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

RESEARCH PROJECT EASA.2021.C38, MLEAP

FINAL REPORT – EXECUTIVE SUMMARY

# MLEAP: Machine Learning Application Approval

An Agency of the European Union

**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.

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. All logo, copyrights, trademarks, and registered trademarks that may be contained within are the property of their respective owners. For any use or reproduction of photos or other material that is not under the copyright of EASA, permission must be sought directly from the copyright holders.

All images, results, models, and illustrative examples that do not belong to the consortium, are provided with references to proprietary public sources.

Reproduction of this deliverable, in whole or in part, is permitted under the condition that the full body of this Disclaimer remains clearly and visibly affixed at all times with such reproduced part.

**DELIVERABLE NUMBER AND TITLE:** MLEAP Final report – 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 Final report. The full report can be downloaded on the EASA MLEAP webpage.**

| APPROVED BY: | AUTHORS | REVIEWER | MANAGING DEPARTMENT |
|---|---|---|---|
| Olivier GALIBERT | Thiziri BELKACEM<br>Arnault IOUALALEN<br>Swen RIBEIRO<br>Noémie RODRIGUEZ<br>Jean-Baptiste ROUFFET<br>Jérémy BASCANS<br>Quentin SIGNE | MLEAP consortium | *Project Manager : Michel KACZMAREK*<br>*Quality Manager : Bernard BEAUDOUIN* |

**DATE: 28 May 2024**

# 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 comprises 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 modules, 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 primary deliverable, a Concept Paper, *'First usable guidance for level 1 machine learning applications'* in December 2021. This latter has been recently updated to EASA Artificial Intelligence Concept Paper Issue 2, published in March 2024, to cover level 2 AI applications. It refines the guidance for Level 1 AI applications and extends the exploration of several concepts, such as *learning assurance*, *explainability* and *ethics-based assessment*. This new issue provides comprehensive guidance for developing and deploying Level 2 AI-based systems, which concerns human-AI teaming applications, including cooperation and collaboration actions where AI systems automatically make decisions under human oversight.

These different versions of the EASA AI concept paper lay the basis of EASA guidance for ML application approval. They set several areas where further research is necessary to identify efficient and practicable means of compliance within a defined *AI trustworthiness* objective. Hence, the framework of learning assurance, namely the *W-shaped learning process*, has been updated. This process serves as a reference for the Machine Learning Application Approval (MLEAP) project, which is initiated to provide a set of recommendations for methods and tools to achieve the different requirements allocated to the ML components of the system.

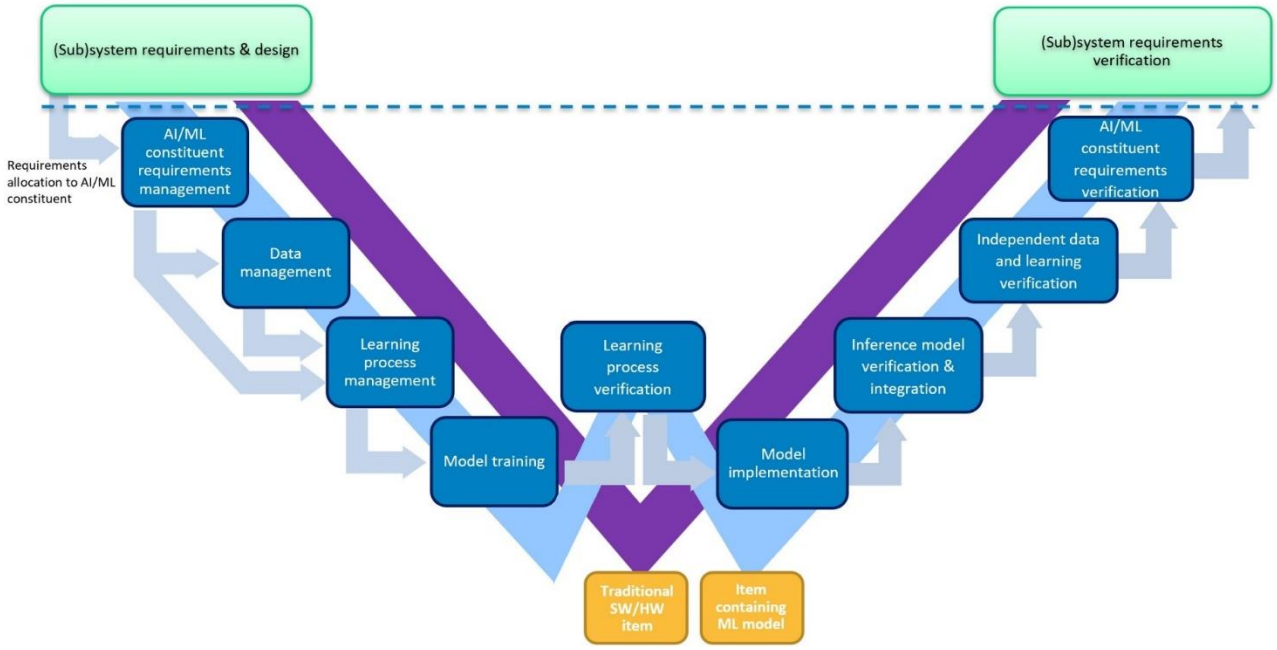


*Figure 1 - Global view of the learning assurance W-shaped process overlapping the non-AI component V-cycle and the safety assessment process. The dashed line separates the system level (upper part) and the AI level (lower part).*

The learning assurance framework is an essential building block of the AI trustworthiness concept, which adapts development assurance principles to learning-based approaches. As shown in figure 1 the learning assurance sets the specific objectives for each development step w.r.t the system level and the AI level of the whole system. Hence, this process adapts the typical software development assurance *V-cycle* to ML/DL-based applications. It allows the structure of the guidance through blocks composing it. The dotted line is here to distinguish 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.

