_MLEAP STAKEHOLDERS DAY

#

LAN

LNE - AIRBUS PROTECT - NUMALIS

EASA

European Union Aviation Safety Agency

Introduction of the Research project Machine Learning Application Approval MLEAP

PROTECT

23/11/2022

AIRBUS

LNE

numalis

Agenda of the day :

Introduction by the EASA and AIRBUS PROTECT

Introduction of the partners : LNE & NUMALIS

Presentation of the work plan : objectives, state of the art and expected output

Task 1 : Methods and tools for the assessment of completeness and representativeness of data sets (training, validation, and test) in data-driven ML and DL

Task 2: Methods and tools for quantification of generalization guarantees for ML and DL models

Task 3: Methods and tools for the verification of an ML algorithm and model robustness/stability

Task 4 : Communication

Conclusion & next steps

Q&A session:

The contractors & round tables > > >

Consortium members :

EASA

Willy Sigl, Xavier Henriquel, Guillaume Soudain, François Triboulet

LAN

LNE - AIRBUS PROTECT - NUMALIS

LNE

Olivier Galibert, Swen Ribeiro, Agnès Delaborde

Airbus Protect

Michel Kaczmarek, Thiziri Belkacem, Jean-Baptiste Rouffet, Jeremy Bascans**, Matthieu Rochambeau**

