This project has received funding from the European Union's Horizon Europe Programme

RESEARCH PROJECT EASA.2021.C38, MLEAP

UNIFIED DELIVERABLE PHASE 2 – EXECUTIVE SUMMARY

MLEAP: Machine Learning Application Approval

Disclaimer

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 deliverable 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.

DELIVERABLE NUMBER AND TITLE:	MLEAP Unified deliverable phase 2 – Executive summary
CONTRACT NUMBER:	EASA.2021.C38, MLEAP
CONTRACTOR / AUTHOR:	Airbus Protect, LNE, Numalis
IPR OWNER:	European Union Aviation Safety Agency
DISTRIBUTION:	Public

This is the executive summary for MLEAP Unified deliverable phase 2. The full report can be downloaded on the EASA MLEAP webpage.

APPROVED BY:	AUTHORS	REVIEWER	MANAGING DEPARTMENT
Olivier GALIBERT	Thiziri BELKACEM Arnault IOUALALEN Swen RIBEIRO Noémie RODRIGUEZ Jean-Baptiste ROUFFET	MLEAP consortium	*Project Manager : Michel Kaczmarek* *Quality Manager : Bernard Beaudouin*

DATE: 21 July 2023

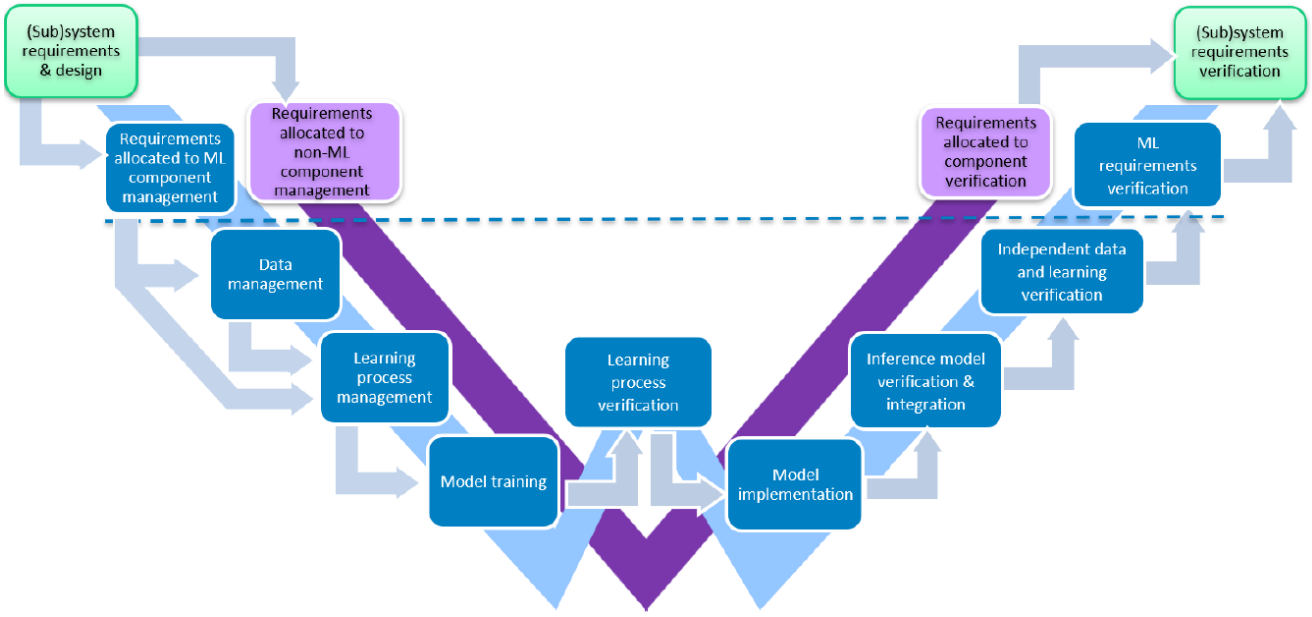
EXECUTIVE SUMMARY

Context

Artificial Intelligence (AI) is becoming ubiquitous and many industrial domains, including aeronautics, aim to harness its promises to improve their performance. The most spectacular progress of contemporary AI comes from Machine Learning (ML) and Deep Learning (DL). These technologies extract and learn behavioural patterns for a given task from corresponding data. This latter is made of a set of samples of the operational context of the target domain and application. However, that same learning process can make it harder for systems including those bricks to be trusted in critical situations. Hence, more adequate approaches need to be developed to build that trust.

In the aeronautics domain, the European Union Aviation Safety Agency (EASA) published its Artificial Intelligence Roadmap in February 2020, followed by the first major deliverable, a Concept Paper *'First usable guidance for level 1 machine learning applications'* in December 2021. This latter has been recently updated to a Proposed issue 02 which was published for consultation in February 2023 to cover level 2 AI applications. These iterative versions of the EASA AI concept paper lay down the basis of EASA future guidance for ML applications approval and set a number of areas in which further research is necessary to identify efficient and practicable means of compliance with the defined 'AI trustworthiness' objectives. Recently EASA updated its Roadmap 2.0, confirming the framework of learning assurance that serves as a reference for the Machine Learning Application Approval (MLEAP) project, which is initiated toward those goals.

