AI SAFETY: PERFORMANCE METRICS AND FAILURE MODES

Claudia Ranieri
Leonardo S.p.A.
Italy

Pietro Ottavio Lecis
Leonardo S.p.A.
Italy

Luca Castaldi
Leonardo S.p.A.
Italy

ABSTRACT

Aim of this work is to update the safety assessment in order to extend the actual safety rigor to Artificial Intelligence (AI) systems. The results are based on the European Union Aviation Safety Agency (EASA) publications, part of the EASA AI Roadmap (Ref. 1), and the starting point of the entire discussion is the *Concept Paper: First usable guidance for level 1&2 application* (Ref. 2). In the guideline, the **Anticipated-MOC-SA-01-7** asks to specify a link between AI performance metrics and safety assessment: this work shows a direct connection between the performance metrics of an AI binary classification algorithm and its Failure Mode, allowing to provide the link required but also to set safety requirements on the performance metrics that will guide the AI development.

NOTATION

*

Symbols

F Learning Algorithm

f True unknown function to approximate

$\widehat{f}$ Trained Model

X Input Space (Input Dataset)

Y Output Space

$\mathbf{x}$ Input Sample

y Output Prediction

*

Acronyms

ACC Accuracy

AI Artificial Intelligence

ALTAI Assessment List for Trustworthy AI

EASA European Union Aviation Safety Agency

EU European Union

FC Failure Condition

FHA Functional Hazard Assessment

FM Failure Mode

FN False Negative

FP False Positive

ML Machine Learning

MOC Means Of Compliance

MLEAP Machine Learning Application Approval

NSE No Safety Effect

PPV Positive Predictive Value

TN True Negative

TP True Positive

TPR True Positive Rate

INTRODUCTION

EASA documentation

The common european goal is to obtain a human-centric development of AI in order to maximize the benefits of AI systems and at the same time prevent and minimize the associated risks.

To support the realization of this objective the European Union Commission introduced the notion of Trustworthy AI, which is applicable not only to the single AI components, but also to the overall context of the AI system, including all the