Numalis

Arnault Ioualalen, Noémie Rodriguez

EASA
European Union Aviation Safety Agency

MLEAP project
1st Stakeholders day

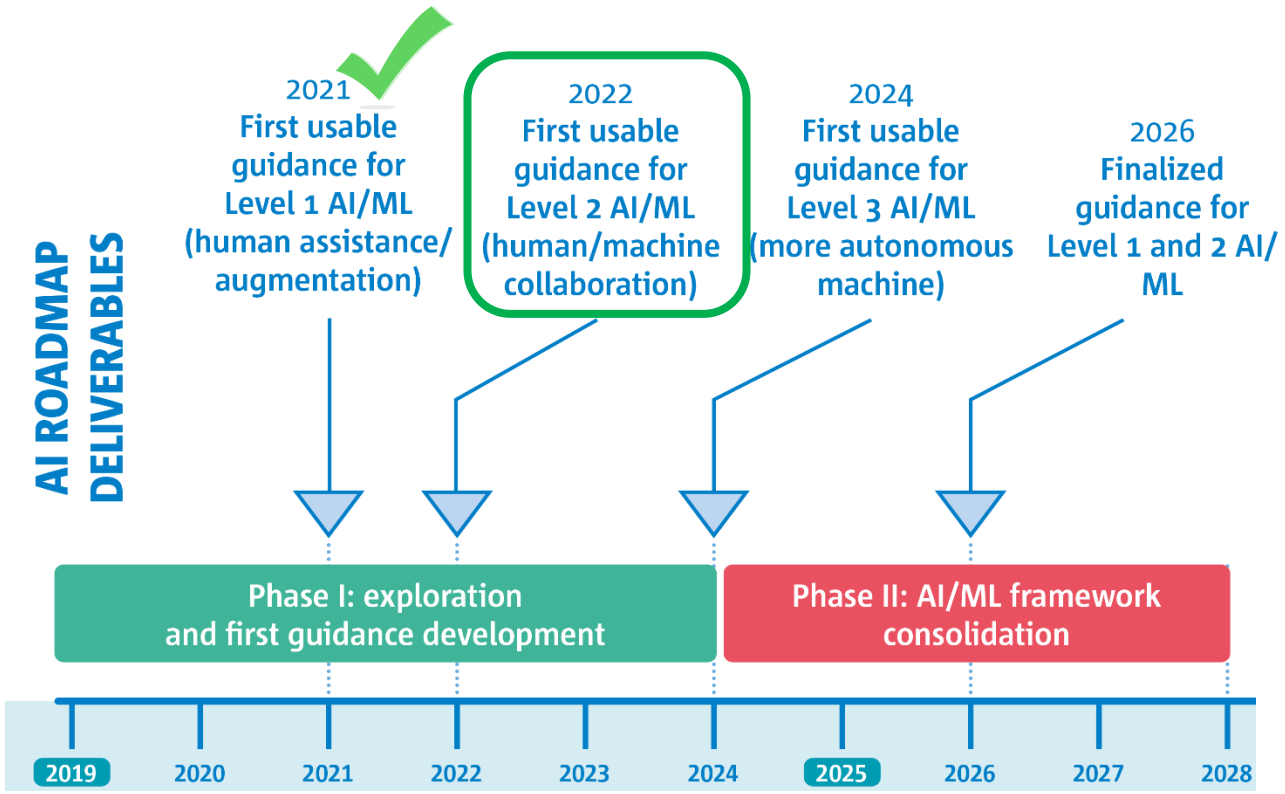
EASA-AI Roadmap

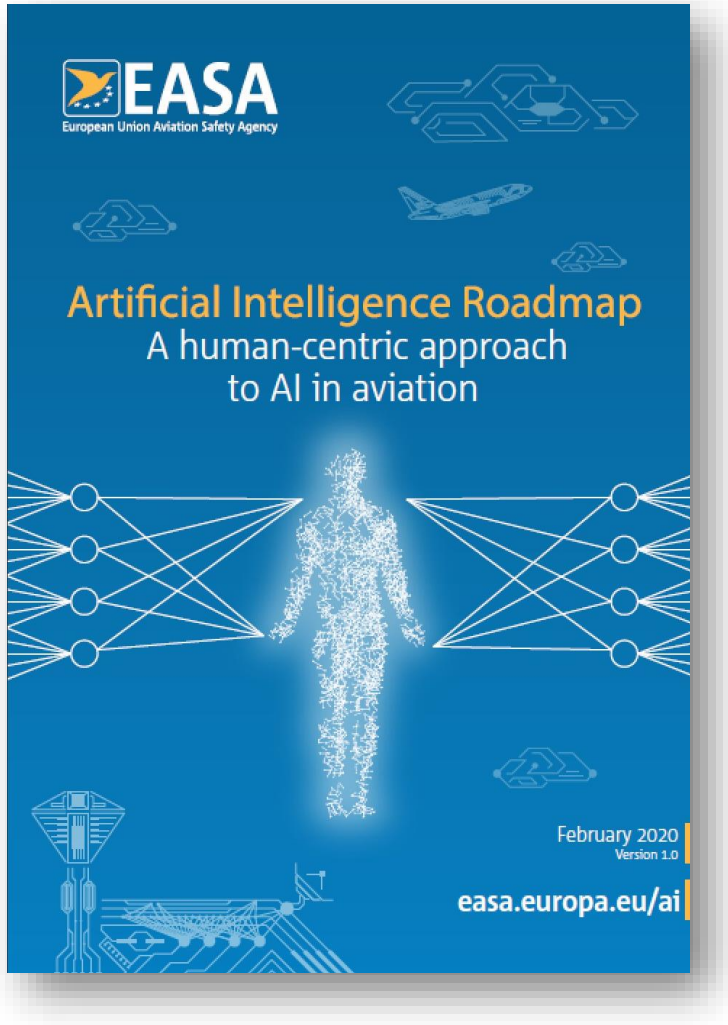
23rd November 2022

An Agency of the European Union

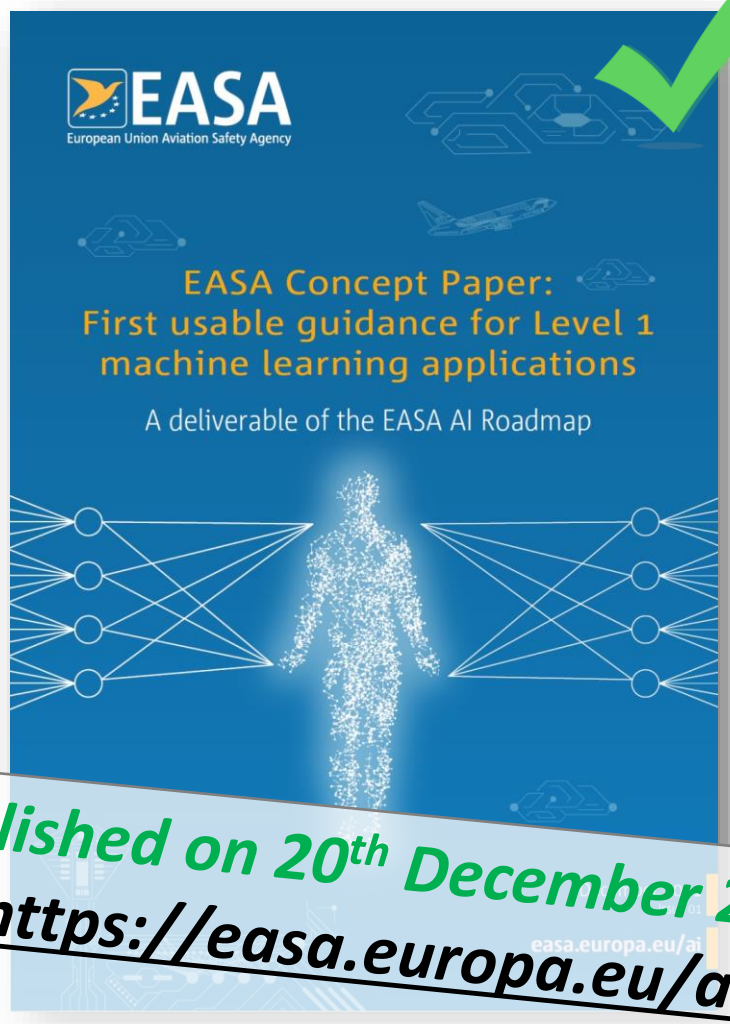
EASA AI Roadmap – Towards AI trustworthiness

- → Impact on all aviation domains
- → Common issue: level of trust in AI performance
- → « AI trustworthiness » concept is the key