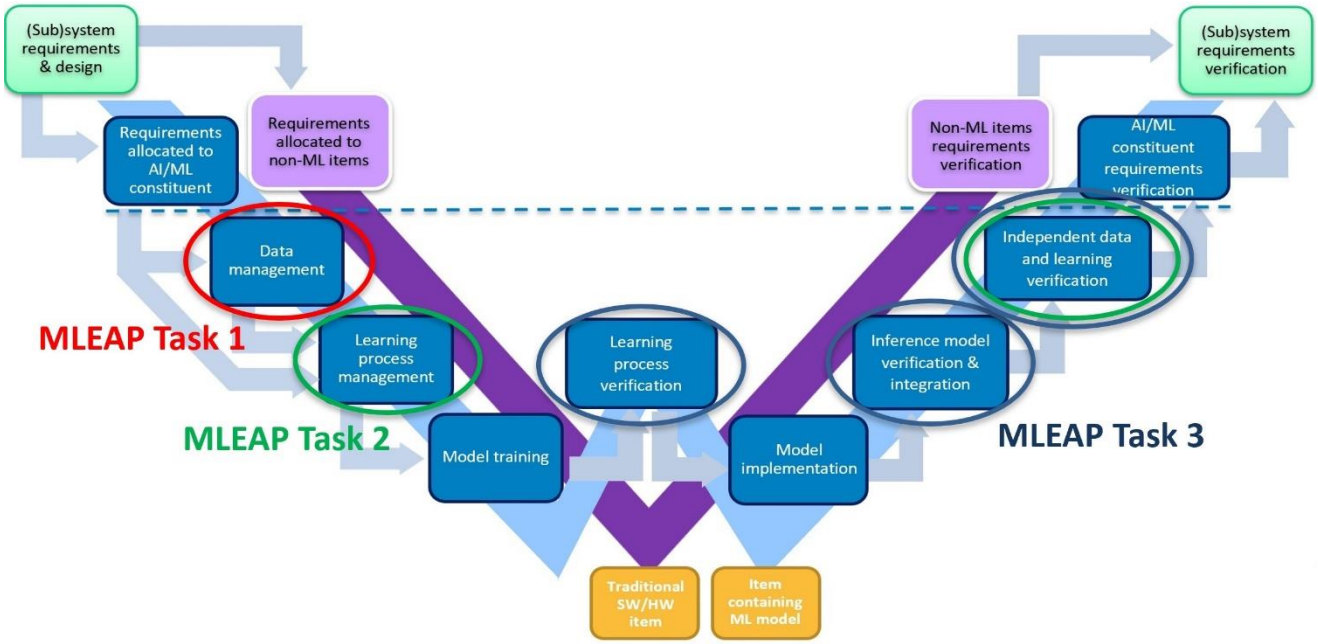
An essential building block of the AI trustworthiness concept is the learning assurance framework which provides an adaptation of development assurance principles to learning approaches. It is instantiated in the EASA concept paper through the so-called W-shaped process, as shown in figure below, and which sets the specific objectives for each step of the learning process.



Global view of the learning assurance W-shaped process overlapping the non-AI component V-cycle process and the safety assessment process

The W-shaped cycle adapts the typical development assurance *V-cycle* to ML/DL. It allows to structure the guidance through blocks composing it. The dotted line is here to make a distinction between the use of traditional development assurance processes (above) and the need for processes adapted to the data-driven learning approaches(below), where the learning assurance processes start below the dotted line. The W-shaped process is concurrent with the traditional V-cycle that is required for development assurance of non-AI components.

Focussing on the development of the AI components, the MLEAP project has been tailored to investigate the challenging objectives of the W-shaped process. Funded under the Horizon Europe framework, MLEAP aims to promote AI blocks of the W-shaped process by carrying out three main tasks, each of which will serve one or more parts of the whole process, as shown below:



The positioning of the various parts of the MLEAP project in the W-shaped process

These three tasks under development, during the 2-year life of this project, correspond to:

- **Task 1:** Dealing with data completeness and representativeness, with the handling of the corner cases. It focuses on data quality verification and sets selection grid of a set of methods that can be used to make sure that the ML pipeline is being developed among a trustable representation of the target domain and application by a complete and representative data set.
- **Task 2:** Deals with the characteristics related to the reliability of the ML model built. Hence, this task revisits the model development, through the handling of the generalization properties, and how the learning process can be leveraged to promote the model's ability to scale up the performances to unseen data during training. Hence, this task revisits the development pipeline to promote the models' performances and reduce regression after implementation in the target system.
- **Task 3:** Focuses on the model evaluation to verify the targeted features, in terms of robustness and stability of the performances noticed during the evaluation. Since the model may be confronted with changes in the representation of input data and/or disturbances in

the real world, it is necessary to check that the model is acceptably robust and that its performance is stable despite corrupted or naturally noisy data.

This interim report offers a set of anticipated concepts for the evaluation and certification of AI-based systems supporting the EASA roadmap deliverables, and helps industry stakeholders in planning new strategies for deploying AI in their human and technical organizations. The full report can be downloaded on the EASA MLEAP webpage. This summary provides an overview of the various points covered in the project's various tasks so far and gives a summary of the work carried out, including analysis of the state-of-the-art, discussion of the various existing methods, and the first directions identified which are now being explored.

Report contents and main findings

In order to provide the various stages of the W-shaped process with analysis and recommendations, the work carried out is structured into six different chapters in the whole document: Note that it is the first public intermediate report issued by the MLEAP project, and a second and final version will be issued by May 2024.

The 1st chapter corresponds to the introduction. It provides a detailed description of the research directions issued in this project while defining the boundaries of the expected work. Besides, it highlights the definitions and terminology the work is based upon. Those topics happen to have variable definitions from one document to another. In order to clearly scope the boundaries of the project deliverables, the terminology, and scientific and technical definitions that are used in the project are first refined. A comparison between the different uses and references is also provided (mainly in relation to the intended use in the EASA documents). Further, the performance evaluation metrics used in the evaluation parts of the work are described, highlighting for every measure what kind of performance is targeted. Since different metrics can be used to evaluate both: the generalization performances to unseen data during training, and the performance stability and robustness toward data and/or environment changes, a dedicated section describes the main differences between those metrics and how they can be used to evaluate performances.

