Airbus Protect Artificial Intelligence Conferences & networking >>

Awareness Session conference/ MLEAP & Beyond

WG-114 Köln Plenary

Paving the way for the future of Artificial Intelligence in Aviation

EASA European Union Aviation Safety Agency

LNE Laboratoire National de Métrologie et d'Essais

numalis

MLEAP project: [Machine Learning Application Approval]

June 30th, 2023

AIRBUS PROTECT

Disclaimer

MLEAP Project is funded by the European Union.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Union Aviation Safety Agency (EASA). Neither the European Union nor EASA can be held responsible for them.

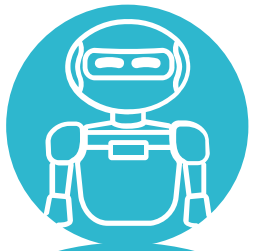
This presentation has been carried out for EASA by an external organisation and expresses the opinion of the organisation undertaking this deliverable. It is provided for information purposes. Consequently it should not be relied upon as a statement, as any form of warranty, representation, undertaking, contractual, or other commitment binding in law upon the EASA. Ownership of all copyright and other intellectual property rights in this material including any documentation, data and technical information, remains vested to the European Union Aviation Safety Agency.

Agenda

- **Introduction of the EASA AI program & MLEAP project**
- **Short presentation of the objectives and progresses & limits of Task 1** *(data management)*
- **Short presentation of the objectives and progresses & limits of Task 2** *(generalisation guarantees)*
- **Short presentation of the objectives and progresses & limits of Task 3** (robustness guarantees)
- **Conclusions & What's next after MLEAP?**

Who we are > > >

Consortium members :



MLEAP Team

EASA

Willy Sigl, Xavier Henriquel, **Guillaume Soudain, François Triboulet**

LNE

Olivier Galibert, Swen Ribeiro, Agnes Delaborde, Sabrina Lecadre

Airbus Protect

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

Numalis

Arnault Ioualalen, Noémie Rodriguez

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)

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

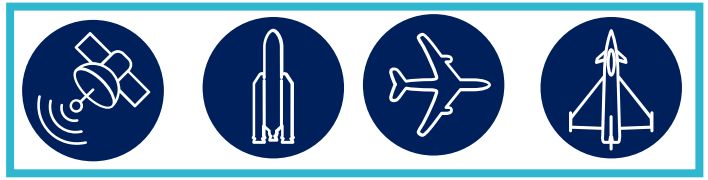
LEIA 1/2/3: testbeds for AI and robotics (simulation, physical, hybrid)

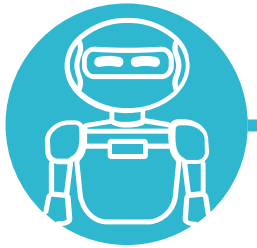

Numalis, the no-guess company

- *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

/ Airbus Protect an {Airbus} company

: What we do

Consulting

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

Training

We are a recognised training organisation

Software

Specialised software supporting end-to-end safe mobility activities

Innovation

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

bringing together outstanding expertise in safety, cybersecurity and sustainability we created a European leader in risk management

… delivering consulting, services & solutions

R&T & software development projects in AI:

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

WG114 Köln Plenary – Awareness Session conference

Introductory notes from EASA technical team

Guillaume Soudain
EASA AI Programme Manager
MLEAP Project Sponsor

Xavier Henriquel
EASA Safety Expert
MLEAP EASA Tech Lead

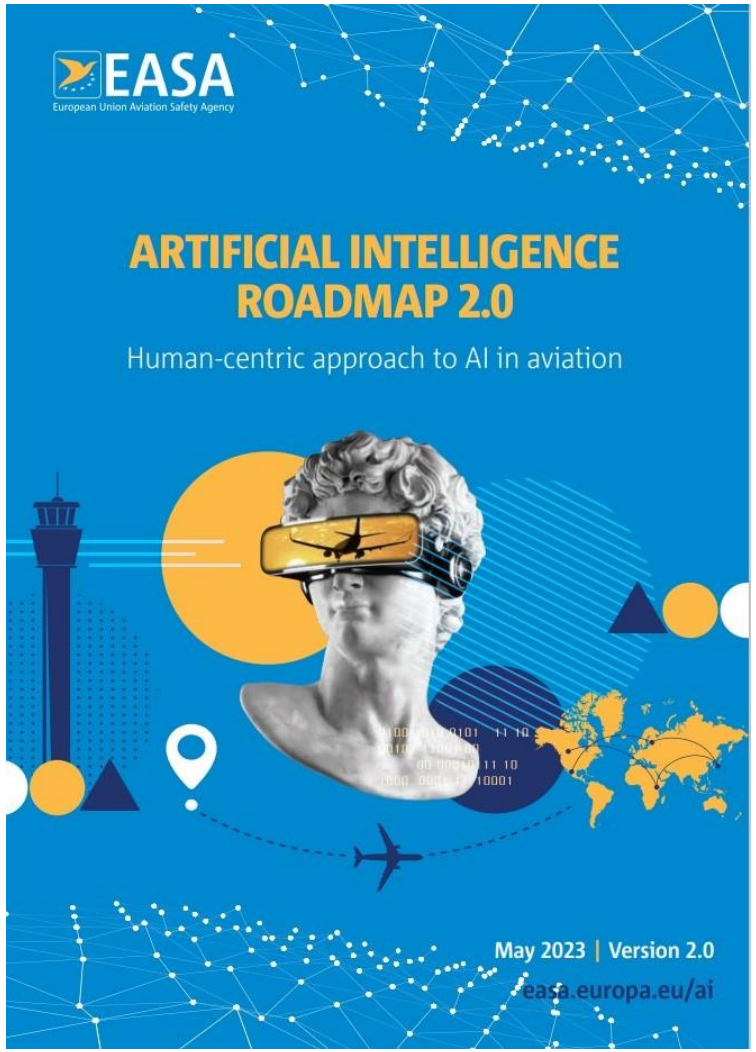
François Triboulet
EASA ATM/ANS Expert
Coordinator

