# MACHINE LEARNING ALGORITHM IN RAILWAY SAFETY CRITICAL APPLICATIONS : A NSA PERSPECTIVE

Julien Boucault (lead presenter)[1], Khaled Bahloul[1], Bruno Dember[1], Francis Dufour[1], Mohamed El Morabti[1], Achraf Gueznai[1], Antoine Hespel[1], Abdou Ngom[1], Jean-Baptiste Ripoll[1], Julien Roger[1], Jordane  Velu[1], Fabien Wilst[1] *

[1] *Établissement public de sécurité ferroviaire (EPSF), 60 rue de la Vallée, CS 11758, 80017 AMIENS CEDEX 1*

*Corresponding address : julien.boucault@securite-ferroviaire.fr.

## INTRODUCTION : AI, DATA AND ML

The term "artificial intelligence" (AI) is used with different meanings depending on the authors. Some authors, such as Laurent Alexandre[1] for example, use it very broadly by imagining an AI capable of surpassing human intelligence, even if he indicates that today the only existing AIs are limited. On the other hand, for Pierre Blanc[2] the term artificial intelligence is a non-sense, and he proposes to replace AI with algorithmic computing. AI would in fact only be algorithms that would accomplish a specific task.

So what are we really talking about ? Yann LeCun defines artificial intelligence as "the ability for a machine to perform tasks generally performed by animals and humans: perceive, reason and act. It is inseparable from the ability to learn, as observed in living beings. Artificial intelligence systems are nothing but very sophisticated electronic circuits and computer programs. But the storage and memory access capacities, the calculation speed and the learning capacities allow them to "abstract" the information contained in enormous amounts of data. »[3]. This definition emphasizes the link between intelligence and learning and focuses on machine learning (ML) algorithms. Figure 1 shows this link between the broad meaning of AI and ML algorithms.

The increase in the amount of data available has allowed the development of machine learning algorithms which have become more and more efficient. This development makes it possible to consider new applications for these algorithms, which sometimes require the production of new specific data. In the railway sector, machine learning algorithms may be used for systems providing assistance functions to a human operator, for example to carry out predictive maintenance, or for systems aimed at replacing a function performed by a human operator, e.g. reading trackside railway signals for driverless trains.

[1] Laurent ALEXANDRE, « La guerre des intelligences : Comment l'intelligence artificielle va révolutionner l'éducation », éditions J CLattès, 2017

[2] Pierre BLANC, « L'intelligence Artificielle expliquée à mon Boss », éditions Kawa, Novembre 2018

[3] Yann LE CUN, « Quand la machine apprend, La révolution des neurones artificiels et de l'apprentissage profond », éditions Odile Jacob, 2019

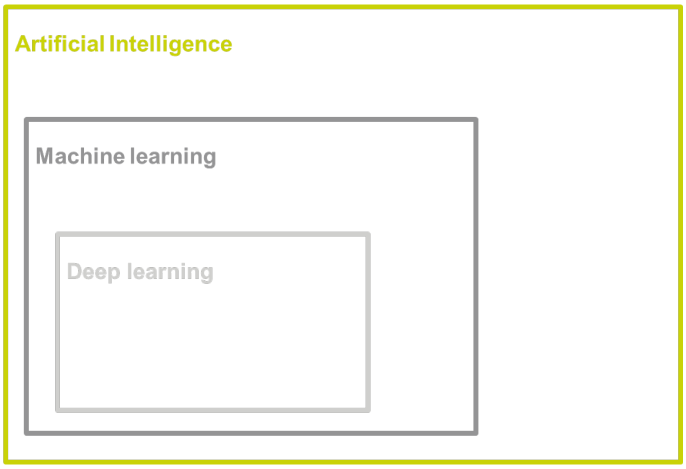

*Figure 1 - Artificial intelligence, machine learning and deep learning*

The different use cases can be classified in two main categories :

1. the case where the machine learning algorithm is directly involved in a security function (i.e., there is no systematic human intervention that would allow a critical look at what is produced by the machine learning algorithm). In this case, a decision involving safety is made without human intervention either by the machine learning algorithm or by a «classic» algorithm which has already been the subject of a safety demonstration but is based on information from a machine learning algorithm. The interpretation of the information carried by the trackside railway signals by a driverless train, for example, falls into this category ;
2. the case where the machine learning algorithm provides information to a human operator who will make the decision. In this case, what is produced by the machine learning algorithm will not directly lead to a safety action but the information it provides will guide the human operator's decision. In addition, an erroneous absence of information transmitted by the machine learning algorithm will not allow the human operator to take a critical look and react accordingly. A rail analysis system to detect cracks and propose to trigger a preventive maintenance action, for example, falls into this category.