The 2nd chapter is dedicated to the description of the use cases that have been selected for the experimental part of the project. It provides, for each use case, a brief background and state-of-the-art analysis, the main objective of the task, the challenges to be addressed in the MLEAP project, and why these are important. The main objective of using different use cases is, on the one hand, to drive and lead the analysis of the state-of-the-art and the methods selection as well as their applicability analysis. On the other hand to evaluate the project findings and make recommendations for AI systems based on real use cases from different domains in aeronautics. These are:

- **Speech-to-text for air traffic control (ATC-STT):** aims to correctly translate the spoken instructions, to ensure that they are well transmitted and received. While the background noise, speech rate, and language accent represent important issues for the STT system, in MLEAP we deal with several accents of spoken English including French and Chinese accents.
- **Automated visual inspection (AVI):** it aims to design a system for the in-service damage detection of aircraft. One of the main challenges is the diagnostic assistance for inspectors to reduce the aircraft maintenance duration, for scheduled and unscheduled events. The main

point is to find acceptable metrics to bring computer vision closer to classical problems such as model development for surface damage detection. Hence, in the MLEAP project, the targeted domain is in-service damage detection including the lighting strike impacts and dents.

- **Airborne collision avoidance system Xu (ACAS Xu):** is an air-to-air collision avoidance system designed for unmanned aircraft (drones). The purpose of an ACAS Xu system is to keep any intruder outside of the desired envelope of the ownship. In this use case, the objective is to produce an ML/DL model that is able to completely fit the discrete input lookup tables.

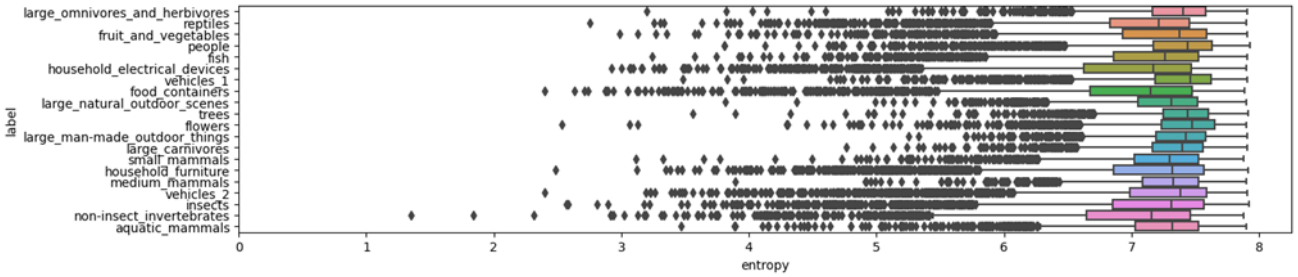
Each of them comes with a set of data and trained ML/DL models. These are either provided by Airbus internal projects or open-source data sets and models, for evaluation and results publication.

Chapter 3 is dealing with the data management aspects of the learning assurance aspects. In particular, methodologies for trying to measure or ensure completeness and representativeness are presented and collated in a selection grid. In addition, approaches to managing edge cases and corner cases are explored. Indeed the EASA guidelines are considering robustness in the sense of changing inputs, including edge cases, corner cases, outliers, out-of-distribution, or even adversarial cases, which can be different in other documents (like the ISO/IEC standards). These changes in the inputs can even be refined in order to distinguish perturbations at different semantic levels, for example at pixel, domain, object, scene, or scenario levels. These different semantic levels of perturbation allow viewing the difficult case, which characterizes the robustness of the system, from different viewpoint which may be linked to the operational design domain definition.
The general problem of completeness and representativeness assessment is broken down into several factors of influence, shedding light on the complexity of the task. More than fifteen factors are identified, and grouped into three categories: technical requirements. i.e. elements influencing the design of the AI system's operational design domain, processes (i.e. of the AI system's development lifecycle), and other data quality requirements (i.e. apart from completeness and representativeness).

Influence factors are used to structure discussions from a normative perspective as well as an operational point of view, by framing the contributions of more than 80 references including academic discussions, methodologies or tools, as well as tools designed for applicative goals.

This extensive work is then summarized into a selection grid isolating around methods that are considered exploitable within the frame of the project. The goal is to test as many methods as possible to provide the future applicant with first-hand feedback and general insight into how completeness and representativeness may be assessed. This work is not prescriptive and does not pretend to be exhaustive.

In the current state of the project, preliminary experiments and conclusions have been obtained on half the methods. Experimentations followed an iterative process by being first tested on a small data set, easy to deploy and manipulate (It uses the ACAS Xu use case, where the data set is sometimes referred to as a "toy" data set). The objective is to get to grips with existing tools or ensuring the correct implementation of the methods.



Boxplots of the distribution of the entropy values of every images in each 20 classes of the CIFAR-100 data set

This figure shows the results of computing pixel-wise Shannon's entropy on the images of the CIFAR-100 dataset (used as a toy dataset). Boxplots are to be read as follows: the line inside the box indicates the median of the distribution (i.e. the value for which 50% of the sample are above and 50% below). The left and right extremities of the box represent the first and third quartiles of the distribution (i.e. 25% of the samples have value below the first quartile and 25% samples have values above the third quartile), respectively. The line at the extremities of the "whiskers" indicate the minimum value (left) and maximum value of the distribution. Finally, the diamonds represent statistical outliers, i.e. single occurrences of values outside the distribution. Note that these statistical outliers might be indicative of outliers in the sense of EASA's Concept Paper. It can be observed that the classes have similar median entropy and range. Nonetheless, they also include a large number of statistical outliers, which could be a starting point for a finer characterization of the data.

