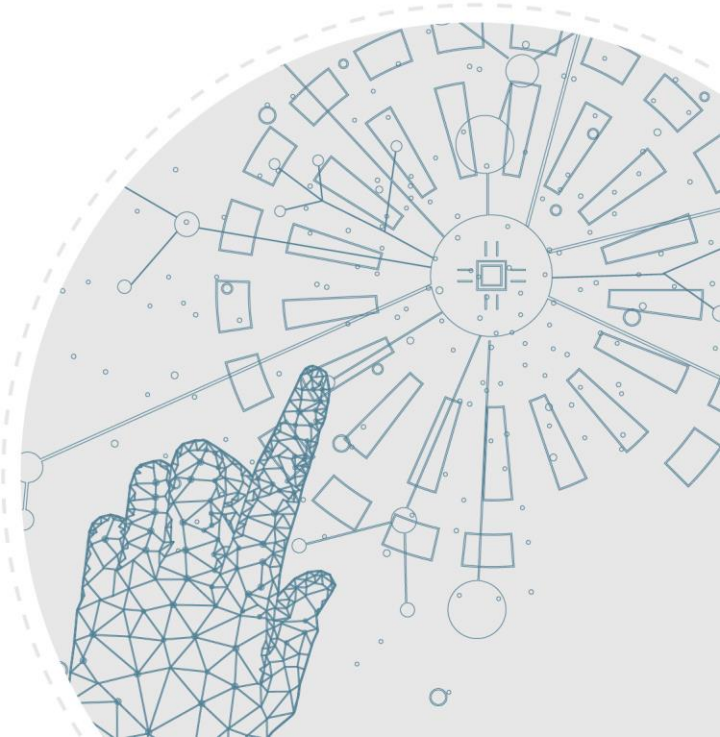
SYS DOCT 001 V1

Doctrine

Artificial intelligence

Conditions for authorising the use of systems incorporating machine learning algorithms

23 October 2023



Foreword

This note was drafted by an internal working group of *Établissement Public de Sécurité Ferroviaire* (EPSF) and reflects solely the position of EPSF.

It represents an initial collection of thoughts on the conditions under which inference models derived from machine learning algorithms could be authorised in safety-critical railway system applications.

EPSF has no expertise in the field of machine learning. This note therefore lays out our current understanding of this subject, in the aim of fuelling debate and enabling informed dialogue between machine learning experts and railway safety experts.

Table of contents

1. Introduction

The availability of increasing amounts of data has led to an increase in the use of machine learning algorithms, which have become increasingly effective. This development foreshadows new applications for these algorithms and consequently for the inference models thus derived. The rail sector envisages using inference models derived from machine learning algorithms in systems providing human operator assistance functions, for example in performing predictive maintenance, or in systems whose aim is to replace a function performed by a human operator, for example reading trackside signals to enable driverless train operation in autonomous or semi-autonomous mode.

However, these inference models derived from machine learning algorithms raise new issues when it comes to demonstrating their safety level. Unlike the “classic” algorithms already authorised for use in railway and urban guided transport systems, the safety level of these inference models derived from machine learning algorithms cannot be demonstrated solely by guaranteeing that the algorithmic rules are complete and correctly coded. “Classic” algorithms, as used in ”expert systems”, 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 will agree with expectations.

For these inference models derived from machine learning algorithms, it is crucial to demonstrate that the learning phase has been sound, enabling the algorithm to determine the “right'” value for each parameter.

In inference models derived from machine learning algorithms, correct application of the rules alone is not guaranteed to produce the expected result. In these machine learning based inference models, the rules incorporate parameters that may or may not produce the right result depending on their value. The value of each of these parameters is not set by the algorithm’s human designer of but determined automatically by the algorithm during the learning phase. Therefore for these inference models derived from machine learning algorithms, it is crucial to demonstrate that the learning phase has been sound, enabling the algorithm to determine the “right” value for each parameter.

Figures 1 and 2 below illustrate the differences in the design and operation of ”classic” algorithms (in green) compared with inference models derived from machine learning algorithms (in grey).

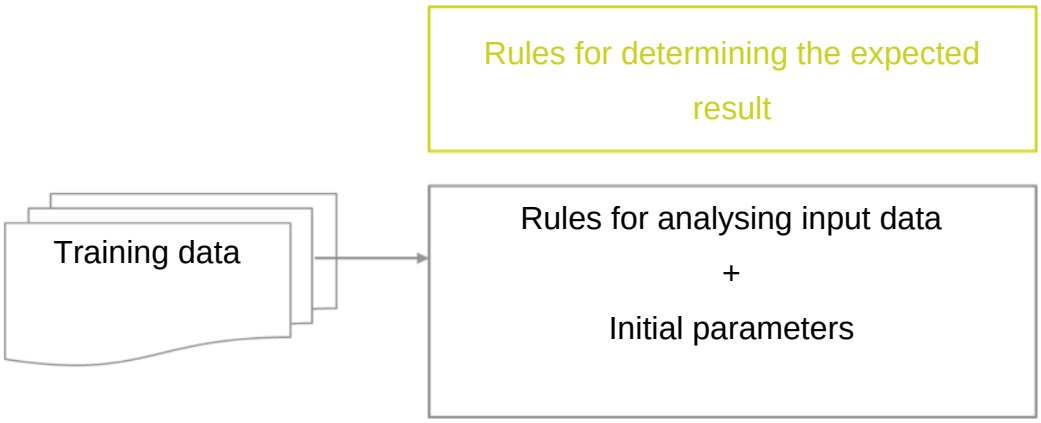


Figure 1 - "Classic" algorithm and machine learning algorithm at the design stage

