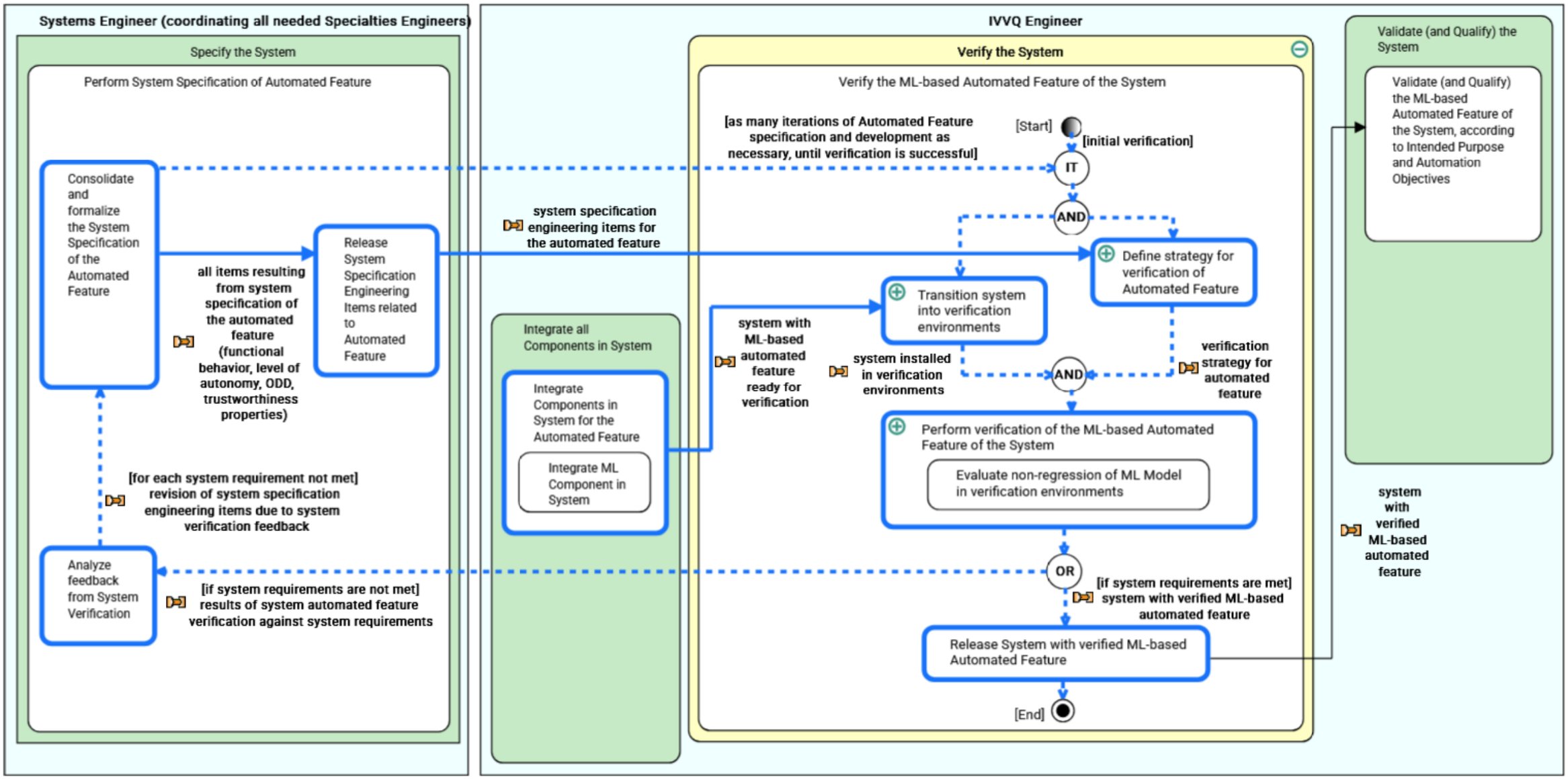
Systems/Software/Algorithm/Data Engineering lifecycle to design a ML-based System



Leveraging the Concepts of Design Assurance for Neural Networks – CoDANN [EASA]