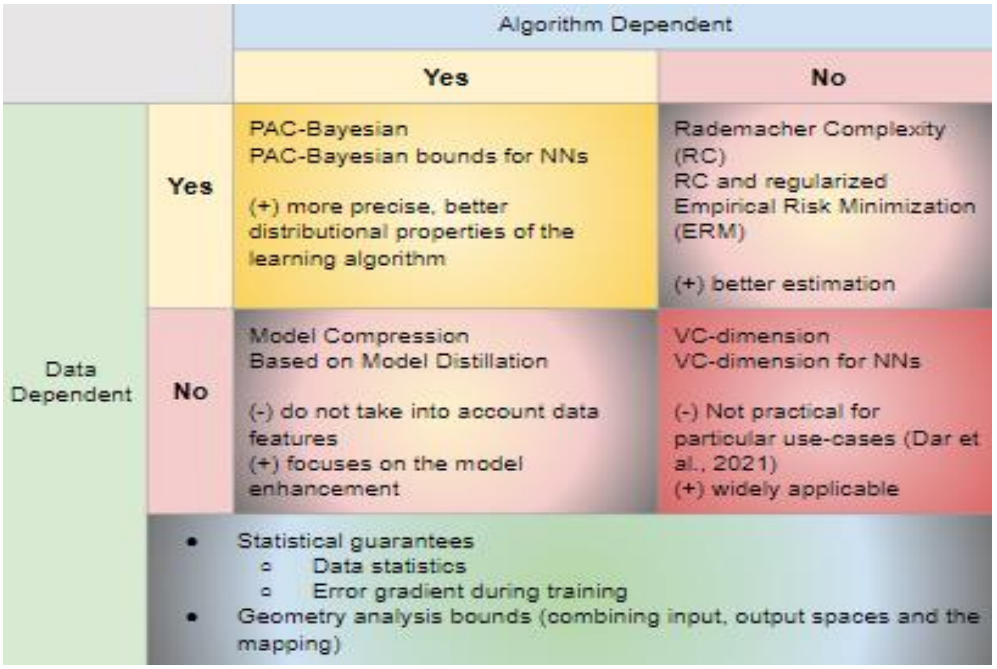
Certain methods could then be applied to larger-scale data sets of task similar to those described in the MLEAP use cases. These data sets are used as an intermediary between toy data sets (for tool validation) and actual use cases (for final analysis) because use cases data sets are harder to access and the scalability of the methods has to be validated beforehand. Moreover, the data sets used are well-known to the experimenters, which allows for more control over the conclusions obtained through the methods. Since no methodology is self-sufficient, this intermediary step is also useful to refine the use of the methods, understanding their limitations and how they can interact with one another to gain the most insight.

Given that the MLEAP project includes three use cases, covering three different AI tasks: classification, object recognition and speech recognition, at various state of progress, experimentations at the present stage allowed to explore all tasks, and thus to identify their specific challenges. Results obtained so far are overall encouraging, with few methods dismissed and interesting potential highlighted. However as mentioned earlier, experimentations have continuously reinforced the idea that there is no one-size-fits-all method, or even a single tool or methodology expressive enough to be used alone on a particular problem.

The next steps of the project will follow the process of testing the remaining methods on a toy data sets, then scaling to larger data sets with experimentations on method combination. Finally, all resulting tools and methodologies should be applied to the MLEAP project use cases data sets for final conclusions.

Chapter 4 of the public deliverable is dedicated to the model development and generalization properties of a trained model. The generalizability of trained models, assessment and evaluation is investigated, including a comprehensive overview and state-of-the-art analysis of existing methods, for ML and DL models, for generalization bounds definition and assessment. Several issues of

generalization and well known ML/DL-related problems of underfitting and overfitting have been investigated. We analysed the existing methods and their limitations to give guarantees about performances of trained models on unseen data. Figure below shows the completed grid, initially defined in the EASA's CoDANN I, and updated with the latest methods reviewed in MLEAP:



State-of-the-art classification of Generalization Bounds methods

After the model training, a set of evaluation measures and metrics can be used to assess the generalizability. The objective is to detect performance dropouts, due to overfitting or underfitting, then boost the generalizability of trained models. While targeting a good model for the industrial application, there is a set of steps to be carried out in order to achieve the desired performance, from the industrial and target system perspectives. Furthermore, by analysing existing AI development approaches in the state of the art and the most common practices in data science, we have identified the major pitfalls and weak practices that can harm the ML/DL system performances. These include, among others:

- The misunderstanding of the generalization bounds, where some norm-based measures negatively correlate with generalization;
- Several common mistakes and pitfalls in practice while developing the ML pipeline, such as the use of inappropriate training objective functions and data representation or split, in addition to inappropriate model complexity and evaluation metrics with respect to the target application and results acceptability;
- A large gap between the expectations from the experimental evaluation compared to the real-world applications, the evaluation metrics of different machine learning applications, such as Mean Squared Error (MSE), precision, and recall, are used to measure only the technical performance of the ML/DL component, however, in the industrial performance assessment how far does the empirical assessment reflect the real model's efficiency?
- The acceptance of the achieved performance : What does it mean to have a 95% accuracy ?
- An appropriate performance indicator to the application domain cannot always be translated by existing evaluation metrics. Hence an adaptation and/or combination of existing processes can be needed to bridge the gap between experimentation and industrial expectations. The classical approach that aims at using a set of technical metrics to assess the model's

performance is limited in terms of capturing performance aspects related to the industry (KPIs) and reproducibility.