In section 2 of this paper, we discuss what is specific about demonstrating the level of safety of sub-systems embedding machine learning algorithms compared to "classic" algorithms. In section 3, we present the holistic approach for demonstrating railway safety and how machine learning algorithms will fit into this approach. In section 4, we discuss the issues that have been identified from a national safety authority (NSA) point of view.

## AUTOMATIC METRO VS AUTONOMOUS TRAIN, WHAT IS SO SPECIFIC WITH MACHINE LEARNING ALGORITHM ?

Fully automatic metros have been operational since the 1980s with the opening of the Port Island Line in Kobe, Japan and the VAL (véhicule automatique léger) in Lille, France. One can wonder why it is such a revolution to do the same thing with autonomous trains. Part of the answer lies in the use of machine learning algorithms.

Fully automatic metros require a specific communication system between the metro and the track. That means that not only the metro needs to have specific equipment but also the infrastructure. On the contrary, the idea for autonomous trains is that they can run on existing tracks without modification. This implies, for example, that the train is able to "read" trackside railway signals. To do so, the train may use a perception subsystem embedding machine learning algorithms and these kinds of algorithms raise new questions regarding the demonstration of their safety levels.

Indeed, unlike the "classic" algorithms already authorized in rail transport and urban guided transport systems, the safety level of these machine learning algorithms cannot be demonstrated only by guaranteeing that the rules described by the algorithm are complete and correctly coded. Safety demonstrations of "standard" algorithms are based on the verification of a set of rules developed by human beings. If these rules are complete, correctly coded and correctly executed, the result obtained will be the expected result.

For machine learning algorithms, the correct application of the rules alone does not guarantee achieving the expected result. In these machine learning algorithms, the rules use parameters whose values are critical in obtaining the expected result. The value of each of these parameters is not fixed by the human designer of the algorithm but determined automatically by the algorithm itself during the learning phase. For these machine learning algorithms, demonstrating that the learning phase was correctly performed and that the algorithm was able to determine the "correct" value for each parameter is therefore critical.

Figure 2 below schematizes these differences between "classic" algorithms and machine learning algorithms.

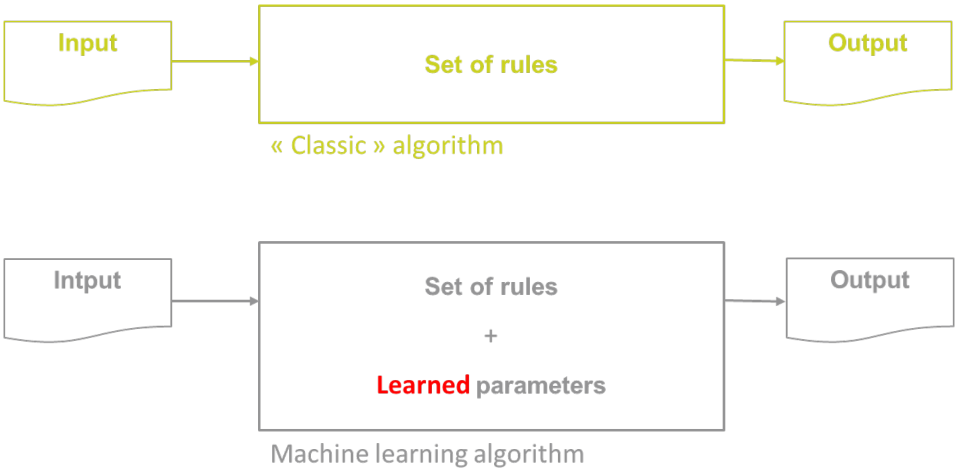

*Figure 2 – Difference between « classic » algorithm and machine learning algorithm*

Broadly, a machine learning algorithm must therefore learn "on its own" to perform the task assigned to it before it can be used operationally, in the sense that it must determine the optimal value for each of its parameters. To do this, it relies on a learning database that must be adapted to the task it has to perform. The three main types of machine learning are:

- supervised learning;
- unsupervised learning;
- reinforcement learning.

In the case of supervised learning, all input data in the learning database is labeled with the expected result. The automatic learning algorithm will therefore apply its internal model to all input data that is in the learning base, compare the result obtained with the expected result and, if the result obtained is different to the expected result, modify its internal parameters. It is for example this type of learning that is used for image classification.