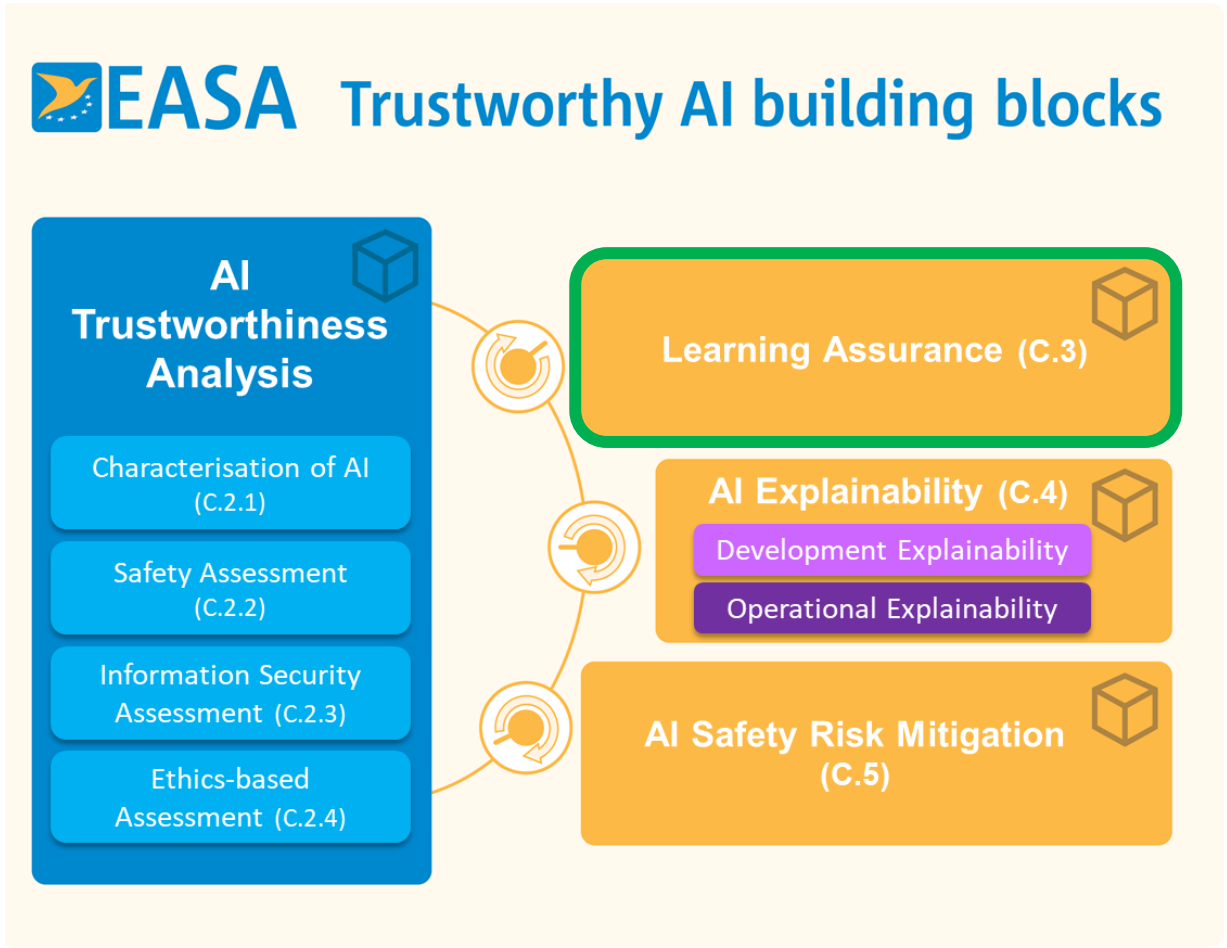




2021 - EASA guidance for Level 1 AI/ML* applications



Published on 20th December 2021
https://easa.europa.eu/ai



* AI/ML = Artificial Intelligence / Machine Learning

2022+ - TOP3 challenges for Level 2 AI/ML guidance

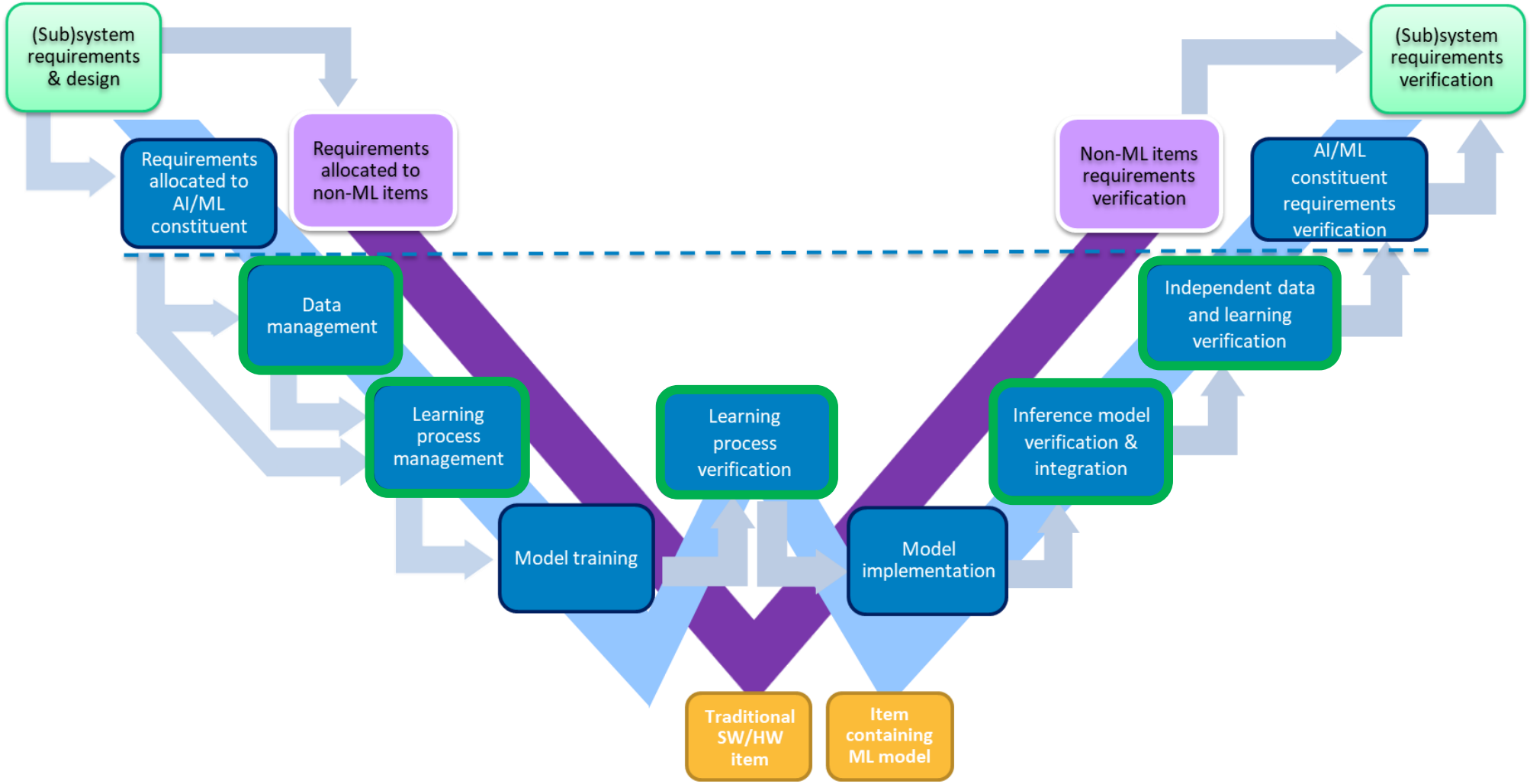
1. **Anticipate means of compliance for Learning Assurance objectives on ML Model guarantees (generalization and robustness)**
   - → Exploit the Horizon Europe Research project MLEAP on ‘Machine Learning applications approval’

2. **Operational explainability & human centric aspects of AI**
   - → Foster confidence in the system by developing specific HF guidance

3. **Ethics-based assessment – social & societal aspects**
   - → Evaluate and refine guidance based on use cases

Partnering on research projects is a key driver for the guidance!

W-shaped assurance process



Advanced MOCs for Learning Assurance

Machine Learning Application Approval (MLEAP) project

Objective

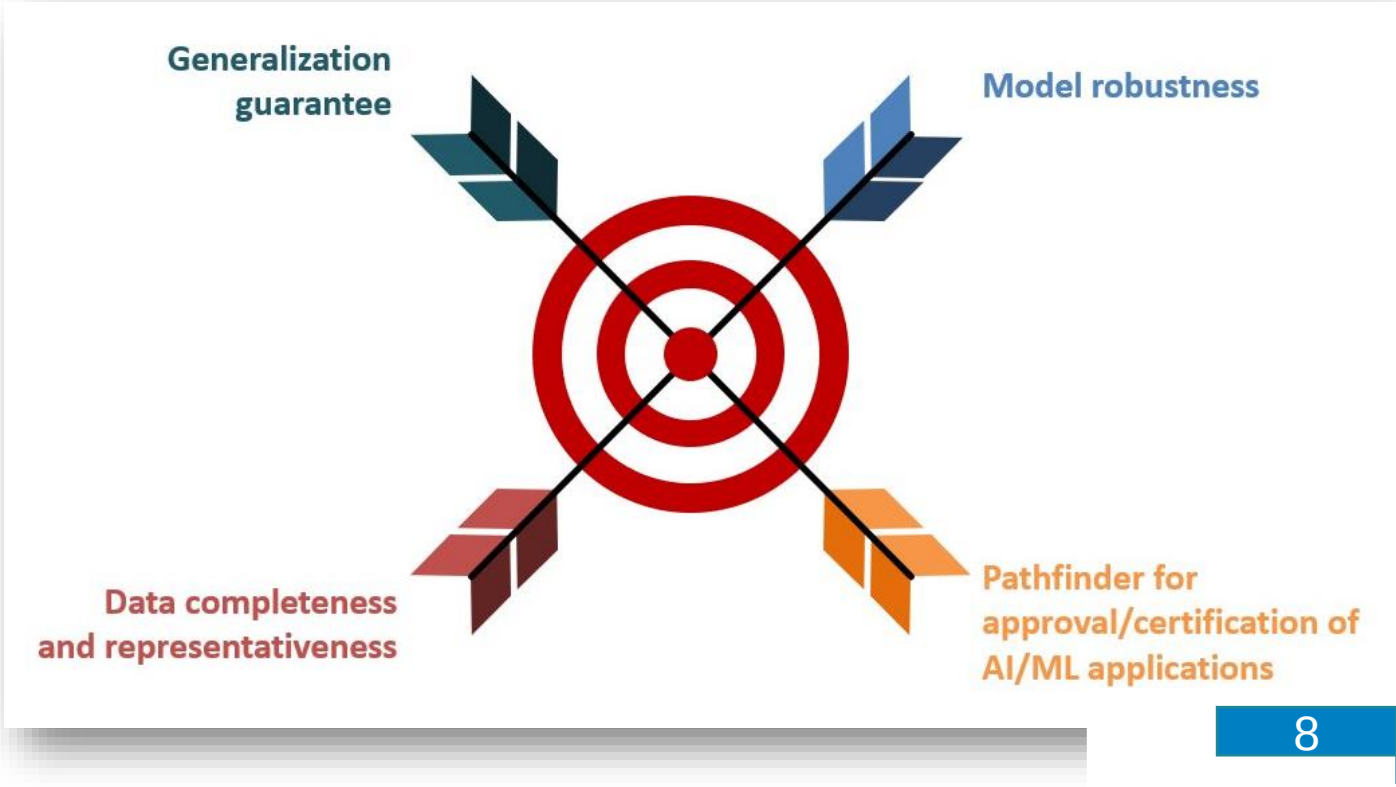
"[..]Streamline certification and approval processes by identifying concrete means of compliance with the learning assurance objectives of the EASA guidance for ML applications with a specific focus on Level 1B and Level 2 as defined in the EASA AI Roadmap[..]"