Hence, the generalization evaluation is ultimately more crucial than originally thought. We have made a detailed analysis of the state-of-the-art of ML/DL generalization evaluation and provided our main observations about the cited methods, concerning the generalization assessment, issues detection, and methods of improvement.

In order to cope with the different issues discussed above, we set the main research and technical questions to be answered by task 2, to build a generalizability promoting pipeline:

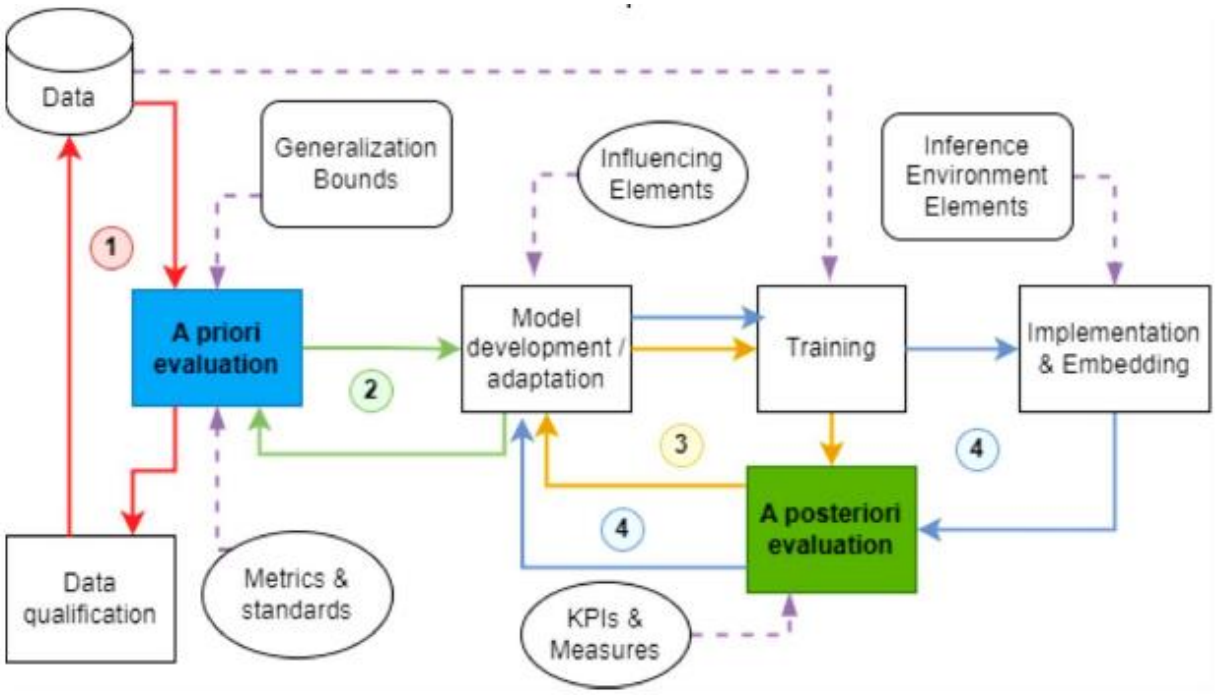
- How to deal with overfitting/underfitting in industry? To deal with this problem, several techniques have been developed to help the ML/DL models generalize better. Although the original model may be too large to generalize well, regularization techniques help limit learning to a subset of the hypothesis space, where resulting models will have manageable complexity. By combining different methods, it makes the model generalize better, independently of the generalization type (domain-based, multi-tasking, OOD-based).
- How to bridge the gap between experimentation and industrial expectations? To bridge the gap between the empirical and industrial processes, we need to leverage the evaluation metrics in the way that they reflect the targeted performances, and integrate the KPIs in the training objectives and the evaluation pipeline as well.
- How to cope with common data processing and evaluation mistakes? To ensure a more rigorous evaluation pipeline, we suggest that the complete roadmap, from the data preparation and qualification step to the model validation and release, benefits from some software engineering best practices, such as the scenarios building for test, and the iteration on the process of data set improvement, evaluation benchmarks as well as the verification and validation process of the final trained model.

After setting the research directions, we defined an ML/DL development, depending on the target application and the task being addressed, the evaluation of ML/DL models is two folds:

- ***A priori evaluation.*** Where the listed development and design pitfalls and mishandlings of the task should be evaluated, in order to prevent harmful learning and development process of the model, as well as the generalization guarantees evaluation and bounds identification. Finally, with respect to the target system requirements, a hypothesis on the performance requirements of the ML/DL model should be made;
- ***A posteriori evaluation***. Where a set of technical metrics should be used, w.r.t. target task, along with a set of the domain specific (business) key performance indicators that verify how well the model can meet the expectations of the final application, in addition to the generalization bounds verification. Finally, the hypothesis on the performance requirements of the ML/DL model will be either verified, and hence validate the resulting model, or compared to the performance of the obtained model and then allow to identify ways to optimize the model better.

Hence, The pipeline[1] is made of four main steps:

1. Data evaluation and qualification (related to Task1). Aims to determine a minimal size of data needed, perform the quality evaluation (completeness and representativeness), define the enhancement operations (data augmentation, processing, cleansing, balancing, and splitting);
2. Model development and adaptation. It takes into account the data constraints (size of inputs and type, alignments…), leverage the mappings between the inputs and outputs, include the generalization bounds estimations in the model design and architecture enhancement, as well as the metrics selection and acceptability criteria definition ;
3. Model training and evaluation on the optimized data set. It includes a definition of the benchmark including a set of industrial KPIs, to define and/or select adapted evaluation measures and thresholds. In the a posteriori evaluation of the trained model (related to Task 3), an empirical and/or statistical evaluation of generalization and robustness is made;
4. And finally, performance verification in the target environment to verify the performances after implementation, take into account different environment and system elements impacting performances with regard to the system/target performance requirements. At this stage, an important drop in performances can drive a step-back to the training and even to the model design and adaptation, to make sure that there will be less surprises after the model integration in the target system.