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 in

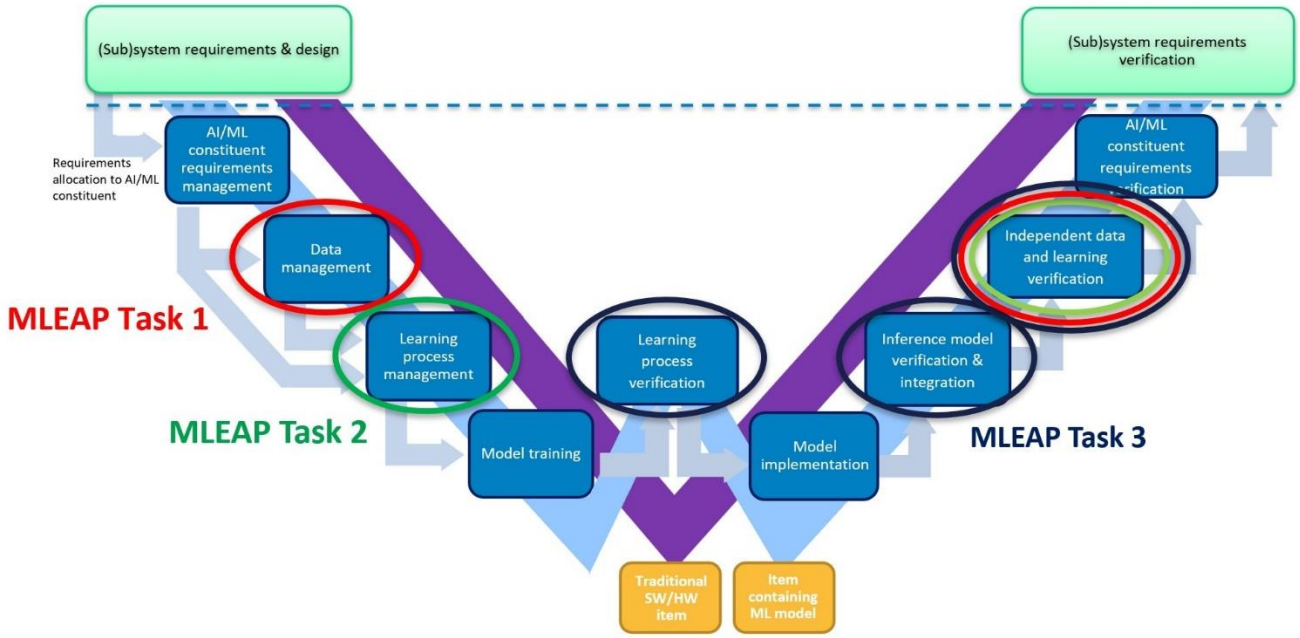


*Figure 2 - The positioning of the various parts of the MLEAP project in the W-shaped process*

The three tasks shown in Figure 2 carried out during the 2-year life of this project, correspond to:

**Task 1:** Dealing with data completeness and representativeness and handling the corner cases. It focuses on data quality verification and proposes a selection grid of methods that can be used to make sure that the ML pipeline is being developed with a trustable representation of the target domain and application. This representation corresponds to the construction of a complete and representative data set. Hence, this task concerns the two main steps of the W-shaped process, including the data management tasks (quality and volume assessment, preparation, and processing…) and the independent data and learning verification of the produced model during the development process. A set of empirically verified tools and methods for data qualification are expected, along with recommendations on how the completeness and representativeness of the datasets can be assessed, in addition to leveraging the learning behaviour of trained models to enhance the data quality.

**Task 2:** It deals with the characteristics related to the reliability of the built ML/DL model. This task revisits the model development by handling the generalisation properties. It explores how the learning process can be leveraged to promote the model's ability to scale the performances to unseen data during training. To do so, generalisation assessment tools and performance enhancement techniques are investigated, avoiding overfitting and underfitting. Hence, to reduce regressions after implementation in the target system, this task focuses on the learning process management and verification, the model training, and the performance verification as concerned steps of the W-shaped process. Evaluated generalisation bounds and other tools for generalisation assessment and assurance are expected, along with a generic AI development pipeline focussing on model performances and how to drive the development steps to achieve the target performances.

**Task 3:** Focuses on the model evaluation to verify the targeted features in terms of robustness and stability of the performances measured in the assessment. Since the model may be confronted with changes in the representation of input data and 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. Hence, this task concerns the learning process verification, the inference model verification and integration, and the learning verification as the main steps of the W-shaped process. Therefore, several questions about the robustness and stability of trained models must be answered, including edge and corner case handling and outliers' management. To do so, the expected outcome is empirically evaluated methods and approaches handling these aspects, in addition to recommendations on how the robustness and stability can be verified during the development process and what needs to be done to avoid weak performances.

This final report offers a set of anticipated concepts for evaluating and certifying AI-based systems supporting the EASA roadmap deliverables. It helps industry stakeholders plan new strategies for deploying AI in their human and technical organisations. The final public report can be downloaded on the project's page. This executive summary provides an overview of the various points covered in the project's tasks. It summarises the work to address the objectives mentioned in the EASA's concept paper (EASA, 2024), for which MLEAP is expected to bring answers. The report proposes a set of studies, including analysis of the state-of-the-art, selection and discussion of several methods, experimental results, and the main conclusions and recommendations verified empirically to meet the objectives explicitly highlighted in Chapter 1 of the full report.