EASA AI Roadmap – Towards AI trustworthiness

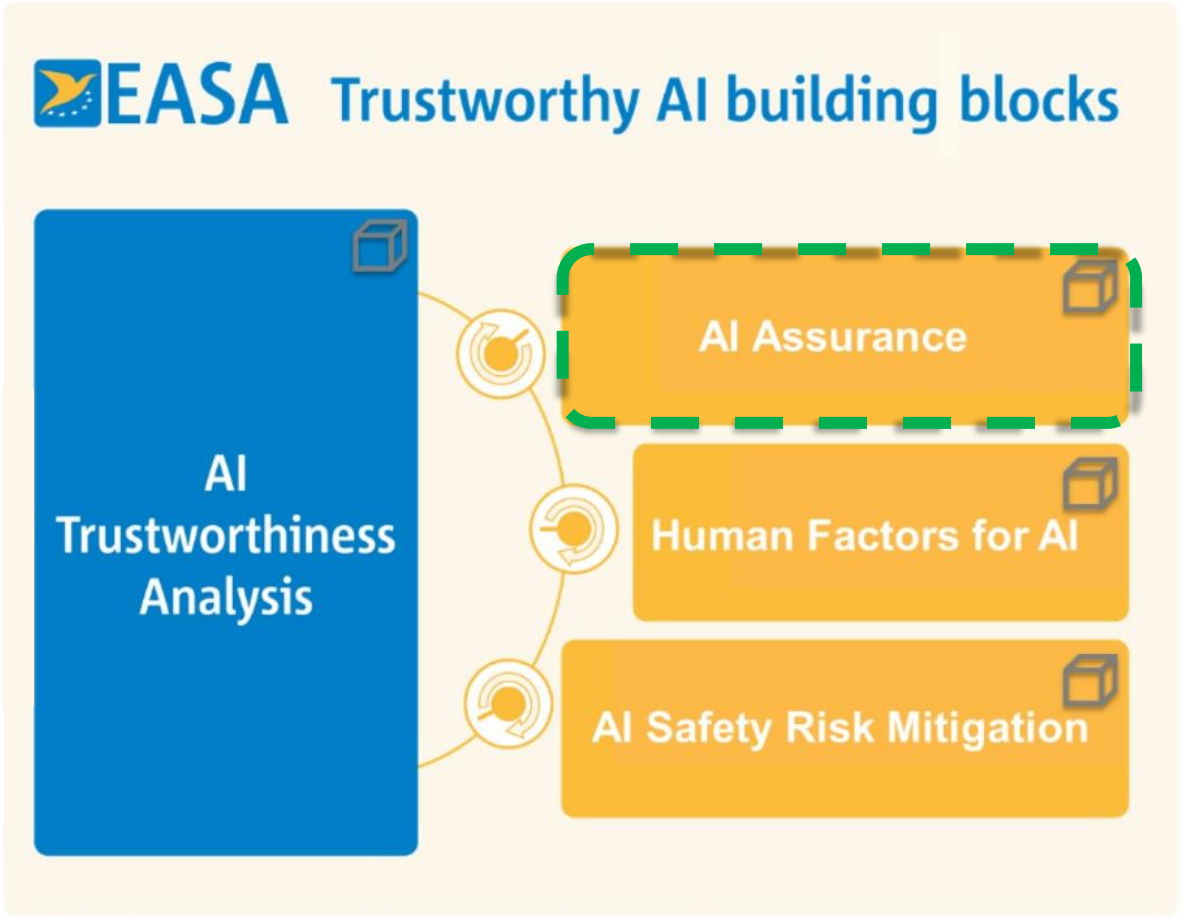
→ Impact on all aviation domains

→ Common issues for safety-related applications

→ « AI trustworthiness » concept is the key!



EASA guidance for Level 1 & 2 ML* applications



EASA

* ML = Machine Learning

TOP3 challenges for Level 1&2 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' and ForMuLA IPC with Collins Aerospace ART