A General Framework projecting the identified methods for performance evaluation and verification, on the W-shaped process

Note that this pipeline is so far focused on generalization assessment and how to conduct the development in a way to reduce the drop on model performance after implementation. This process

[1] This process is designed to an offline models training setting. It applies to supervised models development as well as an unsupervised setting, and can be further extended to online training settings (e.g. lifelong learning frameworks).

will be developed more in the next phases of the project, based on findings related to the data evaluation and preparation (Task 1), to take into account elements that could harm the model performance, and consider recommendation about robustness and performance stability evaluation (Task 3), to validate the model after its implementation.

Chapter 5 explores the issues of robustness and stability with first a global view of evaluation approaches then a specific overview of formal and analytic methods as applied to models.
The literature on stability and robustness is not entirely homogeneous across standards or the state of the art. For example the concept of stability, in the sense of an algorithmic property differs between the ISO/IEC literature and the EASA documents (such as the concept paper, the CoDANN reports). If stability is present in the ISO/IEC literature, it is largely absent from the ISO/IEC technical literature on information technology (even outside the subcommittee studying AI). It usually refers to a property of a material or a mechanical device that is not applicable in the current context. However the concept of robustness, and even the robustness in the context of artificial intelligence is far more present. However, the different concepts of robustness that the EASA distinguishes (robustness of the training algorithm, the trained model or the inference model) are more or less aligned however between the two. They both refer, to some degree to the performance of a machine learning model holding even in presence of changes in its input. This notion is then considered along the life cycle of the machine learning model, for example using the W cycle described in the EASA concept paper, or another life cycle model, for example the one used in ISO/IEC 22989. In both cases, it is possible to see the correspondence between phases and the properties to be assessed during these phases.

The main conceptual difference between the notions of stability and robustness in ISO/IEC standards (from the JTC 1 / SC 42) and the EASA concept paper is that the EASA CP separates them into two different concepts, whereas ISO/IEC tends to unify them under the same name of robustness. For the EASA concept paper, robustness is based on the notion of input adversity, whereas stability is more focused on normal inputs. ISO/IEC views them in the same way since the robustness has to be defined in “any” condition, so it is both true for adverse or normal inputs.

With the current state of technology, most of the robustness and stability properties that can be verified are mostly at local properties (except in some specific cases where the AI dimensionality allows some global verification to be done). This limitation does not prevent a meaningful process of validation to take place. For this, the properties must be properly defined. For example, properties may express some form of stability of a machine learning system (maximum stable space), others may express some form of bounded behaviour reachable (reachability), or some form of local interpretation (relevance). These properties can be assessed using different methods, each with its own advantages and drawbacks, as well as its level of industrial maturity. Chapter 5 analyses statistical, formal and empirical approaches that can be used. For each a survey is done to distinguish which techniques can be used, what tools are identified and what is their industrial availability and maturity.

In order to evaluate the robustness or the stability of a ML system, it is possible to apply a statistical methodology. In short, the general method consists in choosing the data set to evaluate the ML system on, as well as the metrics that will be calculated. To do this, the general method will select one or more metrics in order to consider them together. These metrics will then be applied on the machine learning system using the testing data in order to assess its robustness or stability properties. Performing a testing protocol is not unique to machine learning models and considerations include the setup of the testing environment, what and how to measure, and data sourcing and characteristics. During testing, planned data sourcing and availability of computational resources are

important considerations due to the sometimes-massive amounts of data and computational resources required by machine learning models.

To constitute this corpus of data, corpus amplification can be used under the control of the operational design domain definition. Not only the operational design domain can describe perturbations that can affect the input data, but it is also possible to expand the coverage of the test set data. Beyond perturbations that can affect the input data, the test data must also reach possible edge or corner cases. For this purpose, using either white box or black box testing techniques can be used to generate edge or corner cases.

It is also possible to rely on formal methodology to assess the robustness or stability of machine learning components. Several approaches are available, for example using solver, abstraction interpretation, reachability, or model-checking techniques.
Solver can rely on different representations of the property to be tested, such as a mixed-integer linear programming problem, (a logical formula using satisfiability modulo theory or satisfiability modulo convex). They encode all computations of a given machine learning model as a collection of constraints and then use these constraints to prove robustness properties. Depending on the machine learning model's architecture, these methods can be complete or incomplete.

Abstract interpretation relies on a theory that constructs controlled approximations that can be built using different representations of the domain, such as boxes, pentagons, octagons, templates, polyhedrons, zonotopes, etc. It is used to provide an incomplete, deterministic, and white-box method that can verify the robustness of large machine-learning models. Abstract interpretation proposes an inherent trade-off between precision and scalability.

Reachability techniques allow us to verify the impact that a machine-learning model has over time on an overall system. It can be used in deterministic or non-deterministic environments. In deterministic environments, it combines solvers on a closed-loop system to determine an over approximation of the reachable set of the system at the next iteration of the loop. In non-deterministic environments, it is combined with probabilistic model checking to determine the probability of reaching a set of states. Probabilistic model checking determines the probability of reaching a certain set of states from a given initial state using dynamic programming. By adapting this framework to work with cells rather than single input states, it is possible to obtain an overapproximated probability of reaching a set of states when using a machine learning system.