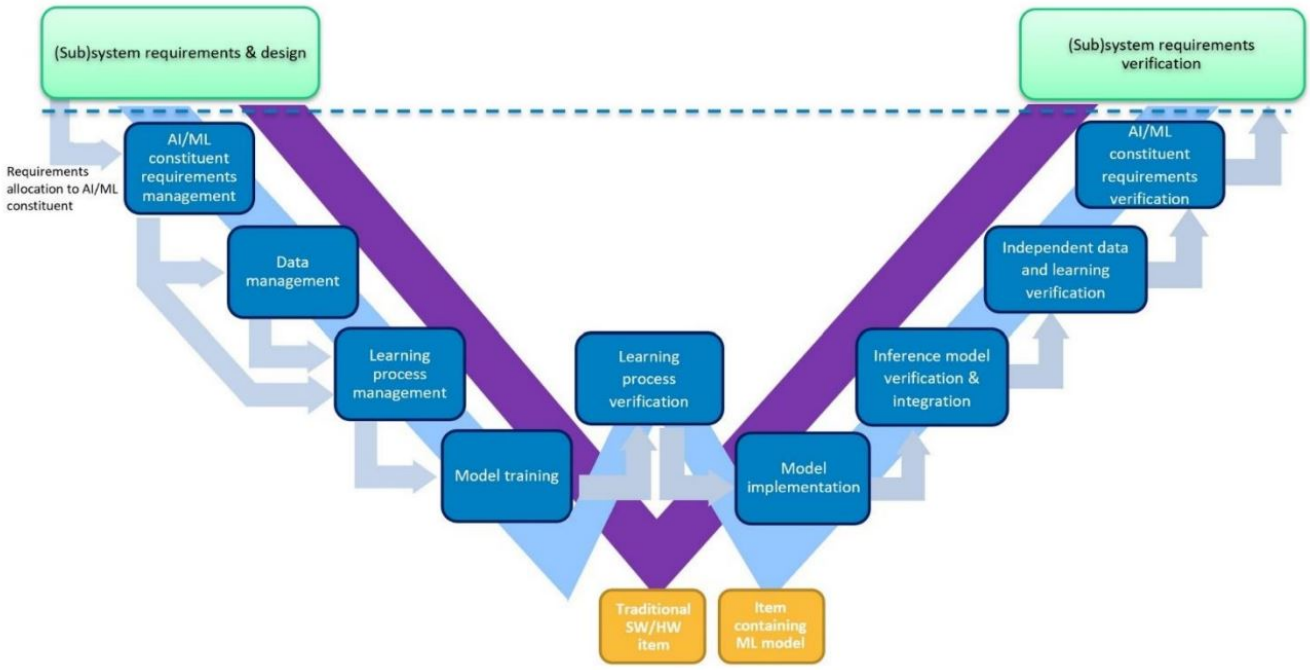

Figure 1. Learning Assurance "W"-Shaped Process Detail. The dashed line separates the system level (upper part) and the Item level (lower part). Over the dashed line the process is the same for AI and traditional systems. Credit: (Ref. 3)

actors and processes that contribute throughout its entire life cycle.

The investigation of this concept led to the EU document: *Assessment List for Trustworthy Artificial Intelligence* (Ref. 4), where a set of questions are presented to perform an initial evaluation of an AI application. The questions are intended for a general use, which means that organizations can modify the list to better capture the relevant elements of a particular application.

These documents delineate the type of AI that will be certifiable in the future; and it is based on this concept, that the European Union Aviation Safety Agency (EASA) is working to introduce proper aeronautical regulation to allow the safe use of AI.

The EASA AI Programme officially started in 2020 with the publication of the document *Artificial Intelligence Roadmap: a human-centric approach to AI in Aviation* (Ref. 1). The Roadmap delineates the path that the Agency wants to follow to gain knowledge on the AI based systems.

One of the latest publications is the *Concept Paper: First Usable Guidance for Level 1&2 Machine Learning Applications* (Ref. 2), which address the AI problem by splitting it into multiple objectives, these aim to respond directly to the questions formulated by the European Community in the Assessment List for Trustworthy AI (ALTAI) (Ref. 4).

This work addresses part of the following safety objective identified in the Concept Paper (Ref. 2), and in particular, is focused on the associated **Anticipated-MOC-SA-01-7**.

Objective SA-01:"*The applicant should perform a safety (support) assessment for all AI-based (sub)systems, identifying and addressing specificities introduced by AI/ML usage.*";

Anticipated-MOC-SA-01-7: *Links between performance metrics and safety assessment*

"*When a quantitative safety (support) assessment is required to demonstrate that the safety requirements are met, performance metrics should provide a conservative estimation of the probability of occurrence of the AI/ML constituent failures modes. Performance evaluation performed as part of the learning assurance per Objectives LM-09 (for the trained model) and IMP-06 (for the inference model) is fed back to the safety assessment (support) process.*"

Safety Assessment

A safety assessment is a systematic process used to ensure the safety and reliability of civil airborne systems and equipment.

It involves identifying potential hazards, assessing their risks, and implementing mitigation measures to reduce the probability of failures. This process is crucial to demonstrate compliance with certification requirements.

The Safety Assessment Process is described in ED-135/ARP4761A (Ref. 5)/ (Ref. 6), titled "*Guidelines and Methods for Conducting the Safety Assessment*", the document provides a structured framework for conducting safety assessments throughout the entire life-cycle of an airborne system, ensuring that potential hazards are identified, risks

are evaluated, and appropriate mitigation measures are implemented to achieve an acceptable level of risk.

The ED-135/ARP4761A guideline is strictly related with ED-79B/ARP4754B (Ref. 7)/ (Ref. 8), "*Guidelines for Development of Civil Aircraft and Systems*, where a structured approach to the development is presented, resulting in a "V"-shaped process that applies to software and hardware items.

The investigation of a trustworthy AI, has led EASA to an alternative "W"-shaped process of learning assurance for AI item that is intended to overlap the traditional non-AI component cycle and the safety assessment process (see Figure 1).

A link between AI Performance Metrics and the Safety Assessment as required by A-MOC-SA-01-7, is fundamental to obtain clear safety requirements that can guide the AI development from the beginning of the process, and should be systematically verified throughout the design implementation.

The Safety Assessment Process is the starting point of this work, which is based on concepts and definitions provided in documents ED-135/ARP4761A (Ref. 5)/ (Ref. 6) and ED-79B/ARP4754B (Ref. 7)/ (Ref. 8).