In the case of unsupervised learning, the machine learning algorithm will itself determine the characteristics corresponding to the different classes from the data in the learning database. This type of learning can be used to form groups of elements with common characteristics (clustering).

In the case of reinforcement learning, the operating principle is close to that of supervised learning but with an additional time dimension because the algorithm only knows if it has reached the objective after several operations. The algorithm must therefore estimate after each operation whether it has reached its objective in order to be able to adjust its internal parameters. It is this type of learning that has, for example, been used to teach an algorithm to play the game of Go.

In this paper, we will focus only on **machine learning algorithms trained with supervised learning designed to perform a perception task**.

To conclude this section, like “classic” algorithms, the choice and the coding of the set of rules are important for machine learning algorithms. However, unlike “classic” algorithms, the safety level of the algorithm cannot be assessed only by regarding the set of rules. The learning phase, that encompasses the learning database and the learning techniques, is also critical in establishing confidence in the machine learning algorithm. The constitution of the learning database is critical for the machine learning algorithm to be effective. The training database must be completely representative of the problem. In particular, it is necessary to be careful with regard to the biases that could come with this learning database and which could be reproduced by the algorithm. The choice of the learning technique is also critical as it will allow or not to reach an optimum for the desired task.

Note : in this note, we will use the term “accuracy” to describe the output from the machine learning algorithm. An output data will be considered accurate if: i) it corresponds to what is expected ; ii) it is transmitted within a period compatible with its use. To take an example outside of the rail field, the output of an image classification algorithm will be considered accurate if, when a photo of a cat is presented as input, the algorithm indicates that the most likely class of the image is “cat” within the allotted time.

## DEMONSTRATING SAFETY, THE HOLISTIC APPROACH OF RAILWAY

The fundamental principle of railway safety is to maintain the overall safety level of the railway system over time. Because of this fundamental principle you have to ensure that the introduction of a new subsystem or the modification of an existing system does not downgrade the overall safety level of the system. This principle of non-regression of the safety level is included in the European regulations, in particular directive (EU) 2016/797 of the European Parliament and of the Council of 11 May 2016 on the interoperability of the rail system within the European Union and directive (EU) 2016/798 of the European Parliament and of the Council of 11 May 2016 on railway safety. For the interoperable rail system, the compliance with this principle relies on two pillars: on the one hand, the authorization of fixed installations, vehicles and rail operators, and on the other hand, the monitoring of how these fixed installations and vehicles are operated, maintained and modified as well as the feedback on safety events.

Authorizations in the railway sector relate either to vehicles or to fixed installations. There will therefore be no authorization of equipment encompassing a machine learning algorithm as such. The authorization will cover only the entire "vehicle" system or the entire "fixed installation" system. It should be noted that, in the context of this note, the term system refers to the scope of the authorization (vehicle or fixed installations). As a reminder, the authorization also takes into account the safe integration of this system into the rail system within the meaning of Annex I of Directive (EU) 2016/797.

The description of this "vehicle" or "fixed installation" system is therefore essential. It should give an overview of how the system works and describe how the subsystem encompassing the machine learning algorithm fits into this system and how it contributes to the functions of this system. It must also explain the operating and maintenance conditions of this system with particular attention to human and organizational factors. The objective of this description of the system vis-à-vis the subsystem encompassing the machine learning algorithm is to identify :

- the input data of the subsystem encompassing the machine learning algorithm;
- the output data of the subsystem encompassing the machine learning algorithm;
- the way in which output data is used by the system under consideration and, where applicable, by the human operator to fulfil the expected functions.