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

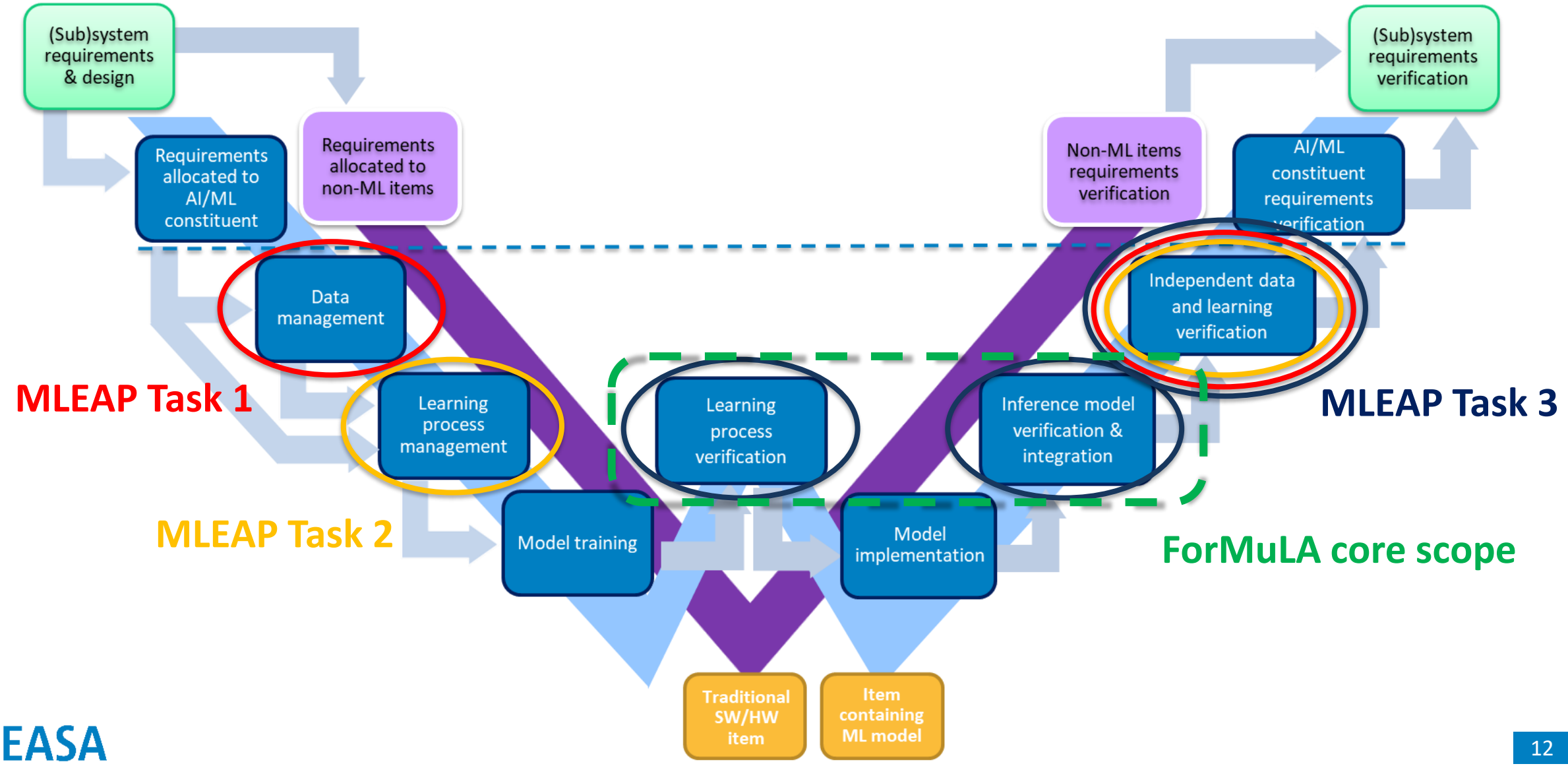
2. **Operational explainability & human centric aspects of AI**

   →Foster trust in the human-AI teaming by developing specific Human Factors guidance.

3. **Ethics-based assessment – social & societal aspects**

   →Evaluate and refine guidance based on use cases

W-shaped Learning Assurance concept



Machine Learning Application Approval (MLEAP) project

Objectives

*"Streamline certification and approval processes by **identifying concrete means of compliance with** the learning assurance objectives of the **EASA guidance for ML applications***

Budget

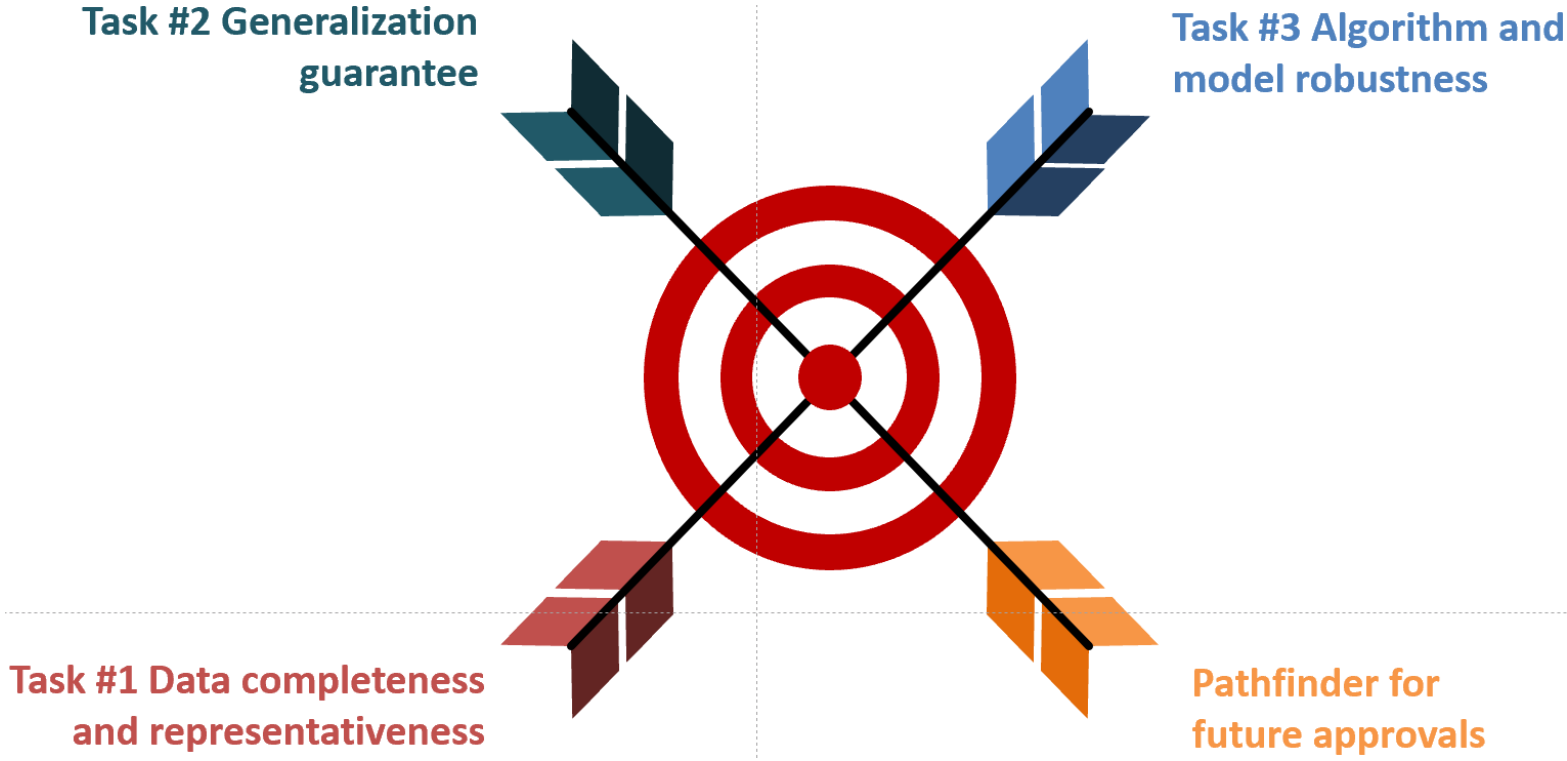
1.475 Million Euros
funded by EU Horizon Europe

