

Combining Data-Driven and Knowledge-Based Al Paradigms for Engineering AI-Based Safety-Critical Systems

Engineering AI-based critical system induces various challenges

Data & Knowledge Engineering

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

Algorithm Engineering

- ✓ Specifiability
- ✓ Traceability
- Correctness / Validity
- ✓ Accuracy
- ✓ Complexity
- ✓ Transparency
- Vulnerability

Human-Al Interaction

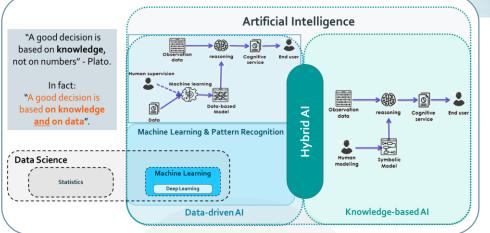
- ✓ Usability
- ✓ Interpretability / Explainability
- ✓ Human-Al dialogue
- Ethics by design

Safety & Cybersecurity

- ✓ Provability
- ✓ Verifiability (test)
- ✓ Robustness
- Integrity / Resilience

Software & System Engineering

- ✓ Repeatability
- ✓ Performance
- ✓ Maintainability
- ✓ Auditability
- Monitorability



Revisiting all engineering disciplines to propose a sound deployment of AI components within safety-critical systems

Formalize the various engineering roles / activities needed to specify / design /

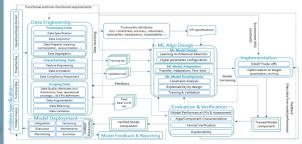
develop / deploy and maintain an AI-based safety critical system

Trust assessment def. framework

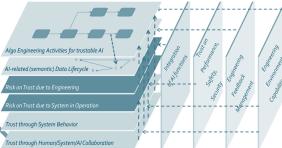
A 1st end-to-end framework to assess Trust through **Risk Mgmt & Assurance Case**

> Trustworthy attributes definition & associated KPIs

Trustworthy ML Alao Engineering



The Trustable AI Engineering Analysis Framework



Juliette MATTIOLI (Thales) Gabriel PEDROZA (CEA) Souhaiel KHALFAOUI (Valeo & IRT SystemX) Bertrand LEROY (Renault & IRT SystemX)

Identify roles and needs of the different stakeholders

Trustworthy AI

concepts,

attributes & KPIs

dimensions of trust in relation to the activities

Trust

def. & KPI

Define the

Risk assessment

management

Identify risks

induced by

engineering

activities and

means of

mitigation

Identify trust

properties of the

outputs of these

activities and

the means to

guarantee them

Generic IVVQ

Strategy

Assurance

Cases

Trustable AI Engineering Taxonomy