The figure 3 below represents these elements

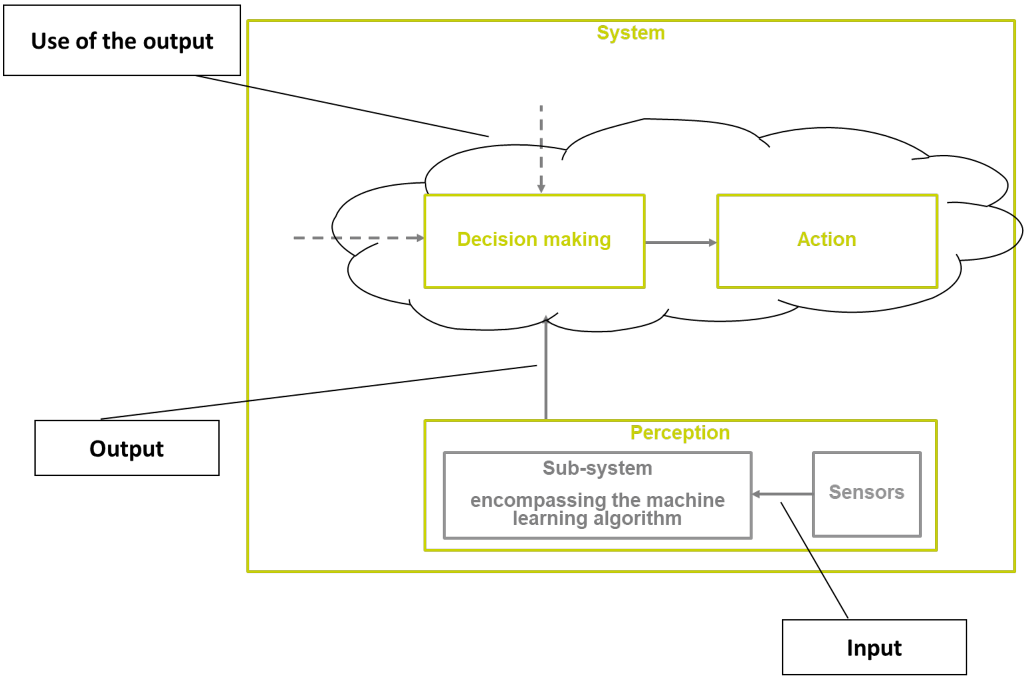

*Figure 3 - Holistic approach with ML algorithm*

In order to illustrate the importance of the holistic railway approach, we will consider the simplified case of a subsystem whose function is to detect and identify obstacles located at track level in the gauge of the train. The input data for the obstacle detection subsystem encompassing the machine learning algorithm is, in this example, the images of a camera filming what is in front of the train. At output, the subsystem transmits a prediction of what is in front of the train: no obstacle, human, animal, tree, rock, smoke. Depending on the output, the train will react as follows (considering that all the other parameters remain unchanged):

- no obstacle: no change in the traction setting ;
- human, animal, tree, rock: triggering of emergency braking ;
- smoke: traction is set to reach the maximum permitted speed.

In this example, we can see that it is not the accuracy of every prediction from the subsystem embedding the machine learning algorithm that will be useful for the safety demonstration but rather the accuracy of its ability to detect if there is an obstacle or not in front of the train, as well as the accuracy of its ability to detect if that obstacle in front of the train is smoke. These levels of accuracy will be integrated into the safety demonstration to ensure that the risk of collision and the risk of fire are covered. In other words, it is not possible to assess the sub-system embedding the machine learning algorithm without taking into account the whole system.

This systemic approach leads to question the notion of area of use. Today a train is already authorized within an area of use regarding technical compatibility. With systems using a machine learning algorithm, this notion may have to be generalized. The area of use in nominal and degraded mode will specify the limits of use of the system and therefore of the subsystem embedding the machine learning algorithm. This area of use may indicate, in particular, the maximum operating speed, the range of light conditions (night / strong sunshine), the range of climatic conditions (snow, fog, etc.), and any user constraints placed on the operator and/or the maintainer. The safety demonstration will have to provide the guarantee that the system is not used outside its area of use.

In the case of a machine learning algorithm assisting a human operator, the systemic approach should also include the interface, in a general sense, between the system and the human operator. The description of this interface should make it possible to exhaustively map and qualify the elements made available to the human operator so that they can understand and take a critical look at the output transmitted by the system, associated with a level of "trust". It should also make it possible to understand the conditions under which the human operator will be required to interact with the system in a nominal and degraded operating situation. Particular attention should be paid to human and organizational factors.

## CHALLENGES AND ISSUES AHEAD

In order to authorize a system (e.g. a train or a railway line) with a subsystem embedding a machine learning algorithm with the high level of safety expected for any railway application, 3 main challenges have been identified from the French National Safety Authority (NSA) perspective :

1. the subsystem embedding the machine learning algorithm must be **certifiable** so that it can be taken into account in the system safety demonstration (this includes the operation and maintenance phases to ensure that the level of safety is maintained over time) ;
2. the subsystem embedding the machine learning algorithm must be **auditable** ;

3. once commissioned, the action of the subsystem embedding the machine learning algorithm must be **reproducible**.

Some of these challenges seem within arm's reach, others may require additional research and development.

At this stage, 4 requirements have been derived from these three challenges and for each requirement several issues have been identified.