Confiance ai

Verification Phase


Systems Engineer (coordinating all needed Specialties Engineers)
Specify the System
Perform System Specification of Automated Feature
Consolidate and formalize the System Specification of the Automated Feature
all items resulting from system specification of the automated feature (functional behavior, level of autonomy, ODD, trustworthiness properties)
Release System Specification Engineering Items related to Automated Feature
[for each system requirement not met] revision of system specification engineering items due to system verification feedback
Analyze feedback from System Verification
[if system requirements are not met] results of system automated feature verification against system requirements
system specification engineering items for the automated feature
IVVQ Engineer
Integrate all Components in System
Integrate Components in System for the Automated Feature
Integrate ML Component in System
Verify the System
Verify the ML-based Automated Feature of the System
[as many iterations of Automated Feature specification and development as necessary, until verification is successful]
[Start]
[initial verification]
IT
AND
Define strategy for verification of Automated Feature
Transition system into verification environments
system with ML-based automated feature ready for verification
system installed in verification environments
AND
verification strategy for automated feature
Perform verification of the ML-based Automated Feature of the System
Evaluate non-regression of ML Model in verification environments
OR
[if system requirements are met] system with verified ML-based automated feature
Release System with verified ML-based Automated Feature
[End]
Validate (and Qualify) the System
Validate (and Qualify) the ML-based Automated Feature of the System, according to Intended Purpose and Automation Objectives
system with verified ML-based automated feature