Property	Definition
Stability (of the training algorithm, trained model and inference model)	$\|x' - x\| < \delta \Rightarrow \hat{f}(x') = \hat{f}(x)$
Bias	$bias^2(\mathcal{F}, n) = \mathbb{E}_{x \sim \mathcal{X}}\left[(\overline{f_n}(x) - f(x))^2\right]$
Variance	$var(\mathcal{F}, n, x) = \mathbb{E}_{D \sim \mathcal{X}^n}\left[\left(\hat{f}^{(D)} - \overline{f_n}(x)\right)^2\right]$
Relevance	Acceptability of contribution of each dimension of the input vector
Reachability	$\mathcal{E}^n\left(x, \hat{f}^n(x)\right) \notin Z$

Properties to be assessed during the evaluation

Finally, model checking is a method to prove that a formal expression of a theory is valid under a certain interpretation. A theory is expressed by a vocabulary of symbols comprising constants,

functions, and predicates to build sentences that state assertions about the intended semantics of an idea. A theory can either be expressed by sentences of predicate logic or expressed by data patterns. Machine learning models are algorithms designed for the discovery and use of data pattern models. The data pattern model is checked against the input.

Empirical methods can also be an option to evaluate robustness and stability properties. Contrary to statistical or formal, they rely at some level on human expertise and expert judgment to make their evaluation.
For example, in the case of a posteriori testing techniques, the field truth is ambiguous. Since it is not possible to determine all possible correct answers a priori, a posteriori evaluations are performed. That is, human annotators or automated measures look at the outputs of the systems to determine whether they are "acceptable" or "incorrect". In field trials, machine learning is integrated into a system that operates in an environment that is realistic for the application. In this context, data acquisition and sourcing are integral to the design and execution of experiments. Finally, benchmarking is a technique to evaluate a machine learning based system.

It is the first step in building confidence in an AI solution based on machine learning models, but it could introduce elements of subjectivity, such as in the tagging or annotation of test data sets by expert practitioners.
Each method has a different suitability and ease of use with respect to the properties to be proven.

Statistical methods are well suited for evaluating stability, bias, and variance, but are not helpful for relevance or reachability properties. Formal methods are well suited for stability, relevance and reachability. Finally, empirical methods have moderate suitability for each property. Also, each method may have differences in terms of ease of setup.

Statistical methods may be the most straightforward way to analyse these properties. However, they require a lot of preparation work in order to set up the right data sets. Also, any attempt to sample exhaustively is immediately limited by the high dimensionality of the input space. In terms of tools, many libraries provide the necessary functions to evaluate statistical metrics. However, few tackle the data issues associated with such methodologies.

Formal methods, while being promising to overcome this limitation, suffer in part from some scalability issues that few tools can overcome. The available tools can vary in terms of maturity. Most are still academic tools; but a few industrial solutions are starting to emerge. These methods offer stronger properties in terms of robustness and stability; however, they are often limited in the scope on what they can prove over the input space.

Finally, empirical methods may be considered as the most practical in the sense that they require the system to be up and running to be evaluated. However, these approaches can only provide a black box understanding of the system properties. Unlike statistical or formal approaches, they do not allow to evaluate the required properties of the system with the same level of confidence. Their use may be considered for applications of low criticality, depending on the objective that the system has to meet.

	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			
Corner case exploration			

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

A priori assessment of suitability of the different types of methods

Overall, any meaningful evaluation of robustness and stability would benefit from a combination of techniques. In this way, the process would cover as much of the input space as possible while maintaining operational feasibility.

Main results and perspectives

Data quality is a difficult topic, science-wise, because of the inherent cost which comes with doing research in the field. Completeness and representativeness are usually not handled per se, and almost no dedicated tools exist. Thus, there is a need to build indicators from more general metrics (such as entropy) or by leveraging different tools (like sample similarity). Intrinsically, the domain is a difficult one, because an objective estimation of completeness or representativeness requires knowing the exact extent and distributions of the phenomena to observe. In addition, there is a necessary trade-off between representativity and case coverage, since rare cases need to be amplified to be modeled correctly. Hence, the Chapter 3 of the whole public report provides an analysis of the requirement for the operational design domain to set the expectation for the representativity-coverage trade-off. Hopefully, the array of tools and methods described in the selection grid should give AI developers a chance to document and justify if the trade-off holds.

The generalizability of trained models assessment and evaluation is investigated in the 4th chapter of the report. The generalization of a ML/DL model is highly dependent on the data quality and the learning process. This latter has been reviewed, while analysing methods to avoid along the way under/over-fitting, taking into account the impact of the quality and volume of the data needed for training. We presented methodologies to right-scale the complexity and capacity of the models depending on the scope of the task under development, and the volume and nature of input data, while measuring the level of generalization reached by a training session. Furthermore, an operational proposal is presented on how to leverage the project findings and the selected methods within the

W-shaped process. The objective of the proposed pipeline is to revisit the development and implementation of ML/DL models in industry while taking into account data-level evaluation, learning-level verifications and performances evaluation with respect to industrial expectations, and finally the verification of the requirements after implementation. This approach should be extended to the whole set of tools and methods developed/selected on MLEAP for the next phase.

Measuring the quality of the training step takes part in the larger question of the evaluation of the resulting trained and inference models. Such an evaluation is driven by a number of guarantees that need to be gained on the models to ensure an adequate level of confidence in its intended function at a given level of performance. The chapter 5 of the deliverable focuses on two specific guarantees of stability and robustness of machine learning models. We present multiple approaches, from pure performance measures with empirical, data-based approaches to the validation of explicit properties, in particular of stability, through an array of analytic or formal methods. Those methods, while sometimes difficult to put into practice, allow for very powerful analysis of the behaviour of the models, including at runtime, allowing monitoring of the whole system in a live setup. Hence, these evaluation methods on the trained models' robustness and performance stability can be leveraged in the pipeline developed in chapter 4, to ensure better performances after implementation.