Budget

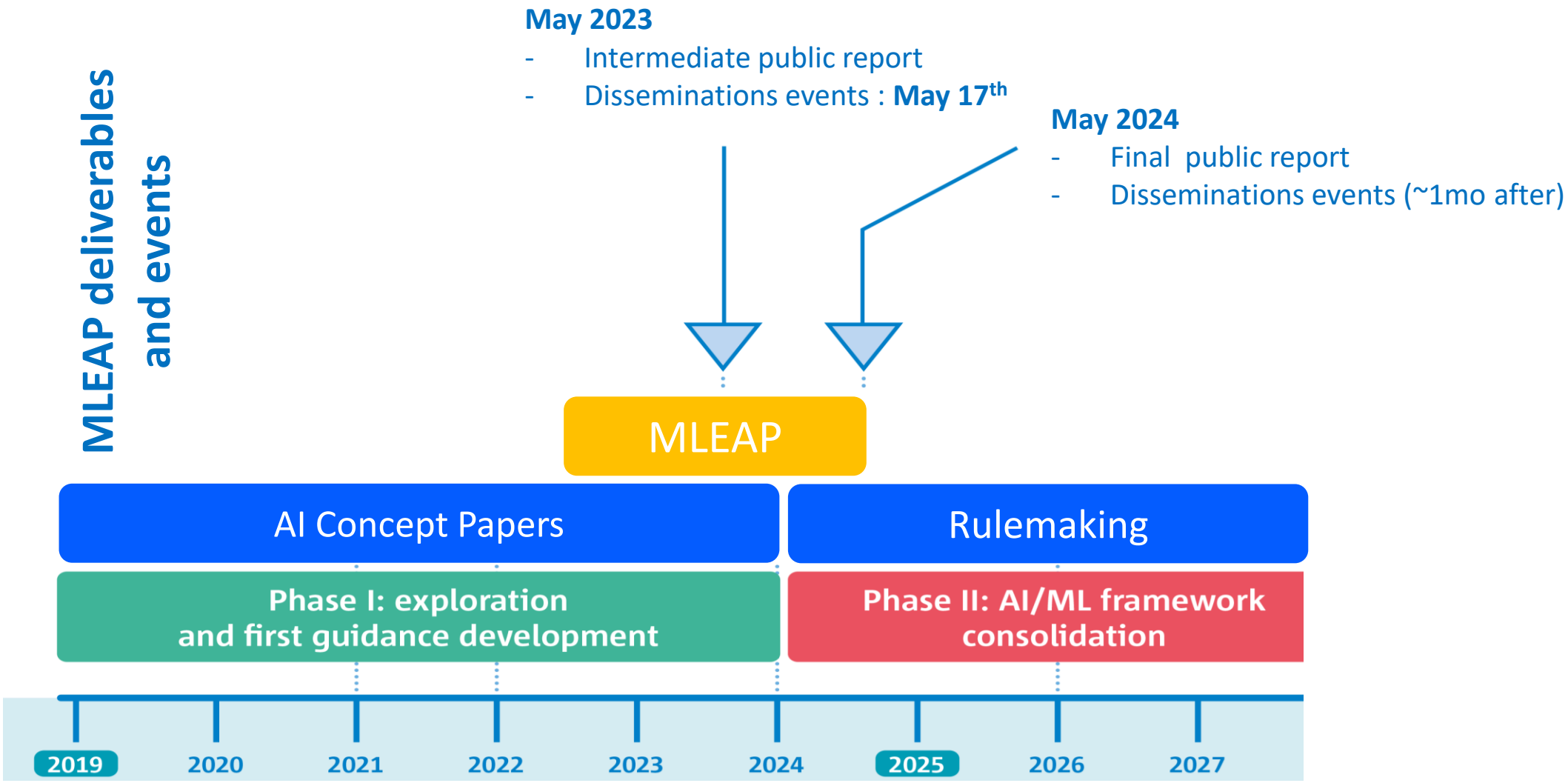
1.475 Million Euros funded by
Horizon Europe

Timeline

May 2022 - May 2024.
- Concurrent project with AI guidance
- Public deliverables



MLEAP timeline



: What we do

AIRBUS

PROTECT

Consulting

on Safety, Cybersecurity and Sustainability to optimise performance and support our customers on regulatory compliance and certification

Innovation

We are involved in research projects & member of institutional working groups

Involved in various R&T and software development projects in Artificial Intelligence:

DEEL project for IRT Saint Exupéry and ANITI
Confiance AI project
EPI project for IRT SYSTEMX (Consortium with STELLANTIS, NAVAL Group, EXPLEO, LIP6)
PRISSMA project for French Ministry of Transportation

Software

Specialised software supporting end-to-end safe mobility activities

Training

We are a recognised training organisation

Experts in

Industrial Risk Management

for over

35 years

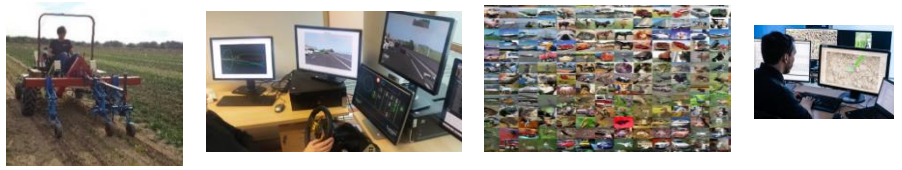
LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS

Founded in 1901
Appointed by French government on testing, certification and metrology for Industry (all sectors)

AI evaluation Department

- Development of evaluation standards
- AI systems testing
- Development of certification schemes
- Development of testbeds
- Professional training for industry