# Report content and main findings

The work is structured into seven chapters in the document to provide analysis and methods/verification recommendations for the various stages of the W-shaped process. Note that this executive summary does not include all the details about the work done; more information, results, and conclusions can be found in the public deliverable.

Chapter 1 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, including terminology definition, the position of the MLEAP project with regards to the W-shaped learning process, the EASA's targeted objectives, as well as the scientific protocol followed during the activities' implementation of this project. The scientific terminology the work is based upon has variable definitions from one document to another. To scope the boundaries of the project deliverables, the terminology and the scientific and technical definitions used in the project are first refined. The objective is to have a shared understanding of the different aspects involved in the activities. A comparison between the various uses and references is also provided (mainly about the intended use in the EASA's documents).

Chapter 2 is about the metrics used in the evaluation parts of this work concerning data quality and volume evaluation, as well as the model's performance evaluation. It describes the different measures, highlighting what kind of performance is targeted for each measure. Since different metrics

can be used to evaluate both the generalisation performance to unseen data during training and the performance stability and robustness toward data and environment changes, a dedicated section describes the main differences between those metrics and how they can be used to evaluate the performance.

Chapter 3 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 on the concerned task and a state-of-the-art analysis highlighting the position of the functions in the project, its main objectives, the challenges to be addressed in the MLEAP project, and why these are important. The goal 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 to address the main challenges of the project, as well as their applicability analysis, and on the other hand, evaluate the project findings and make recommendations for AI systems based on actual aviation use cases. These are:

**Speech-to-text for air traffic control (ATC-STT)** aims to correctly transcribe the spoken instructions by air traffic controllers to ensure that they are well transmitted and received by the pilots. In this use case, the data corresponds to recorded instructions provided along with the corresponding textual transcriptions. While background noise, speech rate, and spoken language accents are essential issues for an STT system, these issues must be considered in an ATC-STT application, which is a critical use case. In MLEAP, we deal with several spoken English accents, including French and Chinese. The objective is to achieve acceptable performances based on the transcriptions analysis (10%-word error rate), regardless of the noise, accents, and speech rates.

**Automated visual inspection (AVI):** This aims to design a system for detecting aircraft damage in-service. 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, the targeted performances have been set and focus on detecting all (most of) the damages (95% accuracy). In the MLEAP project, the objective is to analyse the existing pipeline and provide recommendations to meet the expected performances in the targeted domain, which 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 can completely fit the discrete input lookup tables. In the MLEAP project, we consider the analysis of several models designed for this purpose, focusing on the performance analysis to meet the objective of the use case, as well as the data analysis based on different sources.

Each of the above cases uses a data set and trained ML/DL models. Airbus internal projects or open-source data sets and models either provide these. The open-source and Airbus models and data have

been compared (in terms of quality and performance, as well as compliance with the use case's expectations), enabling analysis of a broader range of application contexts.

Chapter 4 deals with data management in the aspects of learning assurance. In particular, assessment methods for 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 consider 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 input changes can even be refined to distinguish perturbations at different semantic levels, such as pixel, domain, object, scene, or scenario levels for images. These different semantic levels of perturbation bring insight into the problematic cases and characterise the system's robustness from various viewpoints that 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 task's complexity. 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 and operational point of view by framing the contributions of more than 80 references, including academic methodologies and approaches, as well as integrated tools designed for applicative goals.

This extensive work is then summarised into a selection grid isolating methods considered applicable within 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.

Preliminary experiments and conclusions have been obtained on the most pertinent methods. Experiments followed an iterative process, first tested on a small data set that is easy to deploy and manipulate (sometimes referred to as a “toy” data set). The objective is to get to grips with existing tools or ensure the correct implementation of the methods.

Specific methods were then applied to larger-scale data sets of tasks 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 case data sets are more complex 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 reached through the methods. Since no methodology is self-sufficient, this intermediary step is also helpful in refining the use of the methods, understanding their limitations, and determining how they can interact with one another to gain the most insight.

Finally, the pertinent methods were applied to the MLEAP use cases to validate the information they can provide on aviation tasks and data.

The methodologies tested provided results relative to the completeness and representativeness of tabular data and images. Academic approaches requiring prior reimplementation and off-the-shelf industrial tools have also been tested. In both cases, their limitations were experimented with first-hand and discussed in this document to provide a more concrete idea of their usability in an operational context in Chapter 4.

Identified limitations include difficulty accessing direct, quantifiable information about completeness and representativeness due to most methods or tools not being directly designed to address these specific properties. As a result, no tested solution isolates even a subset of the influence factors discussed in the previous sections (while they should ideally isolate them one by one). All methods thus remain fuzzy in the information they provide and require rigorous expert analysis, although they are undoubtedly beneficial to structuring a general assessment approach.

Besides, not all methods fit every data type. Experiments tackled low-dimensional tabular data (i.e., with ACAS-Xu as the target use case) and high-dimensional unstructured data (i.e., images and speech embeddings, although the latter did not yield helpful results). Experimentation allowed us to explore all data qualification tasks identified as relevant to meet the Data Management (DM) objectives. Thus, they need to identify their specific challenges and address them. Overall, the results are encouraging, with few methods dismissed and attractive potential highlighted. However, as mentioned earlier, experimentations have continuously reinforced that no one-size-fits-all method or even a single tool or methodology is expressive enough to be used alone on a particular problem. All the insight gathered from these experiments is synthesised in Chapter 7. A generic way to tackle the assessment of data completeness and representativeness in the general context of the W-shape process is derived. Identified pillars of data completeness and representativeness are the ODD and the trained model: the ODD allows for a priori specifications of data requirements to guide data collection and preparation processes. By contrast, the trained model provides feedback to tune the data set and alleviate biases and other model-specific behaviours.