APPLICABILITY

Binary Classifier

The following investigation and results are limited to AI-based systems that include *Binary Classification Algorithms*.

The goal of a learning algorithm F, is to learn a function $f : X \rightarrow Y$ from an input space X to an output space Y.

An algorithm performs a *Classification* task if the output space Y is a discrete set, in other words, the algorithm gives as output a class $y \in Y$ for each analysed input $x \in X$.

The classifiers can be binary or multi-class, depending on the numbers of the classes: the binary classifiers have only two possible output classes, usually indicated as "*Positive*" or "*Negative*".

Focusing on the Binary case, it can be noted that each instance in the input set of samples, being classified by the model, can be evaluated in one of the following cases:

- **True Positive (TP)**: instance belongs to the class and is predicted as belonging to the class;
- **True Negative (TN)**: instance does not belong to the class and is predicted as not belonging to the class;
- **False Positive (FP)**: instance does not belong to the class and is predicted as belonging to the class;
- **False Negative (FN)**: instance belongs to the class and is predicted as not belonging to the class.

To evaluate the performance of the algorithm, the previous notions are represented in a Confusion Matrix, as shown in Figure 2.

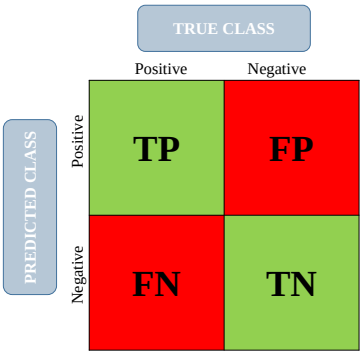

Figure 2. Confusion Matrix for a binary classifier.

The numbers of predictions can be seen from the matrix, allowing a detailed analysis of the classifier performance.

METHODOLOGY

Failure Modes

The first activity to be performed in the Safety Assessment is a Functional Hazard Assessment (FHA), that identifies potential failure conditions and assigns a severity based on their effects on the airplane or its occupants.

The term "*Failure Condition*" refers to a situation where a system's intended function is not being performed as intended. In the context of the safety assessment, the "*Failure Mode*" that led to a specific failure condition is also investigated.

In fact, a "*Failure Mode*" as reported in ED-135/ARP4761 (Ref. 5)/ (Ref. 6), is a specific way in which a system, equipment, hardware item, or piece-part may fail. Having a failure mode over another could result in a different failure condition.

It is important to note that traditionally a FM describes the specific malfunction or error that can lead to a failure condition, (examples for a mechanical component are *fracture, wear, ...* or for an electrical one: *open circuit, short circuit, ...*); but, when speaking of AI it is possible to have a failure condition even in the absence of a specific malfunction of the component.

In other words, it is possible that the AI system works as intended, and the predicted output may still be incorrect. This is due to the fact that the algorithm after the training can be exposed to new unseen input, that may lead the model to a false prediction.

So, when speaking about Machine Learning (ML) algorithm, it is possible to identify failure modes by classifying the type of output predicted by the model.

Given the previous consideration, the failure modes for a binary classification algorithm are:

- FP: The model is generating a false positive output;
- FN: The model is generating a false negative output.

The failure modes correspond to the red diagonal in Figure 2.

Performance Metric

In Machine Learning a performance metric represents a quantitative measures used to evaluate a model's performance, accuracy and effectiveness.

The metrics can be used with different scopes through the entire development; in fact, during the initial training phase, choosing a metric and eventually an associated target means to select a standard that is followed through the learning process and has the goal to optimize a specific behavior of the model.

Advancing to generalization and verification phases, the metrics are used to verify different performances, such as stability or robustness.

Taking into account that the complete evaluation of the model needs a general vision and therefore requires the computation of a plurality of metrics, in the preliminary phase it is preferable to have a clear target in mind, which translates in choosing one performance metric over the others.

The performance metrics for a Classification algorithm are various, however, EASA presented a list of performance metric in the Machine Learning Application Approval (MLEAP) (Ref. 3). Among all the classification metric presented in the document, three are directly linked to the identified failure modes of a binary classifier:

- **Precision or Positive Predictive Value (PPV)**:

$$PPV = \frac{TP}{TP+FP} \tag{1}$$

  This metric aims to optimize the correct positive predictions of the classifier, over the total positive prediction.