**<u>Requirement #1 :</u>** A level of accuracy for output data of a subsystem embedding a machine learning algorithm must be determined and demonstrated for a given area of use

To meet this requirement, two main sources of error appear to be identified :

- the first source of error is specific to machine learning algorithms. It is specified below ;
- the second source of error concerns calculation errors due either to an error from the hardware used or to a default in the programming language used. This second source is common to all software. It is not detailed below but is included in the questions that need to be addressed.

The figure 4 below breaks down, theoretically given that some curves may not be known, the error specific to machine learning algorithms. It takes, as an example, the case of an algorithm for detecting a light signal with a "proceed" aspect.

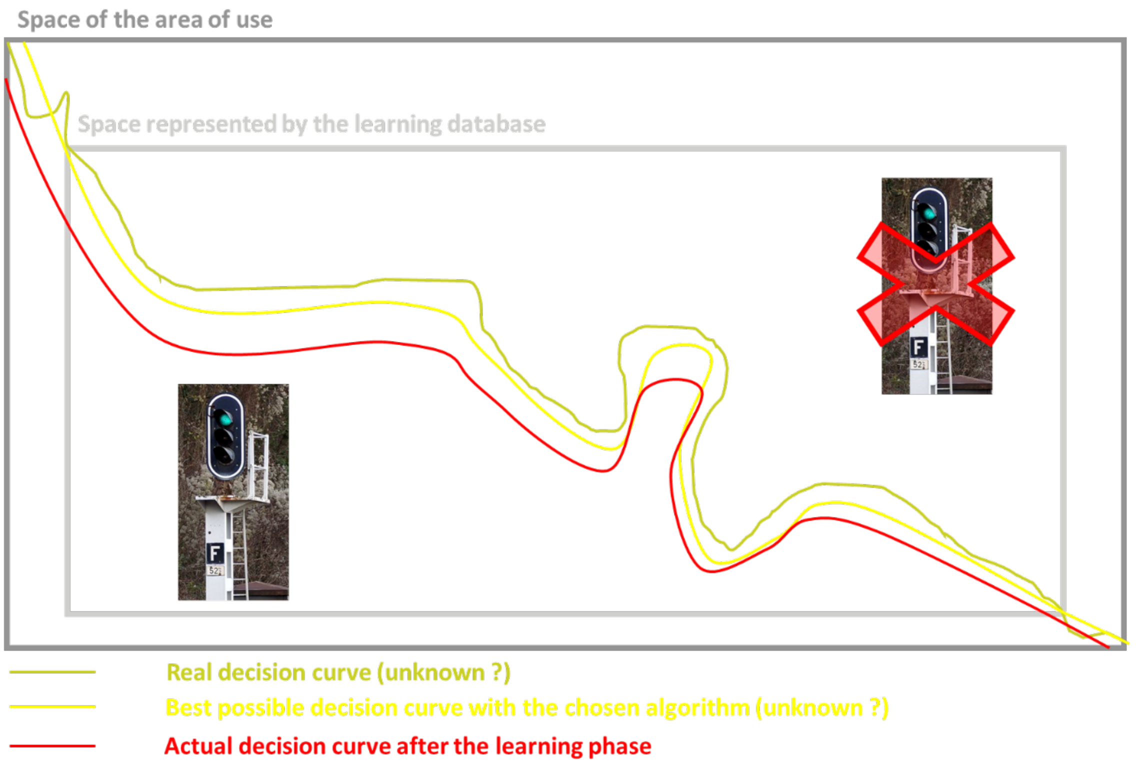
Space of the area of use
Space represented by the learning database
F
521
Real decision curve (unknown ?)
Best possible decision curve with the chosen algorithm (unknown ?)
Actual decision curve after the learning phase

*Figure 4 - Error specific to machine learning algorithms*

An erroneous output result from the subsystem embedding the machine learning algorithm could therefore be due to:

- the choice of an automatic learning algorithm which does not make it possible to stick exactly to the real boundary between all the signals with a “proceed” aspect and all the others. This is represented on the figure 4 by the difference between the green curve and the yellow curve ;
- a learning database that is not representative of all the situations that may be encountered (space of the area of use), which does not allow the subsystem to be efficient for some scenarios ;
- the implementation of learning which does not make it possible to achieve the best possible curve given the chosen algorithm. This particularly applies to deep neural networks for which the loss function is not convex and therefore only a local minimum can be reached during learning.

These last two points make up the gap between the yellow curve and the red curve.