Chapter 5 of this report is dedicated to the development and generalisation properties of a trained model. The generalisability of trained models, assessment and evaluation, is investigated, including a comprehensive overview and a state-of-the-art analysis of existing methods for evaluating ML and DL models and generalisation bounds definition and evaluation. Several generalisation issues 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 the performances of trained models on unseen data. Figure 3 shows the completed grid, initially defined in the EASA's CoDANN I and updated with the latest state-of-the-art methods reviewed in MLEAP.

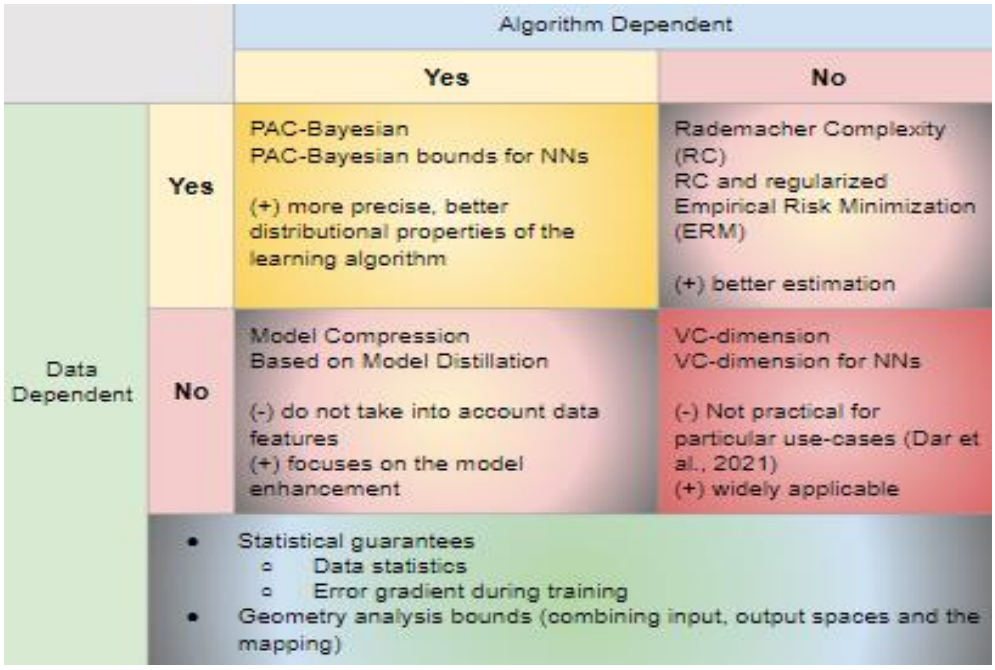


*Figure 3 - State-of-the-art classification of Generalisation Bounds methods*

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

- The misunderstanding of the generalisation bounds, where some norm-based measures negatively correlate with generalisation;
- 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 concerning the target application and results acceptability;
- There is 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, it is necessary to understand how far the empirical assessment reflects the actual model's efficiency and the system-level requirements that should be included;
- The acceptance of the achieved performance and how the system-level monitoring could handle the model's outputs, including errors and uncertainty, but also the meaningful 95% accuracy and the distribution of the remaining 5% of errors;
- An appropriate performance indicator for the application domain is not straightforward, and existing evaluation metrics cannot always translate it. Hence, an adaptation and combination of existing processes can be needed to bridge the gap between experimentation and industrial expectations. The classical approach that uses a set of technical metrics to assess the model's performance is limited in capturing aspects related to the industry (KPIs) and reproducibility.

Hence, generalisation evaluation is ultimately more crucial than initially thought. We have analysed the state-of-the-art ML/DL generalisation evaluation and provided our main observations about the cited methods concerning generalisation assessment, issues detection, and strategies for results improvement.

To cope with the different issues discussed above, we set the leading research and technical questions to be answered by task 2 to build a model's generalisability promoting pipeline:

- ***How do we deal with overfitting/underfitting in the industry?*** To address this problem, several techniques have been developed to help the ML/DL models generalise better. Although the original model may be too large to generalise well, regularisation techniques help limit learning to a subset of the hypothesis space, where the resulting models will have manageable complexity. Combining different methods makes the model generalise better, independently of the generalisation type (domain-based, multi-tasking, and OOD-based).
- ***How can we bridge the gap between experimentation and industrial expectations?*** To bridge the gap between empirical and industrial processes, we need to leverage the evaluation metrics to reflect the targeted performances and integrate the KPIs in the training objectives and the evaluation pipeline.
- ***How do you 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, benefit from some software engineering best practices, such as building scenarios for test and iterating on the process of data set improvement, evaluation benchmarks, and the verification and validation process of the final trained model.

The questions mentioned above have been explored in Chapter 5 of the research. To bring answers to the different raised questions, a two folds experimental process is adopted:
(1) focus on the analysis of generalisation assessment and evaluation,
(2) exploration of the aviation use-cases applying the generalisation assessment and providing a complete alignment between the experimental pipeline and the target objective of each use-case.

To delve into the analysis of generalisation assessment and evaluation, a meticulous examination of how well the trained models generalise across various scenarios and datasets. By analysing the effectiveness of the produced models beyond their training environment, the aim is to gain thoughtful insights into their performances while dealing with unseen data samples. To do so, statistical methods and tools are used to perform this investigation, which lays the groundwork for subsequent analysis into the aviation use cases, where the findings are leveraged to enhance the efficacy and reliability of

the experimental pipeline. Hence, to perform step (1), taking into account the impact of data type and volume, as well as the target task characteristics (the use cases selected for MLEAP), the model's architecture and characteristics, and the state-of-the-art analysis, a set of generalisation bounds have been selected, based on their applicability analysis, for generalisation bounds estimation.