Timeline

May 2022 - May 2024

Research consortium

Airbus Protect - LNE - Numalis



EASA

MLEAP project milestones

MLEAP deliverables and events

May-June 2023

- **First public report** & **Exec Summary**
- Stakeholders day #2 & Dissemination events
  - "EASA AI days 2023" - 17th May 2023
  - "Paris Air Show 2023" - 21st June 2023
  - "*SG34&WG114 Köln Plenary*" - **30th June 2023**

April-May 2024

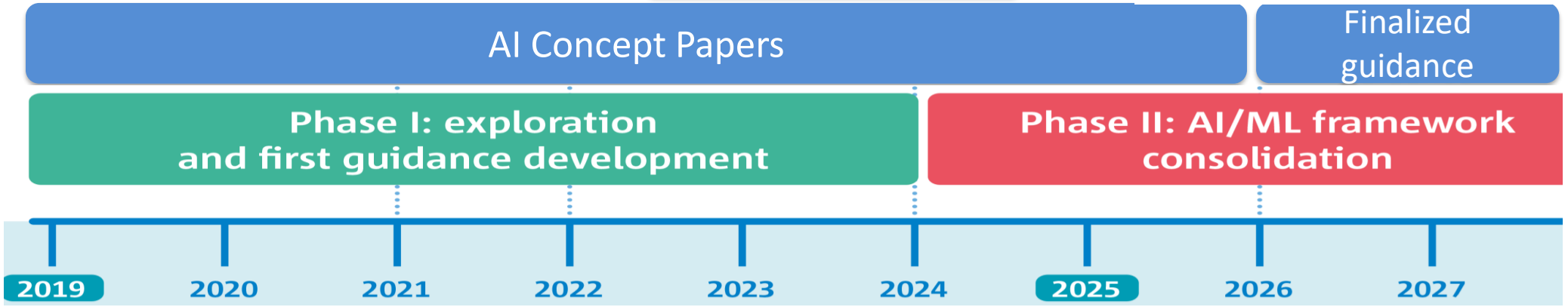
- **Final public report – April 2024**
- Stakeholder day #4 & final event "EASA AI days 2024" - Köln **28/29th May 2024 TBC**

MLEAP

Stakeholders days #1 & #3

#1 – 24th November 2021 – EASA Köln

#3 – 25th January 2024 - Toulouse TBC



MLEAP – Task #1 milestones: Data Completeness & 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 space of the intended application*

Task #1 : Data Completeness and Representativeness

Task #1 objectives (so far)

- State-of-the-art: Provide a list of factors influencing the choice of tools and approaches in order to assess the completeness and representativeness of databases, with corresponding justifications and bibliographical references.
- Synthesis: Present a draft structure of the selection grid for the assessment tools and methods.
- Testing: 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.

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Main take aways of the state-of-the-art

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 actually hard and challenging 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

AIRBUS

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Main take aways of the state-of-the-art

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 > > >

Task #1 : Data Completeness and Representativeness

Synthesis: Building the selection grid

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)
- Labelling (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)

AIRBUS

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Main take aways of the testing phase

No method is **self-sufficient**
They need to be combined to provide meaningful insight

No method is **universal**
The method and their combination must be tailored to each type of task/data

Completeness and representativeness can only be estimated w.r.t ODD specifications
No "absolute measure"

Trade-off between completeness and representativeness for e.g. corner cases

AIRBUS

MLEAP – Task #2 Milestones: Model development

Generalization properties

State-of-the-art analysis:

Available methods and tools to evaluate generalization bounds;
Barriers in generalization guarantees: ML and DL;
Limitation of available methods and common practices;

Identification/selection of suitable methods:

Methods selection;
Projection into the W-shaped approach: ML development pipeline;

Experimentation & Evaluation

} _Task 2: Model generalizability - Summary

Task #2 : Model generalization

Detailed analysis of the state of the art of ML/DL generalization evaluation

Identification of issues related to design and development:

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

Review of methods to boost generalization:

- Regularization, Penalty methods, Data expansion

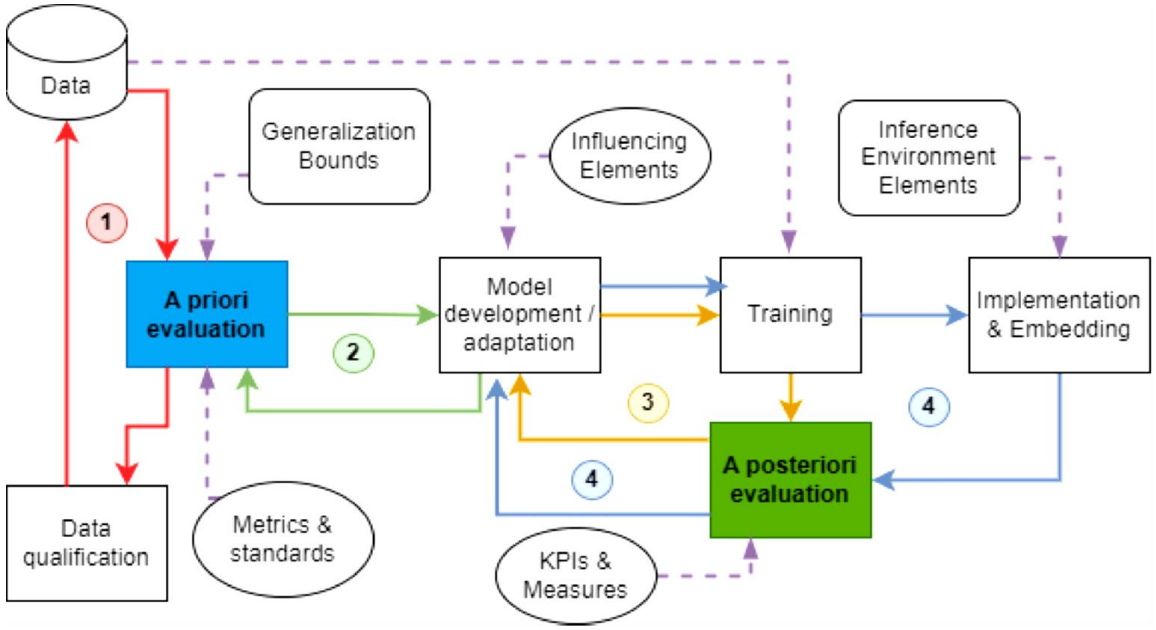
Methods and tools to evaluate generalizability

- A priori evaluation of a model to generalize (Random labelling, data corruption…)
- Trained model behavior assessment (stability, robustness…)
- Statistical evaluation based on empirical measures

Enhanced ML development pipeline

Generalization protocol to capture best practices

(1) Data evaluation and qualification (<=> Task#1)
(2) Model development and adaptation
(3) Model training on the optimized data set (<=> Task#3)
(4) Performance verification in the target environment



AIRBUS

} _Task 2: Generalization limitations - Takeaways

Task #2 : Model generalization

Issues and limitations

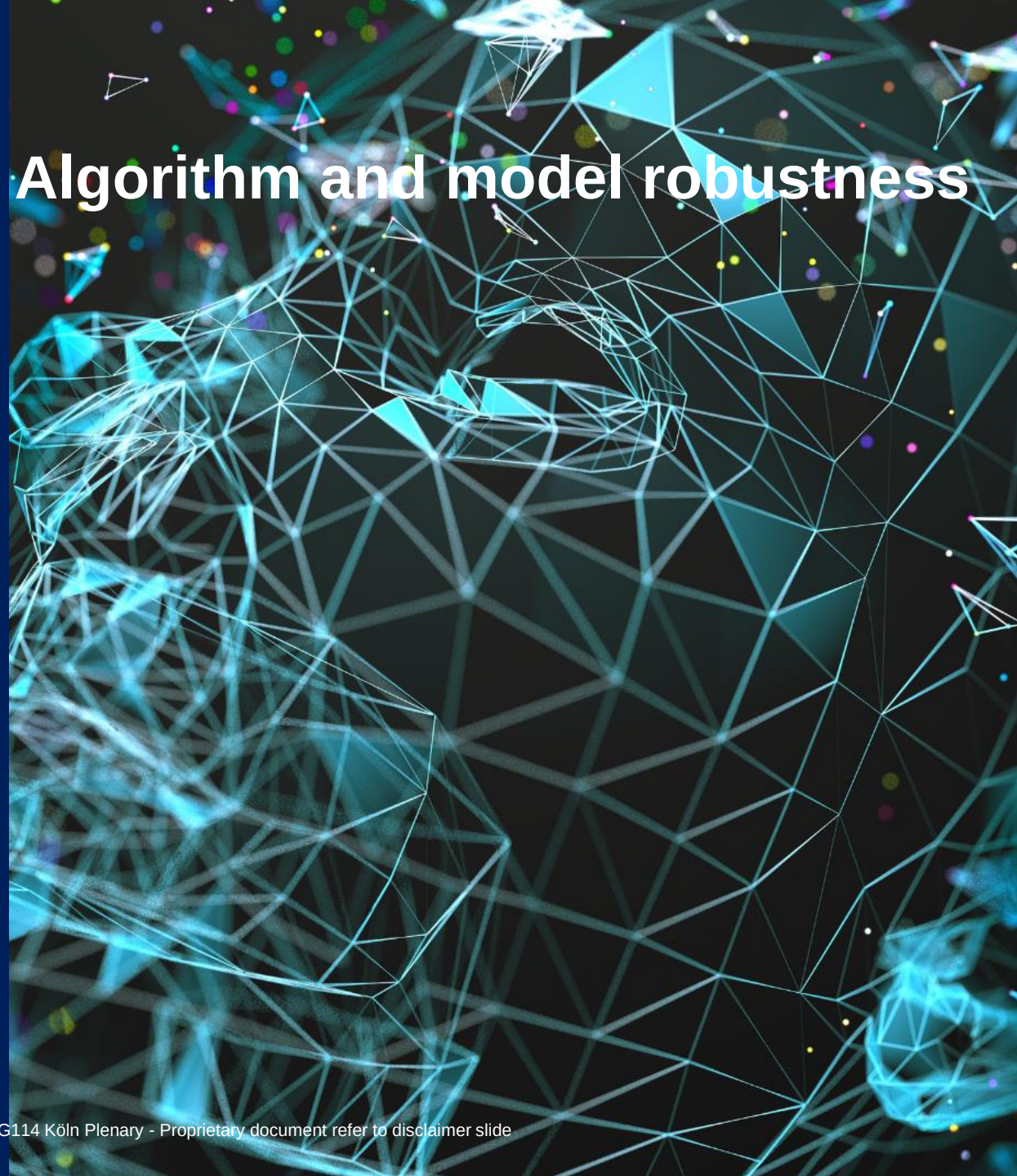
- Based on empirical measures, generalization bounds are efficient statistical tool to estimate level of confidence for low complexity model. In high complexity, bounds are too large to provide appropriate level of confidence
- For high dimensional problems, no mathematical tools available providing guarantees w.r.t. the experimental performance obtained with of such ML algorithms
- Appropriate performance indicators to the application domain cannot ALWAYS be translated by existing evaluation metrics
- Methodologies as best practices at different steps in the W-shaped process