950+ systems evaluated in all major domains of AI and robotics since 2008



Development of softwares for AI evaluation and data preparation

www.lne.fr/logiciels/lne-matics

Certification for AI processes (2021)

NUMÉRIQUE LNE CERTIFICATION

https://www.lne.fr/en/service/certification/certification-processes-ai

LEIA 1/2/3: testbeds for AI and robotics (simulation, physical, hybrid)

LEIA 1 LEIA 2 LEIA 3

AIRBUS

Numalis, the no-guess company

- Formal methods for AI systems
- Markets: Aeronautic, Defence, aerospace, railway, health
- SaaS solution to
  - Measure robustness
  - Explain behavior
  - Prepare compliance of IA
- 20 persons, Montpellier

On-going projects:

- HE MLEAP with EASA
- 2 EDIDP (Defence)
- ESA…

Software:
- AI Robustness
- AI Explainability
- Formal analysis
- Trustworthy AI

Standardization:
- ISO/IEC standard editor on AI robustness
- Contributor to many other projects

Services:
- Standardization ecosystem
- Validation process
- AI Audit

MLEAP Project Architecture > > >

Task #1 : Data Completeness and Representativeness

Task #2 : Model generalization

Task #3 : Algorithm and model robustness

Task #4 : Communication

MLEAP – Task #1 milestones: Data Completeness and Representativeness

Completeness : *A data set is complete if it sufficiently covers the entire space of the operational design domain for the intended application*.

Representativeness : *A data set is representative when the distribution of its key characteristics is similar to the actual input state space for the intended application*

MLEAP – Task #1 milestones: Data Completeness and Representativeness

> > >

Assessment of the criteria influencing the selection of methods and tools for the assessment of completeness and representativeness of data sets (e.g. learning technique, nature and dimensionality of the data, etc.).

Identification or development of efficient and practicable methods and tools for the assessment of completeness and representativeness of data sets (training, validation and test) in the generic case of data-driven ML.

Demonstration of effectivity and practicability of the identified methods on real-scale aviation use case(s).

MLEAP – Task #1 Technical Feedback > > >

- **Influence factors identified:**

Technical requirements
- Intended function
- Model architecture
- Data dimensionality
- Intended level of autonomy
- Intended level of performance
- Intended level of robustness and resilience
- Intended level of stability

Processes
- Data Management requirements (specs)
- Data Quality improvement (augmentation…)
- Data synthesis
- Data sampling
- Labeling
- Pre-processing

Other DQRs
- Balance
- Relevance
- Diversity (discriminative power)
- Diversity (absence of non representative sampling bias)
- Currentness

MLEAP – Task #1
Technical Feedback > > >

80+ sources explored, among which 60+ assessment methods analysed

20 methods selected for testing

Sufficient maturity

In line with the project objectives

Technical requirements
- Intended function
- Model architecture
- Data dimensionality
- Intended level of autonomy
- Intended level of performance
- Intended level of robustness and resilience
- Intended level of stability

6 methods selected (from 11 identified)

Processes
- **Data Management requirements** (2 methods)
- **Data Quality improvement** (3 methods)
- **Data synthesis** (1 method)
- **Data sampling** (1 method)
- **Labeling** (2 methods)
- Pre-processing

11 methods selected (from 33 identified)

Other DQRs
- **Balance** (1 method)
- Relevance
- Diversity (discriminative power)
- **Diversity (absence of bias)** (1 method)
- **Currentness** (1 method)

3 methods selected (from 18 identified)

MLEAP – Task #1 Technical Feedback > > >

Main take-aways

Assessment of data quality in general lacks maturity in the field of AI:

< 10 works are explicitly considering influence factors in their relationship to Completeness/Representativeness
Influence factors and target properties are not studied in a structured way

Exhaustive data quality of the data set may be hard/impossible to attain:

Operations required to enhance data quality attributes may be mutually exclusive (e.g. ensuring relevance can be detrimental to representativeness)
Importance of expert contextual trade-off

MLEAP – Task #1 Technical Feedback > > >

Main take-aways

In literature, the burden of sorting the wheat from the chaff often still rests on the model.

No "off-the-shelf" method to quantify the relationship between a factor of influence and Completeness/Representativeness.

High-dimensionality challenges rarely addressed. Adaptability of the methods to high-dimensional data needs to be explored.

MLEAP – Task #1 Technical Feedback > > >

Next steps:

Exploratory work on the selected methods
Refinement of the selection grid
Application to the project's use cases

MLEAP – Task #2 Milestones
Model Generalization

Comprehensive Overview

Available methods and tools to evaluate generalization bounds;
Barriers in generalization guarantees for a given model: ML and DL;

State-of-the-art analysis and generic evaluation approach proposal

Limitation of available methods and common practices;

Definition of a generic approach for a more effective evaluation;

Further:

Identification or development of efficient methods and tools for the quantification of generalization guarantees in the generic case of data-driven ML

MLEAP – Task #2 Milestones : Model generalization > > >

Generalizability

Definition

Model's ability to generalize the learned knowledge to a new context or environment