The objective of the experimental work on this part is to highlight the differences between these methods and how they can be used to assess the generalisability of miniature models in small datasets. More details about the results and the behaviour of these methods can be found in Chapter 5. The first results have shown that, depending on the model's architecture, the bounds can have different behaviours, and it is not straightforward to generalise conclusions made based on one method's analysis. Nevertheless, this helped clarify several differences between the learning behaviours that models could have during training and how they can converge based on several optimisation and regularisation techniques. After application to the aviation use cases, the same behaviour was observed. The main takeaway is that using specific generalisation bounds in aeronautical scenarios confirms that when applied to moderately sized networks, they can yield precise bounds for the generalisation gap, introducing confidence that the model's behaviour observed with the test dataset will persist. However, we cannot draw definitive conclusions with deep neural networks as the computed bounds are inconclusive.

The two approaches we tested did not yield explicit positive outcomes in generalisation when mitigating the impact of unbalanced datasets on the performance of the ACAS Xu trained model. Nonetheless, enhancing model stability and robustness may have potential benefits.

In the second (2) part of the experimental analysis, the MLEAP use cases have been explored. The objective of the carried-out experiments was threefold:

I. Analysis of existing pipelines, where the models' performances, data construction and targeted objectives have been analysed and compared for each use case. This exercise highlights the main issues that resulted in a gap between experimentation and target application objectives. Several performance alternatives to improve the results have been tested to support the recommendations for a more consistent ML/DL development pipeline;
II. The already selected generalisation bounds, based on a toy use-case, have been evaluated and compared in the aviation use-cases that are of different dimensionalities and using other data types (images, audio, text…);
III. To solve the same task, compare different architectures and alternatives for data quality and model performance enhancement, using open-source models and tools with other approaches. These architectures have been compared in terms of the task's objective implementation and requirements, pointing out the advantages and weaknesses of each solution and how they can be leveraged to meet the industrial objective.

As a result, to analyse a model's performance, several pieces of information can be drawn from the model's behaviour during its training, such as the correlation between the model learning and the model complexity, the training epochs and the training dataset size, as well as the errors that the model makes after the completed training. All these analyses could help understand the elements that impact the performances and manage them better, in addition to identifying the acceptable weaknesses of the model that the system-level monitoring could handle.

Finally, chapter 5 is completed with the generalisation "assurance" definition, where the objective is to give a quantitative verification of the developed model to ensure compliance with the system-level requirements and less impact on safety. Indeed, generalisation assessment in machine learning relies on the model's ability to maintain performance with unseen data. However, providing guarantees is

challenging, as generalisation bounds are based on assumptions about data quality. Various quantitative analyses are necessary to supplement this, including statistical hypothesis testing, simulation under different scenarios, and conformal prediction techniques. Loss optimisation, a wide range of performance measures, rigorous data analysis, and error analysis are crucial. These approaches provide insights into the model's generalisability, aiding decision-making, particularly in critical cases where even a tiny error margin is unacceptable.

Chapter 6 explores the issues of robustness and stability with a global view of evaluation approaches and 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 and 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. 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 between the two. They both refer, to some degree, to the performance of a machine learning model, holding even in the 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-shaped process 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 (CP) is that the EASA CP separates them into two different concepts. In contrast, ISO/IEC tends to unify them under the same name of robustness. For the EASA concept paper, robustness is based on input adversity, whereas stability focuses more on regular inputs. ISO/IEC views them similarly since robustness has to be defined in “any” condition, which is valid for adverse or regular 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 from taking place. For this, the properties must be adequately defined. For example, properties may express some form of stability of a machine learning system (maximum stable space), and 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 advantages and drawbacks, as well as its level of industrial maturity. Chapter 6 analyses statistical (1), formal (2), and empirical (3) approaches that can be used. For each, a survey is done to distinguish which techniques can be used, what tools are identified, and their industrial availability and maturity.

To evaluate an ML system's robustness or stability, it is possible to apply a statistical methodology (1). In short, the general method consists of choosing the data set to evaluate the ML system and the metrics that will be calculated. To do this, the general process will select one or more metrics to

consider together. These metrics will then be applied to the machine learning system using the testing data 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 it, and data sourcing and characteristics. During testing, planned data sourcing and availability of computational resources are important considerations due to the massive amounts of data and computational resources required by machine learning models.

Corpus amplification can be used to constitute this corpus of data under the control of the operational design domain definition. The operational design domain cannot only 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, either white box or black box testing techniques can generate edge or corner cases.

It is also possible to rely on formal methodology (2) to assess the robustness or stability of machine learning components. Several approaches are available, such as solver, abstraction interpretation, reachability, or model-checking techniques.
Solvers 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 them 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 domain representations, such as boxes, pentagons, octagons, templates, polyhedrons, zonotopes, etc. It provides an incomplete, deterministic, and white-box method for verifying the robustness of large machine-learning models. Abstract interpretation proposes an inherent trade-off between precision and scalability.

Reachability techniques allow us to verify a machine-learning model's impact 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 likelihood of achieving a particular set of states from a given initial state using dynamic programming. Adapting this framework to work with cells rather than single input states makes it possible to obtain an overapproximated probability of reaching a set of states when using a machine learning system.

Finally, model checking is a method to prove that a formal expression of a theory is valid under a particular 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. Sentences of predicate logic or data patterns can express a theory. Machine learning models are algorithms designed to discover and use 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 assess. For example, in the case of a posteriori testing techniques, the field truth is ambiguous. Since it is impossible to determine all possible correct answers a priori, a-posteriori evaluations are performed. Human annotators or automated measures look at the systems' outputs to decide whether or not they are "acceptable" or "incorrect". In field trials, machine learning is integrated into a system that operates in a realistic environment for the application. In this context, data acquisition and sourcing are integral to the design and execution of experiments. Finally, benchmarking is a technique used to evaluate a machine learning-based system.