Body of Knowledge
Beta

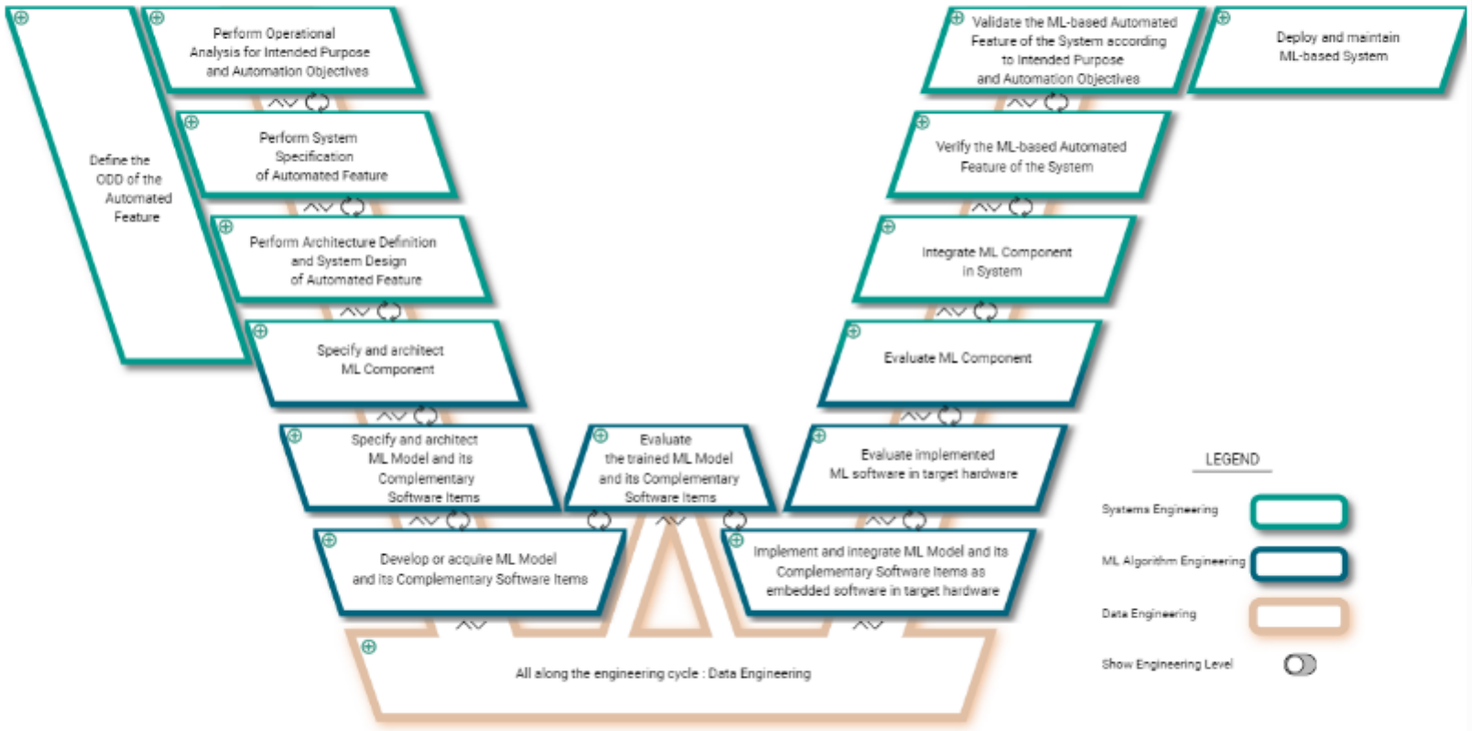
All Roles

Search by title

Activities for the engineering of a critical ML-based system

WELCOME MESSAGE AND TUTORIAL

Activities for the engineering of a critical ML-based system



The Confiance.ai research program revisited the classical engineering disciplines (Systems Engineering, Software Engineering...) in order to be able to handle correctly the integration of the ML technology in critical systems. The engineering processes necessary for ML-based critical systems have been formalized and are presented in this navigable "Body of Knowledge". Each page presents an engineering activity, with links to elements of Confiance.ai's catalog: software tools and documents that support the presented engineering activity.

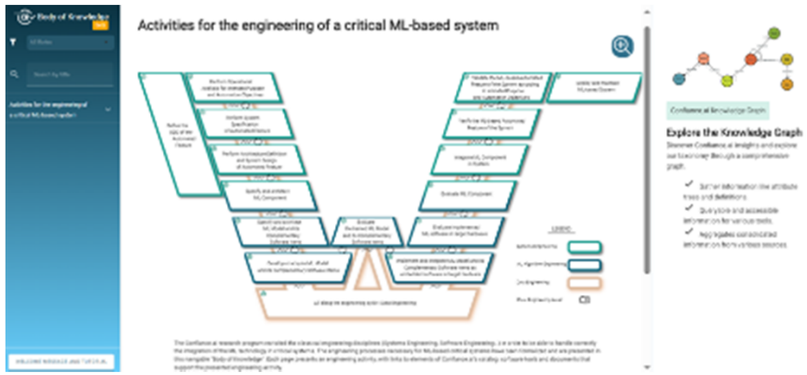
Confiance.ai Knowledge Graph

Explore the Knowledge Graph

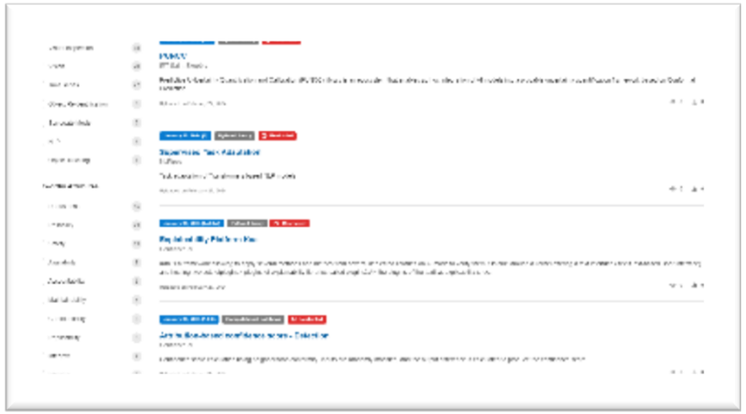
Discover Confiance.ai insights and explore our taxonomy through a comprehensive graph.

- ✓ Gather information like attribute trees and definitions.
- ✓ Queryable and accessible information for various tools.
- ✓ Aggregates consolidated information from various sources.

… integrated in a methodology supported by specific components and tools



The body of Knowledge is available at https://bok.confiance.ai/



The catalog is available at https://catalog.confiance.ai/

- More than 133 documents
  - 34 Methodological Guidelines Released
  - 28 State of the Art
  - 44 Benchmark or Application over use-case
- More than 72 components
  - 32 are open-source libraries or application
  - 26 are full Confiance.ai intellectual property
  - 5 are the property of a partner of Confiance.ai

The body of knowledge that reference these tools and method is now open to the community…

Questions & Answers