Success indicator

Good performances (w.r.t. some criteria) for *Dtest ≠ Dtrain*

Failure indicators

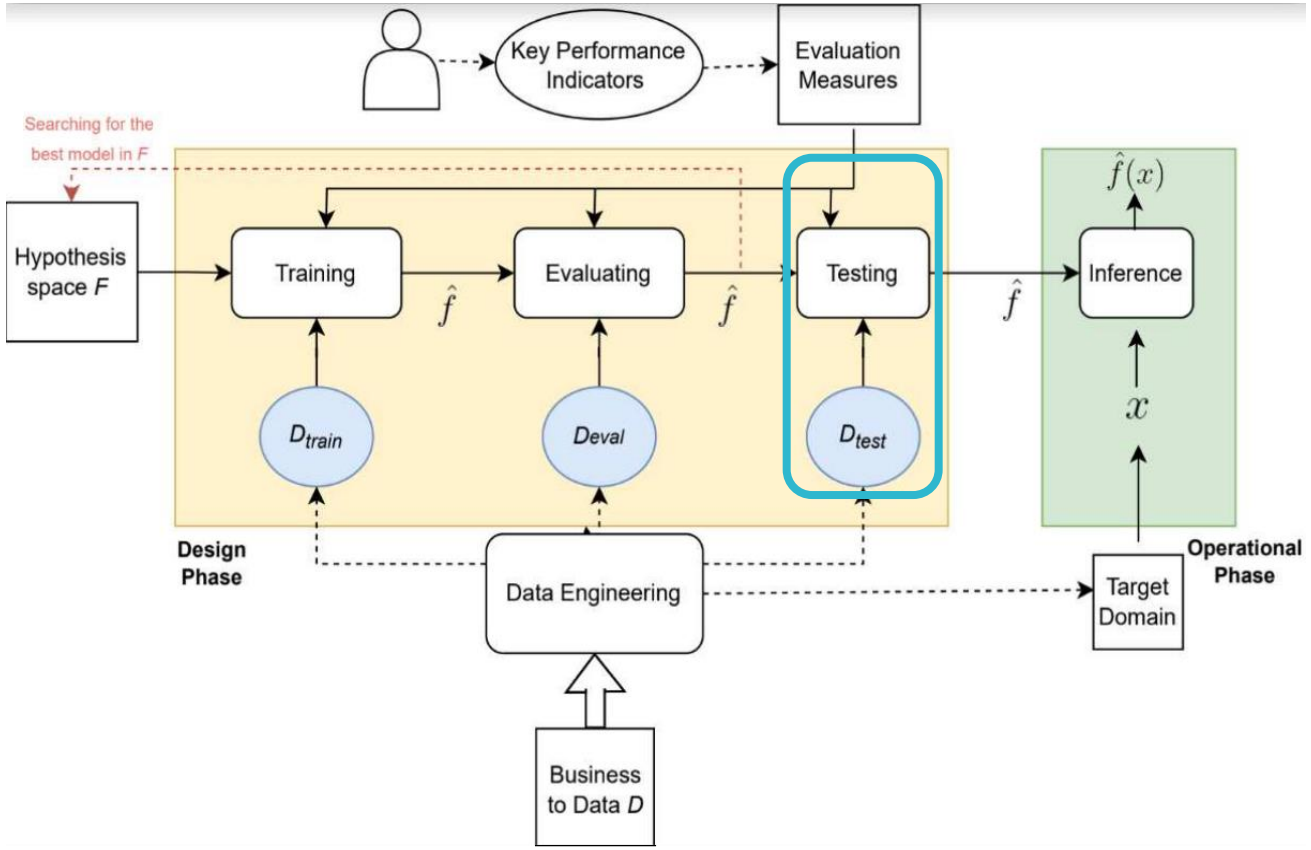
- *Overfitting (e.g.* big model, few data*)*

$$R_{D_{test}}(f) > \hat{R}_{D_{train}}(f)$$

$$R_{D_{test}}(f) = \frac{1}{m}\sum_{j=1}^{m} l(f, x_j)$$

- *Underfitting (e.g.* small capacity, complex task*)*

$$C_f^i < C_f \qquad C_f = sup_D\, I(f \mid D_{train})$$



AIRBUS

MLEAP – Task #2 Milestones : Model generalization > > >

Generalization Bounds

Definition *(CoDANN-2020 → 13 references)*

Statistical tools that take as input various measurements of a model $\hat{f}$ on training data, and output a performance estimate for unseen data *D*

$$G(\hat{f}, D) \leq \sqrt{\frac{func(model\ class\ F\ complexity) + \log(1/\delta)}{\|D_{train}\|}}$$

$$G(\hat{f}, D_{train}) \to 0 \quad as \quad \|D_{train}\| \to \infty$$

Generalization in ML

Based on data set characteristics (e.g. convex hull)

Based on model characteristics (e.g. complexity)

Generalization in DL

Based on NN size, Norms and margins

Uniform stability of the learning algorithm

The theoretical bounds cannot be applied easily due to the over-parameterized setting

Furthermore: Domain generalization practices

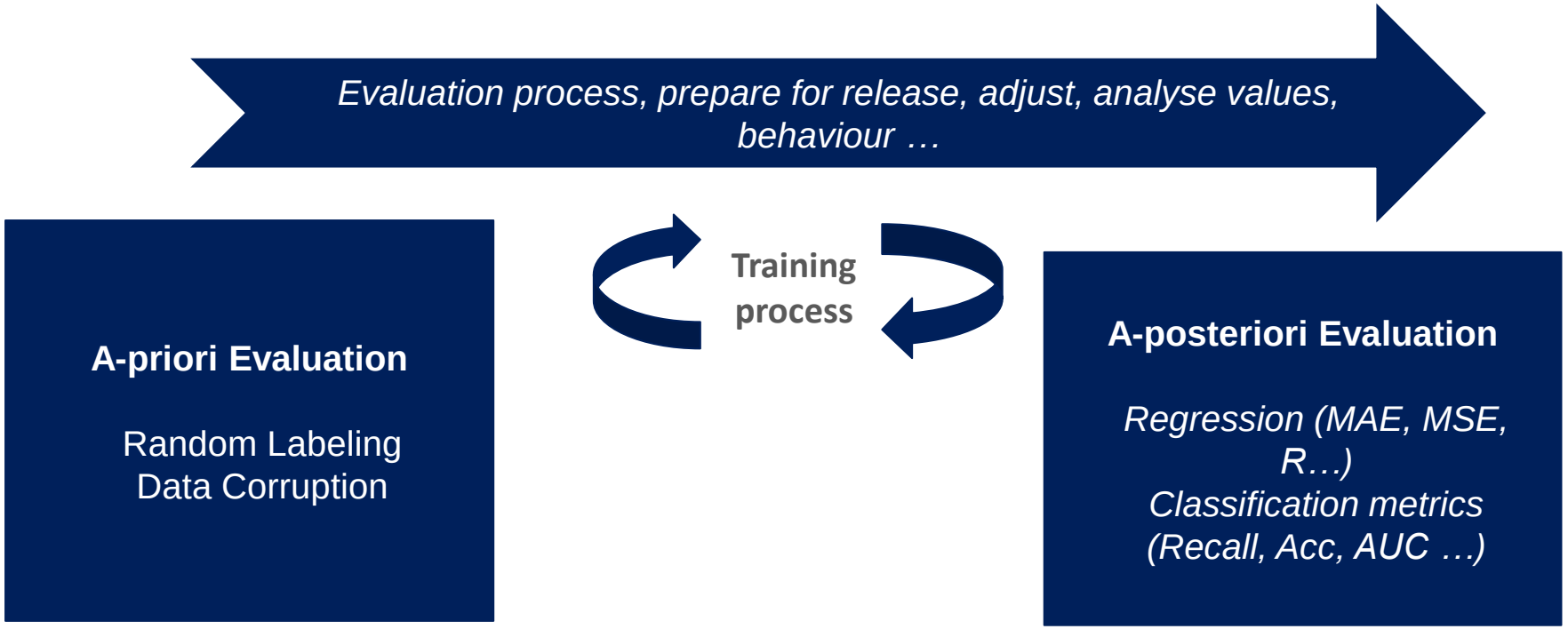
		Algorithm Dependent	
		Yes	**No**
Data Dependent	**Yes**	• PAC-Bayesian • PAC-Bayesian bounds for NNs (+) more precise, better distributional properties of the learning algorithm	• Rademacher Complexity (RC) • RC and regularized Empirical Risk Minimization (ERM) (+) better estimation
	No	• Model Compression • Based on Model Distillation (-) do not take into account data features (+) focuses on the model enhancement	• VC-dimension • VC-dimension for NNs (-) Not practical for particular use-cases (Dar et al., 2021) (+) widely applicable

		• Statistical guarantees
		  ○ Data statistics
		  ○ Error gradient during training
		• Geometry analysis bounds (combining input, output spaces and the mapping)	

MLEAP – Task #2 Milestones : Model generalization > > >

Methods to Evaluate Generalization

After the model is designed, how good it would be?