The main results of EASA IPC ForMuLA report (EASA and Collins Aerospace, 2023) finalized in April 2023 will also be evaluated in the coming phase to assess how it can complement the results achieved so far by MLEAP project.

An updated version of the deliverable is expected to be published in May 2024. It will allow validation of the applicability of the methods on the aviation use cases summarized in the beginning of this document, and focus on the actual operational usability and scaling of the various tools identified.

Description of the consortium

The consortium in charge of the project is a partnership of three entities, one of the aeronautics domains, Airbus Protect, and two transverse, LNE and Numalis.

Airbus Protect is an Airbus independent subsidiary bringing together expertise in safety, cybersecurity, and sustainability-related services. As a risk management company, the aim of this entity is to offer end-to-end advisory, consulting services, training programs, and software solutions. Pairing expertise built through large-scale projects with the latest insights from its own research programs, it supports customers, partners, and their ecosystems in different industrial domains. Airbus Protect is already a trusted partner of customers in high-tech industrial manufacturing, aerospace, transportation and future mobility, energy and utilities, financial services, critical infrastructure, governments, institutions, and defense. The mission of Airbus Protect is to contribute to making its clients' businesses and products safe, secure, and sustainable. Airbus Protect brings together more than 1,400 experts based in France, Germany, the UK, Spain, and Belgium, to create a center of excellence to meet the clients' evolving needs. Airbus Protect combines more than 35 years of experience with industry-leading expertise to deliver services in three areas: Cybersecurity, providing consulting and managed security services to help our clients to establish and maintain persistent cyber resilience; Safe Mobility ensuring the safety of tomorrow's smart mobility solutions and smart cities; Sustainability developing new ways of working, new products and zero-emission energy supplies.
Airbus Protect team implements several Data/AI/engineering projects, including MLEAP:

- SmartPlanif / MaiVA (Maintenance Virtual Assistant): an airline-centric tool, supporting customers by automating/providing an increased level of assistance to activities.
- Climate and energy challenge: aims to provide structured and semantic access to a large amount of data on the climate and energy ecosystem.
- eIODA (Environmental Industrial Operations DAta foundation): aims to create a single source of truth for all departments and Airbus divisions to enable Environmental Official Reporting as well as Environmental Performance Management.

The French National Metrology and Testing Laboratory (in French, "Laboratoire National de métrologie et d'Essais" or **LNE**) is a public industrial and commercial establishment (EPIC) attached to the Ministry of the Economy and Finance. It is the central support body for the public authorities in the field of testing, evaluation, and metrology. Its action aims in particular to examine new products and assess their impact in order to inform, protect and meet the needs of consumers and national industry. In this context, it carries out measurement, testing, characterization, and certification work on systems and technologies to support breakthrough innovations (artificial intelligence, cybersecurity, nanotechnologies, additive manufacturing, radioactivity measurement, hydrogen storage, etc.) for the benefit of the scientific, normative, regulatory and industrial communities. LNE has particular expertise in the evaluation of artificial intelligence (AI) systems. It has carried out more than 950 evaluations of AI systems since 2008, notably in language processing (translation, transcription, speaker recognition, etc.), image processing (person recognition, object recognition, etc.), and robotics (autonomous vehicles, service robots, agricultural robots, collaborative robots, intelligent medical devices, etc.). It participates in the major challenges of AI by developing standards to guarantee and certify these technologies. The collaborative projects that it conducts at the national, European, and international (in particular via its strategic partnership with NIST on AI and robotics), aims first and foremost to define standards, and protocols (using various conformity assessment methods: literature review, testing, on-site audits), metrics and testing environments (databases, simulators, physical or mixed test benches) for AI, are varied and involve it in almost all technical and socio-economic issues, ethical questions, and sociological issues and networks of institutional actors

(programmatic collaborations with the OECD, the High Authority for Health, the Cofrac and the most French Ministries) and industrial partners (agreements with Thales, Dassault, Airbus, Facebook, CEA, etc.) in the field. In December 2020, as an impartial and independent third party, it launched a working group to define in a consensual manner the first AI certification standard

Numalis is a software editing company specializing in the topic of the reliability of AI systems. The goal of Numalis is to allow companies to accelerate the path of AI adoption, by allowing its design, validation, integration, and deployment to be more reliable. Numalis is involved in several industries with safety critical concerns such as Defense, Aeronautics, Aerospace, Railway, (and Healthcare). For them, Numalis provides a unique set of tools and expertise in order to improve the maturity (and ultimately the adoption) of their use of AI technologies in their future systems. Currently, Numalis has developed Saimple, a solution based on abstract interpretation. By using only formal analysis, Saimple allows to measure the robustness of neural network or support vector machines (SVM) models against specific types of perturbation tied to the domain of use employed, visualize in a human readable fashion the robustness across the input space, and extract explainability components from the system. As robustness and explainability are key components in most software quality models as well for the future EU regulation (the AI Act), Numalis aims at develops also standards at the international level to bring uniformity to processes across all industries. These standards are written in order to bring good practices on the use of formal methods on AI and to that effect, Numalis is currently the editor of ISO/IEC standardization documents (the ISO/IEC 24029 series) related to the assessment of the robustness of neural networks. Founded in 2015 in Montpellier, Numalis employs 18 persons who are mostly PhDs and engineers specialized in formal methods and software development.

European Union Aviation Safety Agency

Konrad-Adenauer-Ufer 3
50668 Cologne
Germany

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

An Agency of the European Union