Potential approaches and solutions foreseen

- Leverage the ability of empirical measures to estimate level of confidence for low complexity model in order to assess the one of high complexity models (*eg. Estimation of performance of compressed models*)
- Leverage the multi-criteria methods for models evaluation and performances assessment, a combination of different bounds to assess the same model, make local estimations on high dimensional problems
- Rework the performance indicators to the target application and how they can be translated by existing evaluation metrics
- Identify the limited common practices and development pitfalls leading to a weak performances

MLEAP – Task #3 Milestones: Algorithm and model robustness

Review of methods and tools
Review of methods to identify 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 for the intended application



Task #3 : Algorithm and model robustness

} _Task 3: Model evaluation – Robustness and Stability – Summary

Alignment of different source for the same concepts

Stability and Robustness:

- Stability of the behavior (training, trained, inference)
- Robustness against more adverse conditions

Edge and corner case:

- Corner case, edge case, novelty and anomaly…
- Possible semantic decomposition of the considered cases

Alignment of the literature

- ISO/IEC, EASA concept paper v2, EUROCAE WG114…
- Several vertical but with similar concerns all related to the horizontal approaches (ISO/EC – EU AI Act)
- Overlap of the concepts is favorable to alignment
- Possibility to use requirements from EASA and ISO/IEC

Multiple approaches available

Formal methods

- Solver
- Abstract interpretation
- Optimization

✓ Doable but with local results

Statistical methods

- Combining metrics

✓ Doable but through sampling

Empirical methods

- Field trial
- A posteriori
- Benchmarking

✓ Human intervention needed

} _Task 3: Model evaluation – Robustness and Stability – Takeaways

Task #3 : Algorithm and model robustness

Issues and limitations

- ODD definition is fundamental
  - What attributes?
  - What distribution?
  - What perturbation?
  - …

- Combination of method is necessary
- Formal can cover large part but not remote case
- Statistical can cover remote but isolated cases
- Empirical can cover very rare and human defined cases

- Tool maturity can vary widely
  - Scalability can be an issue, as well as applicability
  - Industrial tools are maturing and do not cover everything

Combination of the approaches

➢ Application on the 3 use cases proposed

➢ Experimentation using the 3 approaches combined

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

WHAT's next for MLEAP?

PROJECT:

MLEAP Final report in 1 year from today

EVENTS:

January 2024: MLEAP Stakeholders day #3

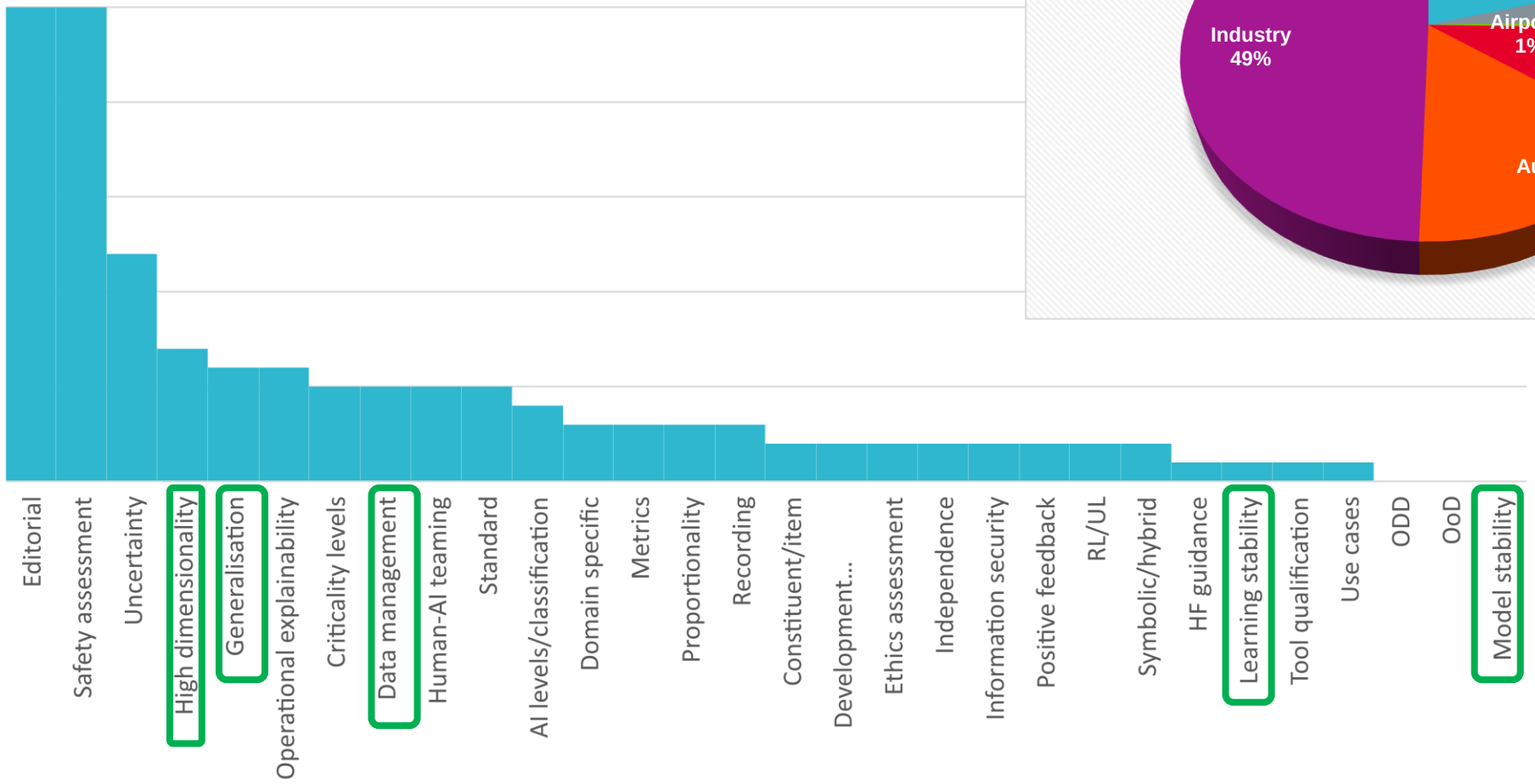
Awareness session conference #2

April 2024: Knowledge sharing conference #2

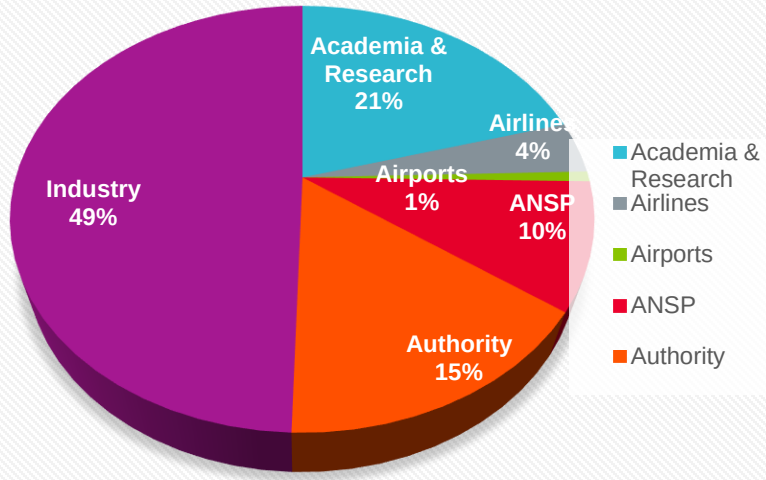
May 2024: MLEAP Stakeholders day #4

EASA AI Concept Paper Proposed Issue 02

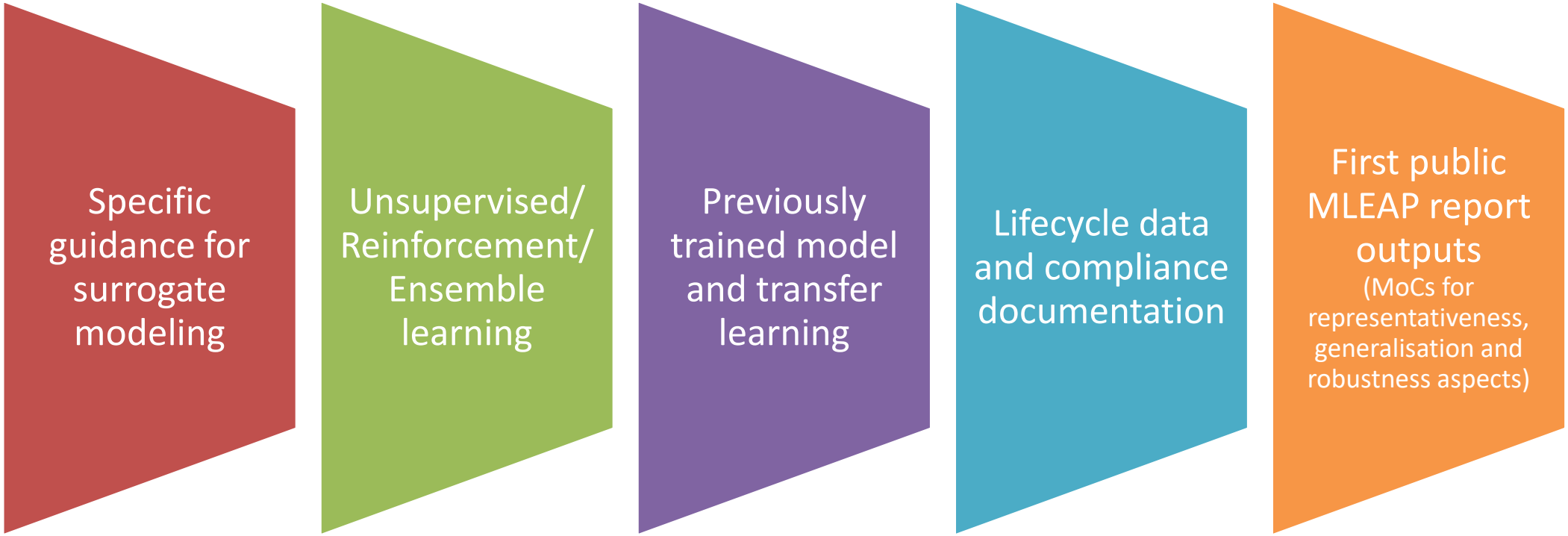
Key themes from the consultation



917 comments from 34 stakeholders

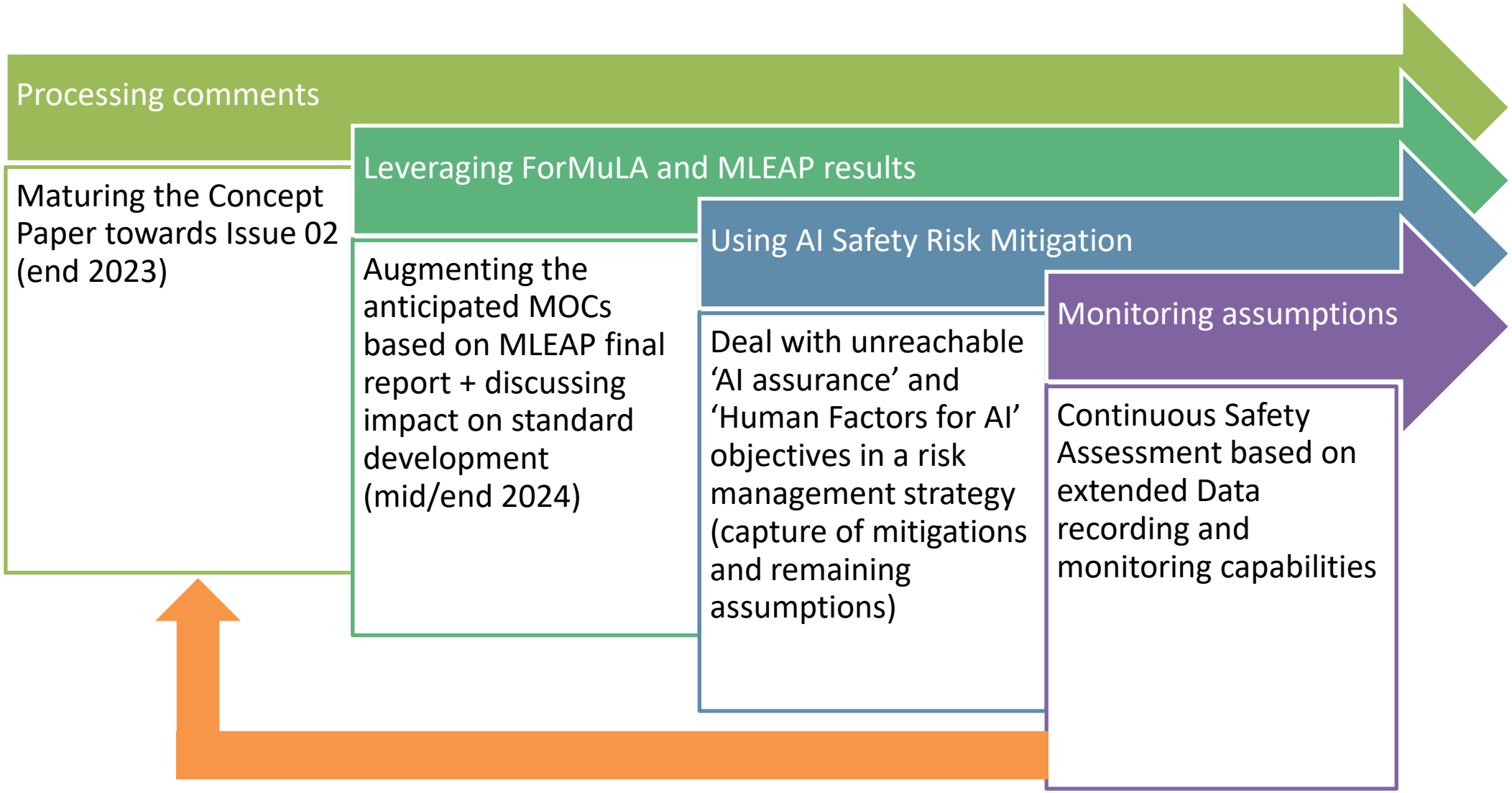


Next steps for Concept Paper Issue 02



→ However, some of the necessary MOCs may still be challenging or non-scalable in high-dimensionality, so what is EASA anticipated approach for first certifications/approvals?

Way forward for remaining challenges