  A model with high Precision will generate a positive output only with high confidence, however, it should be noted that no guarantee will be provided on the False Negative predictions.

- **Recall or True Positive Rate (TPR)**:

$$TPR = \frac{TP}{TP+FN} \tag{2}$$

  This metric aims to optimize the correct positive predictions of the classifier, considering also the number of False Negative predictions.

  A model with high Recall will minimize the false Negative predictions, however, it should be noted that no guarantee will be provided on the False Positive predictions.

- **Accuracy (ACC)**:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

  This metric aims to optimize the correct predictions of the classifier over the total predictions.

Since the metrics reported above are based on the numbers of different predictions of the model (TP, TN, FP, FN), selecting a specific metric can change the probability of occurrence of a failure mode and consequently of the associated failure condition.

In other words, choosing to optimize a behavior of the algorithm through a performance metric minimizes the respective failure mode.

For this reason different metrics can be chosen, depending on the severity of the failure condition caused by the AI specific failure mode (FP or FN).

Precision/Recall Trade-off

Precision and Recall are performance metrics with opposite effect on the classification algorithm training:

- Precision aims to minimize the number of false positive prediction;
- Recall aims to minimize the number of false negative prediction.

The prediction of a particular class is made by the algorithm based on a soft score of the input, how the score is computed depends on the training, architecture and learning choices; not relevant for the scope of this discussion.

What is important, in fact, is that a class is chosen if the score is over a given threshold (t).

Taking as example the easiest case of a binary classifier:

$$\text{Predicted class } (x) = \begin{cases} Positive & \text{if } \widehat{f}(x) \geq t \\ Negative & \text{if } \widehat{f}(x) < t \end{cases} \tag{4}$$

The choice of t is not directly a Safety task, but it is a consequence of the performance metric selection.

The threshold is selected during the training phase based on the metric identified by the safety assessment: choosing an high t, means having high precision and low recall (low FP prediction and high FN), otherwise choosing a lower t brings to the opposite situation.

As the value of t is varied, different values for precision and recall are obtained, its influence on metrics can be seen in Figure 3.

The best classifiers are those that maintain a high precision as recall increases: such classifiers have a high area under the curve.

RESULTS

Metric Selection

The selection of the metric is based on the identified link between Failure Mode and Performance Metric.

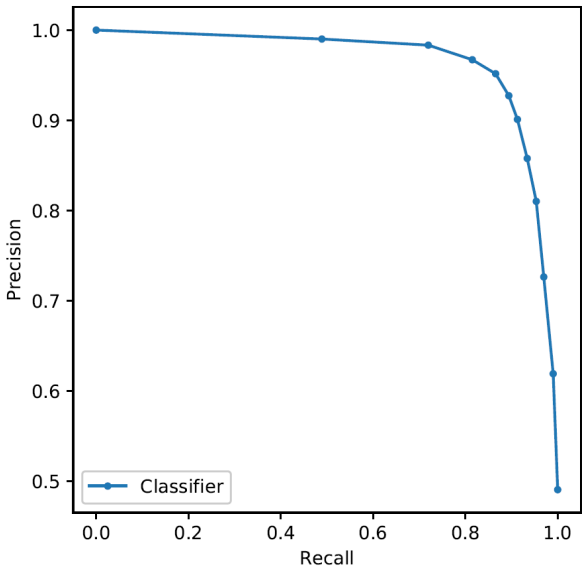

Figure 3. Precision-Recall curve for a binary classifier. The values are obtained by varying the threshold t. Credits: (Ref. 9)

The FM of the algorithm should be evaluated based on their effect on the system, so according on the associated failure condition.