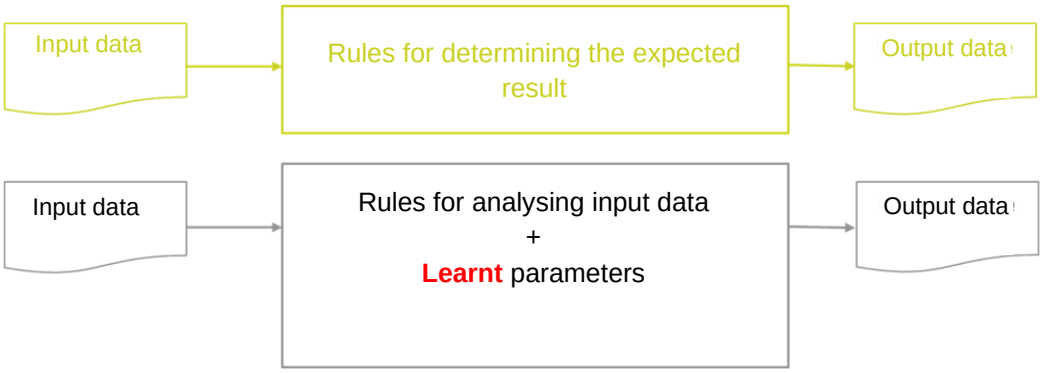


Figure 2 - "Classic" algorithm and machine learning algorithm at the operational stage

Schematically speaking, therefore, an inference model derived from machine learning algorithms must learn to perform the task assigned to it “on its own” before it can be used operationally, insofar as the learning algorithm must determine the optimum value for each of the model's parameters. To do so, the algorithm relies on a training database tailored to the task it has to perform (for example, to learn to recognise railway signals, the database needs to contain images of railway signals). Using the data in this database, the algorithm will determine a value for each of the model’s parameters in order to determine the result being sought. The main types of machine learning include:

- supervised learning,
- unsupervised learning,
- reinforcement learning.

The learning phase must include a strategy covering all actions involved in calibrating all the model's parameters, which could mean using one or more of these main learning schemes.

In the case of supervised learning, each entry in the training database is labelled with the expected result. The machine learning algorithm therefore applies its internal model to each entry in the training database, compares the result it obtains with the expected result, and if the results are different, modifies its internal parameters. A practical example of this type of learning is image classification.

In the case of unsupervised learning, the learning algorithm will itself determine the characteristics corresponding to the different classes using the data in the training database. The idea is that the algorithm will discover the structures underlying this unlabelled data. This type of learning is therefore used to create groups of elements with common characteristics (clustering).

In the case of reinforcement learning, the operating principle is to identify the actions to be taken in order to optimise a quantitative reward over time. It takes several actions for the algorithm to know whether it has achieved the objective (optimising the reward). To determine its internal parameters, it performs experiments, calculating the reward each time. A practical example of this type of learning has been teaching an algorithm to play Go.

Building the training database for the first two types of learning, as well as formulating the reward, are therefore especially important if the inference model resulting from the machine learning algorithm is to

be effective. For both supervised and unsupervised learning, the training database must be properly representative of the problem. We need to take particular care not to introduce any biases when creating the database, as these could be reproduced in the inference model derived from the algorithm.

This note therefore sets out to present EPSF's current thinking on the conditions under which inference models derived from machine learning algorithms could be used in railway applications. It first (section 2) reviews the principles implemented in the railway world to guarantee an ongoing high safety level. It then (section 3) looks at selected primary use cases and the main issues when using these inference models derived from machine learning algorithms in railway applications. Third and last (sections 4 & 5), it looks at the descriptive elements that would be expected in a railway application authorisation submission file and questions the requirements to be met in the case of an inference model derived from a supervised automatic learning algorithm used for a perception function.

2. Context: authorisations, safety monitoring, and demonstration methods in the railway sector

The fundamental principle of railway safety is non-regression, along with the ongoing preservation of the railway system's overall safety level. This fundamental principle notably involves demonstrating that the introduction of a new subsystem or the modification of an existing system will not regress the overall safety level of the system. This principle of non-regression of the safety level is enshrined in European regulations as well as in French decree no. 2019-525 on interoperable railways and decree no. 2022-664 on local railways. This principle is sometimes referred to as *Globalement au moins équivalent* (globally at least equivalent) or by its initials "GAME". At design level, in the scope of an explicit demonstration, non-regression is assessed for each hazard situation whose rate of occurrence must be less than a threshold value determined according to the severity of the accident the hazard situation could cause (occurrence/severity matrix).