Given the current state of the art, it seems difficult to envisage a formal demonstration of the accuracy of a subsystem embedding a machine learning algorithm. Therefore, for the following part of this section, we assume there will be a statistical dimension in the safety demonstration strategy used.

Considering these elements, this first requirement leads to the following questions to be addressed :

Question Q1.1 : How to demonstrate that the space represented by the training database is representative of the real space of the area of use in which the subsystem will operate ?

Question Q1.2 : How to design a machine learning algorithm in such a way as to demonstrate that it is suitable for the desired function, i.e. the output is based on the correct characteristics of the input ?

Question Q1.3 : How to train a machine learning algorithm in such a way as to demonstrate that the learning optimum has been reached ? Is this demonstration necessary for certification ?

Question Q1.4 : How to test the performance of the subsystem embedding the machine learning algorithm? In particular, what volume of tests must be carried out to be statistically significant according to the safety objective that has to be achieved ?

Question Q1.5 : Are the actual programming languages for machine learning sufficiently robust ?

Question Q1.6 : Should the hardware architecture used for testing and operating the subsystem be safe (for example with a 2 out of 3 architecture) ?

**Requirement #2 :** To ensure that the accuracy of the subsystem embedding a machine learning algorithm is maintained over time, a monitoring process will have to be put in place.

Over time, the accuracy of the machine learning algorithm may decrease either because the quality of the data transmitted may change (sensor aging, new sensors with different sensitivity, change of the sensor environment – for example, the tint of the windshield behind which a camera is located) or because the “real space” evolves beyond the algorithm’s generalization capabilities (for example with the introduction of a new chassis for railway lights).

Given these considerations, this second requirement raises the following issues :

Question Q2.1 : What follow-up needs to be put in place to monitor the evolution of the environment in which the machine learning algorithm operates ? Is the process used to qualify a modification of the railway system necessary and sufficient ?

Question Q2.2: Is it possible to continuously monitor the evolution of “real space”?

Question Q2.3: Is a periodic review of the certification of the subsystem including the machine learning algorithm required ?

**Requirement #3 :** In the event of an incident or accident, the action of the subsystem embedding the machine learning algorithm shall be reproducible.

The main objective of this requirement is to be able to reproduce what has been achieved by the subsystem embedding the machine learning algorithm in order to determine whether it plays a role in the occurrence of the hazardous situation. If the output was not adequate, it would allow, with the auditability of the subsystem, to understand what failed.

For that, two conditions have been identified : i) to be able to use the subsystem in the same state as it was in at the time of the event ; ii) to have a recording of the input data at the time of the event.

Given these considerations, this third requirement raises the following issues :

Question Q3.1: Is it necessary and sufficient to freeze learning before commissioning ?

Question Q3.2: What relevant input data should be recorded in order to reproduce an event ? (algorithm input data, raw sensor data, partially pre-processed data, etc.)

Question Q3.3: Like the juridical recording unit (JRU), how could this relevant data be stored securely?

**Requirement #4 :** Human operators will need to have the skills and tools to critically examine the output of the subsystem embedding the machine learning algorithm.

To meet this requirement, you need to consider the time available to perform a critical analysis of the output of the subsystem embedding the machine learning algorithm. If an in-depth analysis is performed and you have time to do so, you can rely on several people with different skillsets, using different tools, including skills and tools from a specialized service provider. On the contrary, if the analysis is performed within a short time-frame, you can only rely on the skills of the operators (one person or a small team) and the tools directly available to determine whether the output is accurate, not accurate or is in doubt.

Given these considerations, this fourth requirement raises the following issues:

Question Q4.1: What training should be provided to teams performing in-depth analysis, on the one hand, and to operators dealing with subsystem embedding machine learning algorithms on the other hand, to enable them to understand the output of the subsystem embedding a machine learning algorithm, in a given case (local explicability) and therefore to make a critical judgement ?

Question Q4.2: Is the subsystem embedding the machine learning algorithm able to safely assess whether it is being used properly in its area of use (for example, with sufficient brightness or not too much fog).

Question Q4.3: To make this critical judgment, do we need tools which are safe and unrelated to machine learning algorithms ? If yes, what would be the useful tools available for each case ?

To conclude, authorizing machine learning algorithms for railway safety critical applications raises new issues that will need to be addressed. Current research in different fields of AI, especially explainable AI (XAI) and R and D railway projects may provide some answers in the future.

**Keywords**: safety ; artificial intelligence ; machine learning ; autonomous train ; predictive maintenance ; national safety authority