It is the first step in building confidence in an AI solution based on machine learning models. Still, it could introduce elements of subjectivity, such as in the tagging or annotation of test data sets by expert practitioners. Each method is different in suitability, ease of use, and 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 much preparation 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.

While formal methods promise to overcome this limitation, they suffer partly from 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 more substantial properties in robustness and stability; however, they are often limited in what they can prove over the input space.

Finally, empirical methods may be considered the most practical because 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's properties. Unlike statistical or formal approaches, they do not allow evaluation of 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 | Not suitable | Suitable | Not suitable (the training algorithm is still probably too large) |

| | | | |
|---|---|---|---|
| Stability of the trained model | Could be used but with limited confidence in the results | Suitable | Suitable |
| Stability of the inference model | Could be used but with limited trust in the results | Suitable | Suitable |
| Bias | Not really well suited | Suitable | Not really well suited |
| Variance | Not really well suited | Suitable | Not really well suited |
| Robustness (Corner case exploration) | Could be used for very specific catastrophic scenario | Suitable | Could be used in combination with statistical approach |
| Relevance | Expert judgment | Not suitable since it requires some form of symbolic analysis | Suitable in combination with empirical assessment |
| Reachability | Not suitable since it requires strong guarantees | Not suitable since it requires strong guarantees | Suitable |
| Scalability | Human intervention needed | Doable but through sampling | Doable but locally |
| Methods | • Field trial<br>• A posteriori testing<br>• Benchmarking | • Combining metrics | • Solver<br>• Abstract Interpretation<br>• Optimization |

*Figure 4 - A priori assessment of the suitability of the different types of methods*

Overall, combining techniques would benefit any meaningful evaluation of robustness and stability. This way, the process would cover as much of the input space as possible while maintaining operational feasibility. Chapter 6 offers a variety of results both on open-source use cases (from space, automotive and healthcare sectors) and the avionic use cases described in Chapter 3. These results verify that the requirements on stability and robustness of the EASA CP are indeed applicable and that the combination of methods allows for good practices to be defined and recommended in the context of the generic pipeline (in Chapter 7).

Chapter 7 represents a global analysis of the project's outcomes to address the already specified objectives for each task. It completes the defined ML/DL development pipeline in task 2 to enhance the model's generalisation, robustness, and performance verification in compliance with a system's target application and AI-level requirements. Chapter 7 recalls the main issues identified in the state-of-the-art concerning the analysis of the standard practice. It then brings a set of elements and a checklist of verifications concerning data management, the ML/DL models development and delivery, and the reinforcement of the model robustness; this chapter allows us to locate the stages of the W-shaped process where the main issues have been identified and how they can be handled, through a set of suggested verifications and solutions.
Before presenting the pipeline implementing the W-shape learning assurance, the dependencies between the system-level and AI-level requirements, in terms of performances and intended

functions, have been first discussed to allow a clear understanding of the correlation between these levels and how the objectives must be aligned. Besides, the results of the experimental analysis concerning data qualification and model evaluation have been served to identify the different aspects that need to be verified at each level (e.g., during the model design, during the datasets preparation, the features selection, the training and the learning behaviour analysis of the model…). This analysis highlights the importance of considering the system-level operating conditions in performance requirements for ML modules, ensuring compliance with system-level safety requirements at the AI level through cascaded requirements. In *Figure 5*, the importance of defining performance expectations in a well-known ODD is emphasised, where performance metrics and error analysis would help investigate if the target objectives were met to build trust toward the trained models.