The severity evaluation is performed as per typical safety assessment, following the ED-135/ARP4761A (Ref. 5)/ (Ref. 6) guideline, as follows:

- **Catastrophic:** Failure conditions which would result in multiple fatalities to occupants, fatalities or incapacitation to the flight crew, or result in loss of rotorcraft.
- **Hazardous:** Failure conditions which would reduce the capability of the rotorcraft or the ability of the crew to cope with adverse operating conditions to the extent that there would be: a large reduction in safety margins or functional capabilities; physical distress or excessive workload such that the flight crew's ability is impaired to where they could not be relied on to perform their tasks accurately or completely; or, possible serious or fatal injury to a passenger or a cabin crew member, excluding the flight crew.
- **Major:** Failure conditions which would reduce the capability of the rotorcraft or the ability of the crew to cope with adverse operating conditions to the extent that there would be, for example, a significant reduction in safety margins or functional capabilities, a significant increase in crew work load or in conditions impairing crew efficiency, physical distress to occupants, possibly including injuries, or physical discomfort to the flight crew.
- **Minor:** Failure conditions which would not significantly reduce rotorcraft safety, and which would involve crew actions that are well within their capabilities. Minor failure conditions may include, for example, a slight reduction in safety margins or functional capabilities, a slight increase in crew workload, such as, routine flight plan changes, or some physical discomfort to occupants.
- **No Safety Effect:** Failure Conditions that would have no effect on safety; for example, Failure Conditions that would not affect the operational capability of the rotorcraft or increase crew workload, however, they could result in an inconvenience to the occupants, excluding the flight crew.

The metric selected should be coherent with the severity classification.

For example, if the effect of a False Positive prediction has a high severity, the best decision is to optimize the Precision of the algorithm and the same consideration can be done with the False Negative and Recall.

However, when both the false prediction generate a severe effect on the system, the choice of the metric can be more difficult: this is due to the Precision/Recall trade-off.

Table 1. FM with different severity (one NSE)

AI FM	Class	FM Metric	Selected Metric
FN	MAJ	Recall	Recall
FP	NSE	Precision	

As explained in the previous paragraph, the two metrics, show an inverse relationship, where improving one of them worsens the other; consequently one of the two can be chosen only if the effect of the other is classified as No safety effect, an example is reported in Table .

When in the case of both failure modes contributing to a safety relevant failure, the choice must be Accuracy, which is the only metric that focuses on all the false predictions and tries to find the best trade-off between the numbers of FP and FN.

Examples of this cases are reported in Tables 2 and 3:

Table 2. FM with same severity

AI FM	Class	FM Metric	Selected Metric
FN	MAJ	Recall	Accuracy
FP	MAJ	Precision	

The possible cases discussed above, are also summarized in the flow chart reported in Figure 4.