"Capture the underlying correlations between input and output space"

- Parametric and sensitivity analysis;
- Combination of several statistical parameters;

-> make sure that the system's behavior is environment independent

AIRBUS

MLEAP – Task #2 Milestones : Model generalization > > >

Methods to Boost Generalization

Regularizations

- Data driven (e.g. batch normalization)
- Model driven (e.g. activation functions)
- Based on the training objective (e.g. make a metric part of the objective function)
- Based on optimization (e.g. initialization, warm-up, pre-training)
- Using regularizers and/or combine the above

Penalty Methods

- Select only useful features (e.g. dimensionality reduction)
- Minimize weights of less important features (weight decay)

Model Reduction

- Network reduction:
  - training
  - pruning (reduce complexity)
  - fine-tuning (recover lost performances)
- Truncation (similar to dropout using position-based scores)

Data Expansion

- Data quality (e.g. completeness) assessment;
- Data volume (using model-based and task-based heuristics or empirical studies) qualification;
- Data augmentation
  - Learnable methods
  - Non-learnable methods

MLEAP – Task #2 Milestones : Model generalization > > >

Generic Approach Proposal
Issues and limitations of existing methods

Misunderstanding of the generalization bounds

- Some norm-based measures negatively correlate with generalization
- Conventional bounds based on uniform convergence or uniform stability are inadequate for over-parameterized models

Common mistakes and pitfalls in practice

- Inappropriate training objective
- Inappropriate data representation, volume, split (train, test, valid), quality (noisy, high sparsity)
- Inappropriate model complexity to perform the task, and evaluation metrics

Gap between expectations from evaluation vs the real-world application

- How far away the empirical assessment reflect the reality about the model efficiency?
- Appropriate performance indicators to the application domain cannot ALWAYS be translated by existing evaluation metrics
- How to define a good model ? what constitutes a good AI/ML model?
- What about the uncertainty tolerance: how the 85% accuracy is good? how the 15% uncertainty is tolerable ?

MLEAP – Task #2 Milestones : Model generalization > > >

Generic Approach Proposal

Objectives

Bridge the gap between experimentation and industrial expectation

- Adopt a multicriteria/additional validation phases;
- Include KPIs (industrial target performance) in the learning objectives and the evaluation metrics as well

Better handle the OOD samples and reduce the impact on the safety of the AI system

Build an enhanced data and model development pipelines reducing the impact of common practices and pitfalls that result in a weak generalization ability of an ML/DL model

MLEAP Project: Next Steps > > >

(1) Data evaluation and qualified (<=> Task#1)

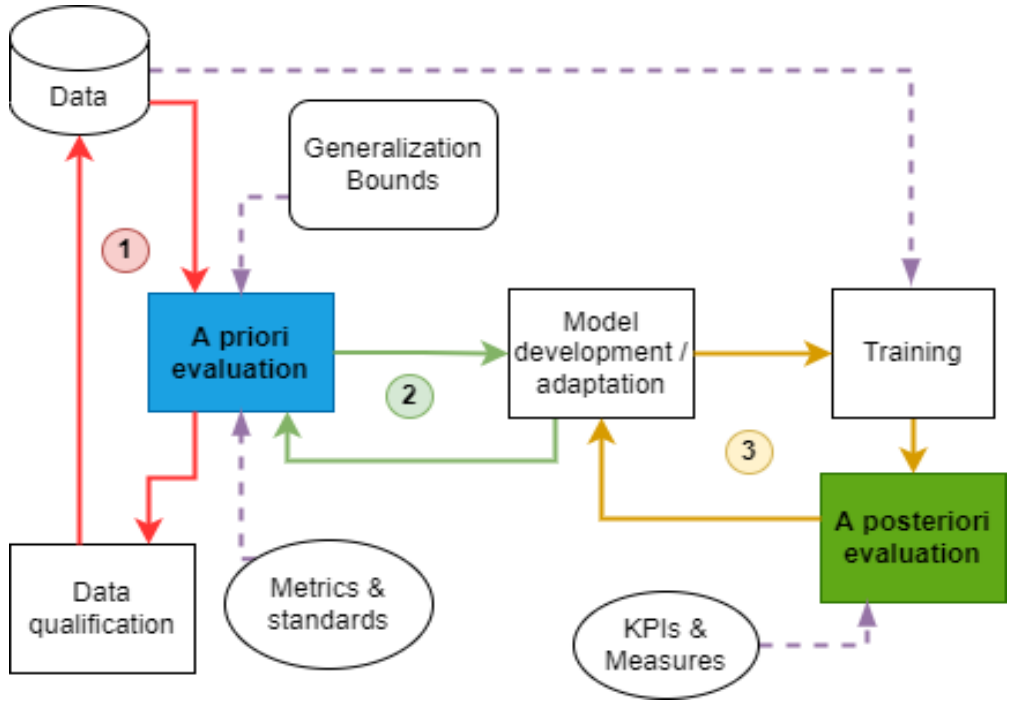
a. Minimal size of data set needed
b. Data quality evaluation (completeness, representativeness)
c. Enhancement operations: data augmentation, processing, cleansing, balancing, and splitting;

(2) Model development and adaptation

a. Data Constraints: data size and type, alignment …
b. The mappings between the inputs and outputs
c. Generalization bounds ;

(3) Model training on the optimized data set

a. Benchmark including a set of industrial KPIs
b. Adapted evaluation measures/metrics/thresholds
c. A posteriori evaluation of the trained model (<=> Task#3): generalization & robustness
d. Measures and loss functions should be adapted to meet the target application objectives



Further: ***Evaluation in three different use cases: STT-ATC, ACAS Xu, AVI***

AIRBUS

MLEAP – Task #3 Milestones
Algorithm and model robustness

- Review of methods and tools
- Identification of corner cases and abnormal inputs
- Identification of sources of instabilities during the design phase
- Identification of sources of instabilities during the operational phase
- Demonstration on a use-case