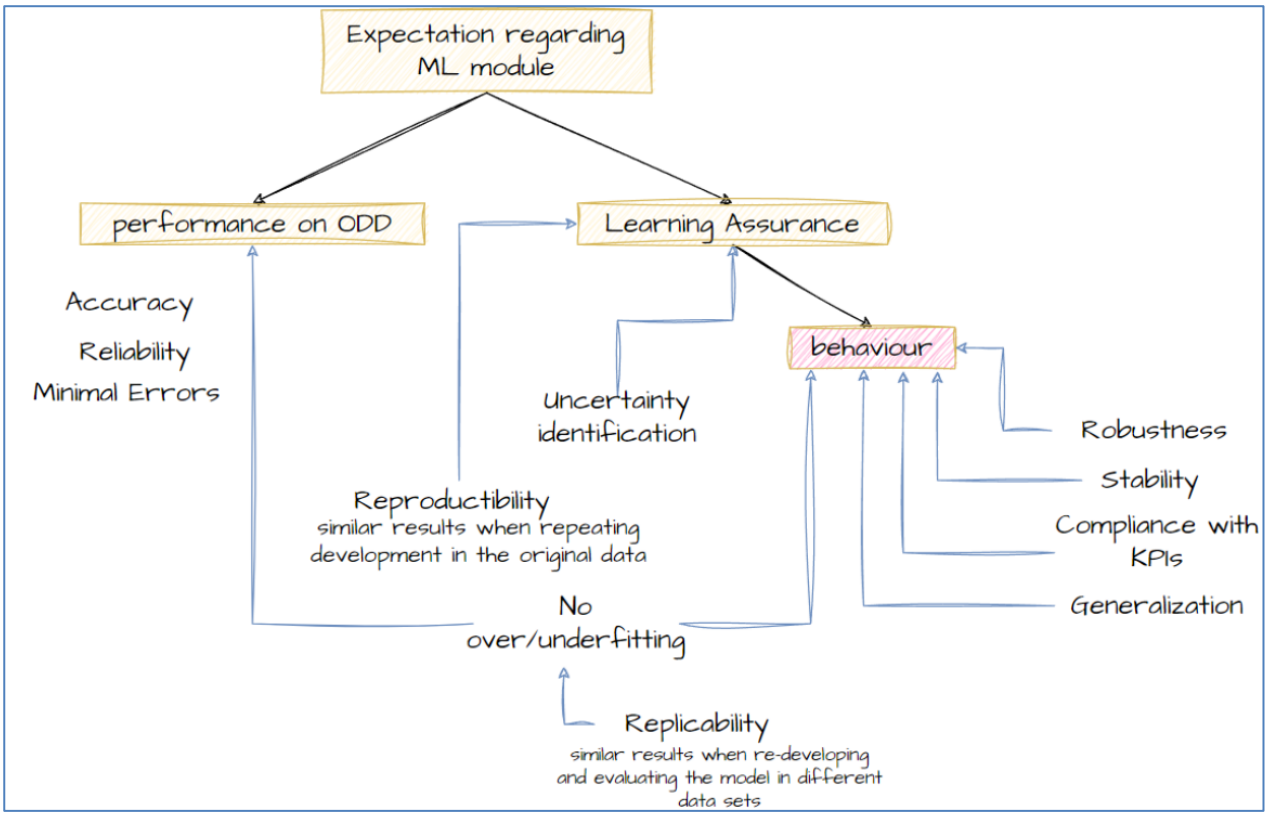


*Figure 5. Expectations regarding the ML module of a system and the related features derived from the development pipeline*

Thus, by being aware of system-level requirements and taking them into account at the AI component level, the latter must enable the trustable behaviour of the trained model. This includes assurance of its robustness, generalisation, a minimum error that does not exceed the system's tolerance limit, compliance with the KPIs, and reproducibility of the results. When verified, this set of measurable qualities will provide a 'guarantee' that the model will perform well in the target system.

A generic development approach is then defined to meet the aforementioned development objectives, including specifying the target application as a mathematical model and defining a complete pipeline. The aim is to implement the W-shaped learning assurance while securing the different stages with a set of verifications to help deliver and validate industrial ML/DL models.
The generic development pipeline is built upon several steps to be performed, depending on the target application and the task being addressed, where the evaluation of ML/DL models is two folds:

***A priori evaluation.*** The listed development and design pitfalls and mishandlings of the task should be identified while evaluating data quality, setting the model target objectives, and determining the generalisation bounds evaluation. The main inputs are the data quality, volume criteria, and generalisation bounds. This first evaluation will specify data requirements regarding completeness, representativeness, and volume necessary for model training. Finally, concerning 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 generalisation 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, which will then allow us to identify ways to optimise the model better.

Hence, the pipeline[1] is composed of six steps:

1. Data evaluation and qualification (related to Task 1). 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…), leverages the mappings between the inputs and outputs, includes the performances influencing elements identified during the ODD analysis in the model design and architecture enhancement, as well as the metrics selection and acceptability criteria definition to be used further in step 3 ;
3. Model training and evaluation on the optimised dataset. It includes a benchmark definition and a set of industrial KPIs to define and select adapted evaluation measures and thresholds. In the a posteriori evaluation of the trained model (also related to robustness verification in Task 3), an empirical and statistical assessment of generalisation and robustness is made;
4. If the objectives are not met based on the validated data and the optimised model, backtracking to the data management and qualification is possible to enhance the data quality and volume to be more adapted to the training requirements while reflecting the ODD definition;
5. After the model implementation, a performance verification in the target environment will be made. This will consider different environment and system elements impacting performances regarding the system/target performance requirements. At this stage, an essential drop in performance can drive a step back to the design phase for a model adaptation to make sure that there will be fewer *nasty* surprises after the model integration in the target system;
6. Finally, after model implementation and during the deployment phase, there could be some updates at system-level requirements from which new AI-level requirements will cascade. Hence, it is necessary to get back to the target application definition and the system

[1] This process is designed to an offline model training setting. It applies to supervised models' development setting, and can be further extended to online training settings (e.g., lifelong learning frameworks) taking into account characteristics of the target application.

understanding to verify if the already validated ML/DL model is still compliant with the target objectives of the whole system. To do so, the verifications and evaluations that have already been made will be rerun for the new application, with updates on measures and metrics selections considered.

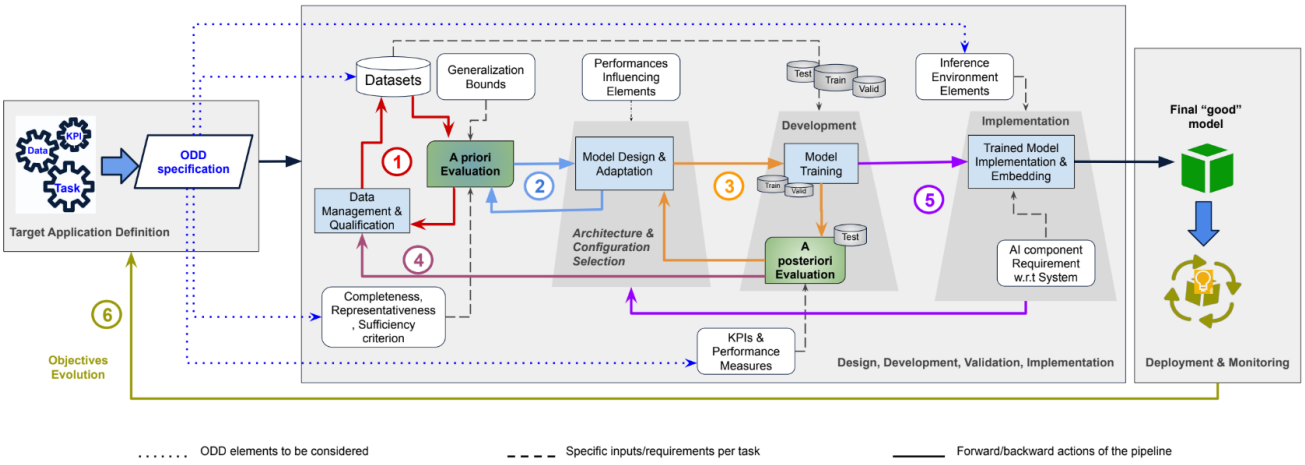


*Figure 6. A general framework for developing and evaluating ML/DL models, implementing the W-shaped learning assurance.*