Perspectives and conclusions

Looking back at our top 3 challenges:

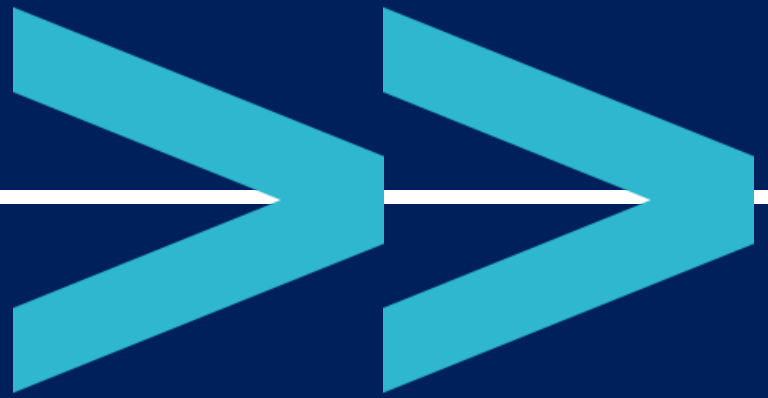
1. **Means of compliance for Learning Assurance**
   → Anticipate additional research needed: any thoughts from the audience?

2. **Operational explainability & human factors for AI**
   → Horizon Europe contribution agreement project will likely be launched to cover the HF for AI aspects, in addition to IPC and MoU activities

3. **Ethics-based assessment – social & societal aspects**
   → Evaluate and refine guidance based on use cases, possibly using focused surveys to aggregate direct feedback from aviation professionals and public.

Q&A

{Thank you}



AIRBUS