MLEAP – Task #3 Progress > > >

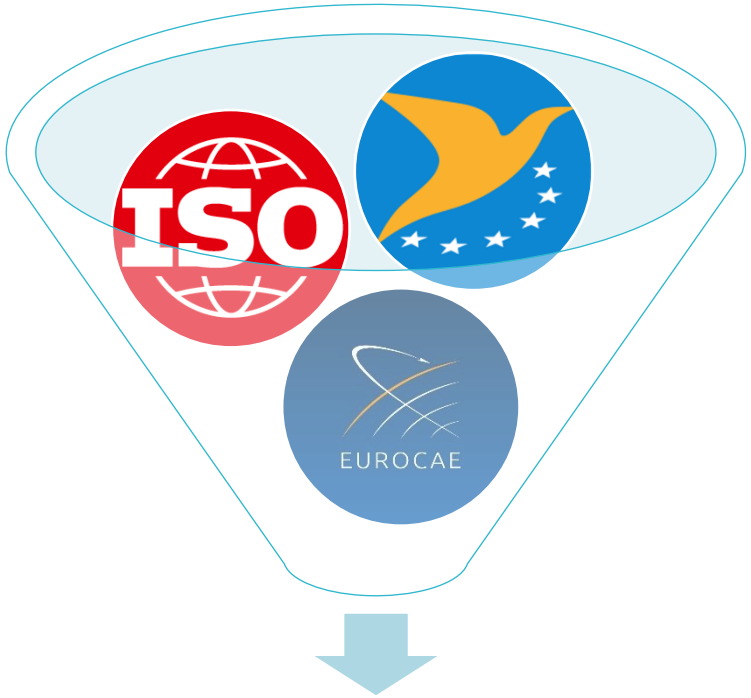
Aligning several sources of the state of the art

- Different concepts robustness, stability, edge cases…
- Different requirements
- Different methods: statistical, formal, empirical

Studying the maturity of the ecosystem

- Applicability to the relevant use-cases
- Scalability of the methods

Preparing the content of the validation tool box



Harmonized state of the art
D3.1 and D3.2

AIRBUS

MLEAP – Task #3 Progress > > >

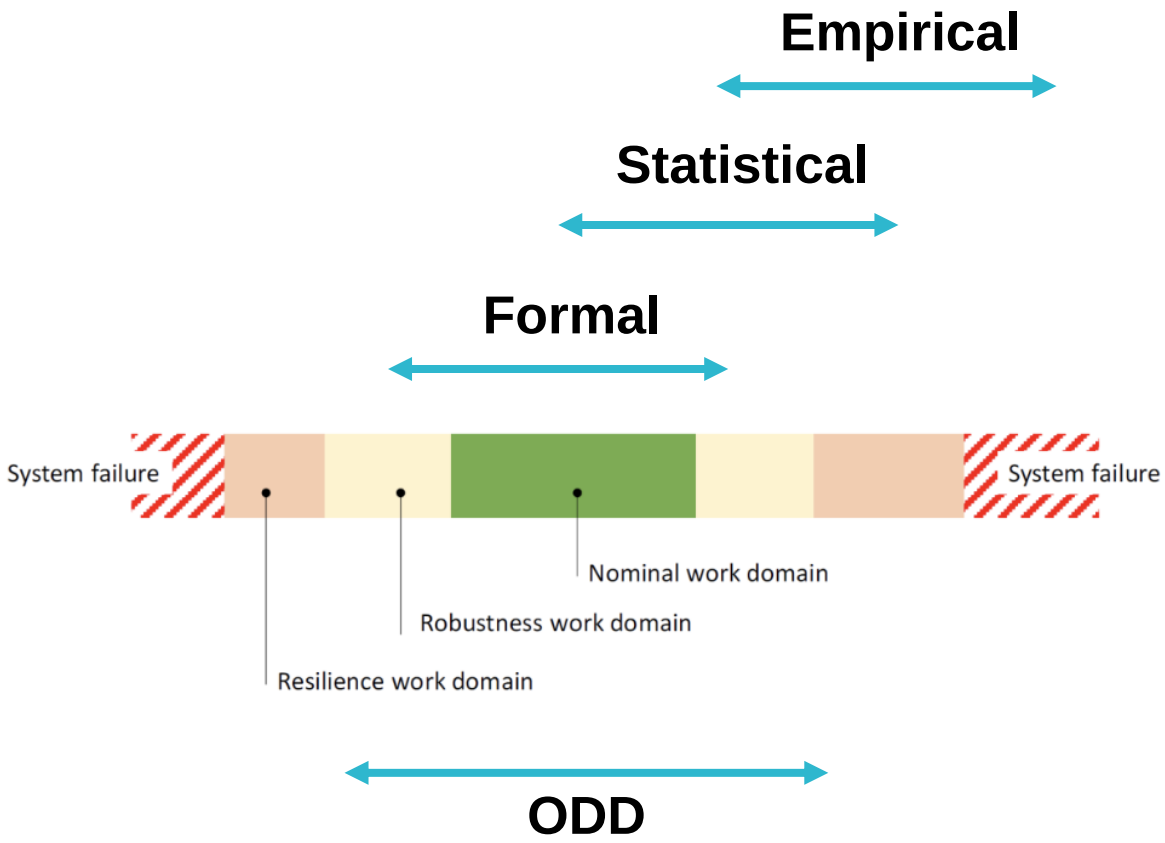
	Empirical methods	Statistical methods	Formal methods
Stability of the training algorithm			
Stability of the trained model			
Stability of the inference model			
Bias			
Variance			
Relevance			
Reachability			

Scalability	Human intervention needed	Doable but through sampling	Doable but locally
Methods	Field trial A posteriori Benchmarking	Combining metrics	Solver Abstract interpretation Optimization

MLEAP – Task #3 Progress > > >

Conceptual alignment is possible
Stability around the nominal conditions
Robustness to more difficult conditions
Resilience to adverse conditions

Methods are complementary
Depends on the ODD description
Combining approaches to match the requirements
…but varying degree of scalability

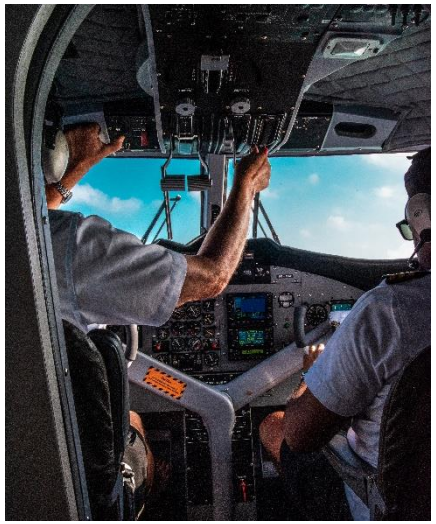


Picture from "DEEL White Paper on Machine learning in Certified System (DEEL Certification Workgroup, 2021"

MLEAP Project: Next steps > > >

Perform a comparative evaluation of methods and tools
to assess their efficiency and make recommendations for possible means of compliance

3 use cases will support the evaluation of this methodology



Speech to text
STT - ATC

Collision avoidance
ACAS - Xu

Automated visual inspection
AVI

AIRBUS

MLEAP – Task #4 Communication

Covers all 3 tasks
To keep you informed on the progress being made !
Share the latest information

Social media platforms
Upcoming events: 4 events over the 2 years of this project

Websites

https://www.lne.fr/fr https://www.protect.airbus.com/ https://numalis.com/

https://www.easa.europa.eu/en/research-projects/machine-learning-application-approval

Q&A

AIRBUS

LNE

numalis

Thank you !

AIRBUS

LNE

numalis