# Main results and perspectives

Science-wise, data quality is a complex topic because of the inherent cost of doing research in the field. Completeness and representativeness are usually not handled per se, and almost no dedicated tools exist. Thus, indicators must be built from more general metrics (such as entropy) or by leveraging different tools (like sample similarity). Intrinsically, the domain is challenging because objectively estimating 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 must be amplified to be modelled correctly. Hence, Chapter 4 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 allow AI developers to document and justify if the trade-off holds. We would be remiss not to emphasise how much more scientific work is needed to reach operational solutions for more system types. At this point, classification systems are reasonably covered, but other system types, such as transcription, are much less so.

The generalisability of trained models' assessment and evaluation is investigated in Chapter 5 of this report. The generalisation of an ML/DL model depends on the data quality and the learning process. This latter has been reviewed while analysing methods to avoid under/over-fitting, considering 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 generalisation

reached by a training session. Furthermore, the selected methods have been evaluated in a toy use case, and their analysis validated in the real aviation use cases. Finally, the selected use cases were investigated, and models and data were first analysed, highlighting the weaknesses developed in existing pipelines. These later are then deep-dived to identify the elements that resulted in a limited performance while comparing the different models in several datasets, pointing out their differences and limitations. Furthermore, to complete the conclusions on the generalisation bounds and the analysis of the gap between the results and the industrial expectations of the use cases, various alternatives were explored, including the estimation of model uncertainty, conformal predictions, and the analysis of errors and their distributions in the test samples. Other model architectures and approaches have also been explored to provide answers to questions concerning the performance limitations of use-case models and their development pipelines.

Measuring the quality of the training step takes part in the larger question of evaluating the resulting trained and inference models. Such an evaluation is driven by several guarantees that need to be gained on the models to ensure an adequate level of confidence in its intended behaviour at a given level of performance. Chapter 6 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, particularly stability, through an array of analytic or formal methods. Those methods, while sometimes challenging to put into practice, allow for compelling analysis of the behaviour of the models, including at runtime, allowing monitoring of the whole system in a live setup. They were proven helpful in various use cases from different sectors and of varying difficulty. The experiments allowed us to verify and expand the findings of ForMuLA IPC (EASA and Collins Aerospace, 2023) regarding formal methods. 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.

Finally, the different analyses of the tasks of this project have been leveraged to construct an operational proposal implementing the W-shaped learning process, which is proposed in Chapter 7. The objective is to leverage the project findings and the selected methods within a revisited development and implementation pipeline of ML/DL models in the industry while taking into account data-level evaluation, learning-level verifications and performance evaluation concerning industrial expectations (KPIs integration in the learning verification and management), and finally the verification of the requirements after implementation. In Chapter 7, a way toward an understanding ML-level requirements verification is first proposed. A qualitative analysis of ML-based applications is defined to help verify compliance with the system-level requirements. It highlights the main issues handled in each step of the W-shaped process and provides a set of verifications and means to overcome them in the generic pipeline. This approach is finally explored in an experimental analysis where different aspects related to data and models have been investigated, highlighting the impact on models' performances.

# Description of the consortium

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

**Airbus Protect** is an independent subsidiary of Airbus that brings together expertise in safety, cybersecurity, digital, and sustainability-related services. As a risk management company, this entity aims 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 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 defence. 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 1,600 experts in France, Germany, the UK, and Spain to create a centre of excellence to meet the clients' evolving needs. Airbus Protect combines more than 36 years of experience with industry-leading expertise to deliver services in three areas: Cybersecurity, providing consulting and managed security services to help our clients establish and maintain persistent cyber resilience; Safe Mobility, ensuring the safety of tomorrow's intelligent 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) is an airline-centric tool that supports customers by automating and providing increased assistance to activities.
- Climate and energy challenge: aims to provide structured and semantic access to many climate and energy ecosystem data.
- 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 and 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 testing, evaluation, and metrology. Its action aims, in particular, to examine new products and assess their impact to inform, protect and meet the needs of consumers and national industry. In this context, it carries out measurement, testing, characterisation, 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 evaluating 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 significant 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) aim 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 the first AI certification standard in a consensual manner.

**Numalis** is a software editing company specialising in the reliability of AI systems. The goal of Numalis is to allow companies to accelerate the path of AI adoption by making its design, validation, integration, and deployment 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 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, visualise in a human readable fashion the robustness across the input space, and extract explainability components from the system. As robustness and explainability are critical components in most software quality models for future EU regulation (the AI Act), Numalis aims to develop standards at the international level to bring uniformity to processes across all industries. These standards are written to bring good practices in using formal methods of AI. To that effect, Numalis is currently the editor of ISO/IEC standardisation documents (the ISO/IEC 24029 series) related to assessing the robustness of neural networks. Founded in 2015 in Montpellier, Numalis employs 18 people, primarily PhDs and engineers specialising 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