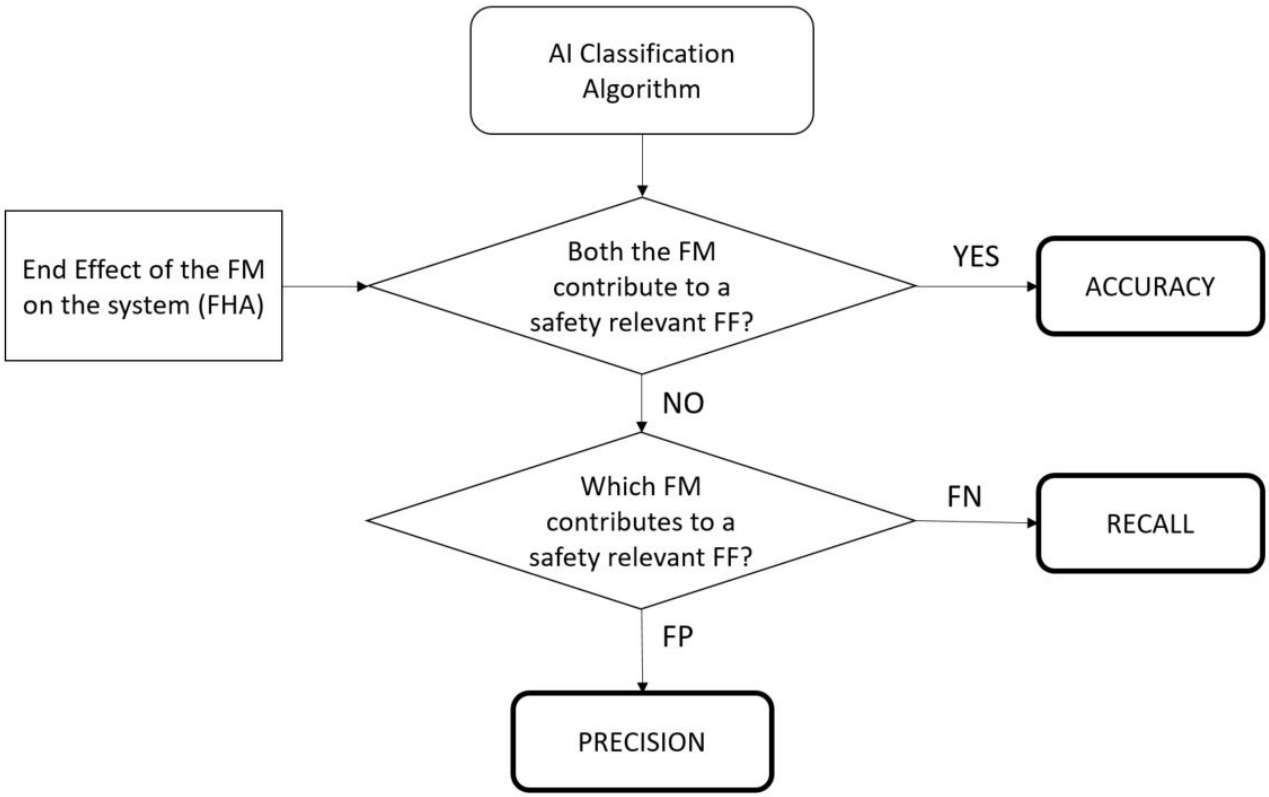

Figure 4. Flowchart: Metric selection

Table 3. FM with different severity

AI FM	Class	FM Metric	Selected Metric
FN	MAJ	Recall	Accuracy
FP	CAT	Precision	

This approach also allows to set a safety requirement on the selected metric: in this way it is possible to fix a safety discriminant target that can be used as reference to end the training phase.

The requirements take into account the severity of the functional failures involved, increasing for a higher severity class.

Possible safety requirements are reported in Table 4, the values are expressed as percentage over the total number of tested samples.

When the selected metric is "Accuracy" the safety requirement must be allocated according to the highest FM severity under investigation.

It should be noted that the quantitative requirement may vary due to specific consideration of different applications and, in general, it should guide the model into the final safety compliance, which will be obtained only at the end of the inference verification and integration and it is subject of the same safety standard of the non-AI process.

Table 4. Safety Requirements

FC Severity	**Precision**(%)	**Recall**(%)	**Accuracy**(%)
Catastrophic	95%	95%	93%
Hazardous	94%	94%	92%
Major	93%	93%	91%
Minor	92%	92%	90%

CONCLUSIONS

As a result of the discussion, the following conclusions are reported:

1. The link between performance metrics and safety assessment (as required by **A-MOC-SA-01-7**) has to be find directly in the failure modes of the algorithm, that for Binary Classifier are identified in the false predictions.

2. The choice of the right safety metric can be performed following the steps described in the Flowchart 4.

3. Selecting a specific safety metric allows also to set a quantitative requirement to follow during the AI training phase, allowing to include the safety discussion early in the process.

Author contact:

Claudia Ranieri - claudia.ranieri@leonardo.com

Pietro Ottavio Lecis - pietro.lecis@leonardo.com

Luca Castaldi - luca.castaldi@leonardo.com

REFERENCES

1. EASA. *Artificial Intelligence Roadmap: a human-centric approach to AI in aviation*, 2 2020.

2. EASA. *Concept Paper: First Usable guidance for Level 1&2 machine learning applications*, 2 2023.

3. MLEAP Consortium. Easa research – *Machine Learning Application Approval (MLEAP) final report*. Horizon europe research and innovation programme report, European Union Aviation Safety Agency, 5 2024.

4. EU. *Assessment List for Trustworthy AI*, 7 2020.

5. EUROCAE. *Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment*, 12 2023.

6. SAE ARP4761A. *Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment*, 12 2023.

7. EUROCAE. *Guidelines for development of civil aircraft and systems*, 12 2023.

8. SAE ARP4754B. *Guidelines for development of civil aircraft and systems*, 12 2023.

9. EASA and Daedalean. *Concepts of Design Assurance for Neural Networks (CoDANN) IPC extract*. Technical report, 3 2020.

CLASSIFICATION & COPYRIGHT