Furthermore, European Commission regulation (EU) No 402/2013 of 30th April, 2013 on the *common safety method for risk evaluation and assessment* sets harmonised design objectives for electrical, electronic, and programmable electronic technical systems in point 2.5.5 of its Annex, as follows:

a) when failure could plausibly and directly give rise to a *catastrophic* accident, the associated risk need not be further reduced if it has been established that the probability of the function's failure is extremely low (10^{-9}/h)
b) when failure could plausibly and directly give rise to a *critical* accident, the associated risk need not be further reduced if it has been established that the probability of the function's failure is low (10^{-7}/h).

However, point 2.5.11 of the same Annex states that *"[...] if for a given hazard the applicant can demonstrate that the existing safety level in the Member State where the system is applied can be maintained with a less stringent design objective than the EU harmonised design objective, then this less stringent objective may be used instead of the harmonised design objective."* The non-regression objective is therefore the very minimum that must be achieved.

For interoperable and local rail systems, compliance with this principle of non-regression along with ongoing preservation of safety levels is founded on two pillars: the authorisation of fixed installations, vehicles, and rail operators, and the monitoring of and acting on feedback.

2.1. Authorisations

For authorisation purposes, risk analysis is the common denominator between authorisation applicants and the safety authorities. This analysis is what will enable the identification of risks and the safety constraints to deal with them.

Risk is defined in Regulation (EU) No 402/2013 as "*the frequency of occurrence of accidents and incidents causing harm (due to a hazard) and the degree of severity of that harm"*, while risk analysis is defined as "the systematic use of all available information to identify hazards and estimate the risk".

The European regulatory framework for risk analysis is set out in Regulation No 402/2013 applying to the railway sector. It notably sets out three risk acceptance principles:

- compliance with best practice rules, primarily regulatory specifications and standards, said compliance being acknowledged as a guarantee of acceptable safety levels
- comparison with a similar system, used in the same conditions, insofar as said system has demonstrated through its operation that it guarantees an acceptable safety level
- explicit risk assessment, which is used when the first two principles cannot be applied and is based on operational safety techniques This principle is especially useful in the case of disruptive innovations, for which no best practice has yet been defined and no similar systems exist. With more and more technologies being introduced, there is an increasing trend in the use of this principle.

It should be noted that in addition to risk analysis, certain best practice rules are mandatory in order to guarantee the interoperability of the rail system. These are the technical specifications for interoperability and the national rules. Applicants for authorisation must therefore ensure that their project complies with interoperability and safety rules and that all the risks associated with the project are covered. To date, there are no requirements relating to machine learning in French interoperability specifications or national rules.

2.2. Safety audits and monitoring

Once the vehicles, fixed installations, and railway operators have been given authorisation, the second pillar of railway safety is feedback in the broad sense of the term, which includes both the analysis of safety occurrences and the audits carried out by the operator (railway undertaking (RU), infrastructure manager (IM)), or by EPSF, enabling ongoing verification of the effectiveness of risk mitigation measures and their effective application.

To this particular end, French govt. order of 4th January, 2016 *on the nomenclature used to classify rail safety events* requires all rail operators to notify the EPSF of safety events occurring during the operation of their services, together with analysis points congruent with the seriousness of the events. This represents an information repository to which more than 20,000 events are added each year, classified according to a taxonomy that helps structure the collected data.

The events identified are analysed in greater or lesser detail depending on their learning potential. This analysis can be carried out by various entities: railway operators, EPSF, BEA-TT.

The systems and subsystems brought into service must therefore be capable of supporting these analysis.

For a system comprising an inference model derived from a machine learning algorithm, the following section discusses the issues associated with the aforementioned two main pillars (pre-operational risk analysis and feedback during the operational phase) in various use cases.

3. Different cases and consideration of their issues

For authorisation purposes, risk analysis is the common denominator between authorisation applicants and the safety authorities.

The use of inference models derived from machine learning algorithms for railway applications raises different issues depending on the use of the algorithm. In the remainder of this document, we will distinguish between the following two types of use:

- That where the inference model derived from a machine learning algorithm ("inference model" below) is **directly involved in a safety function** (i.e. there is no systematic human intervention that would enable a critical eye to be cast on the inference model's output). In this case, a safety-related decision is made without human intervention either by an inference model derived from a machine learning algorithm or by a "classic" algorithm that has already been demonstrated to be safe but which relies on information from an inference model derived from a machine learning algorithm. For instance, the interpretation of information gleaned by a driverless train from trackside signalling falls into this category;

- That where an inference model's output **provides information to a human operator** who makes the decision. In this case, the inference model's output will not lead directly to a given safety action but the information it provides will guide the human operator's decision. Furthermore, if information is expected from the inference model but not provided, the human operator will not be able to take a critical look at it and react accordingly. A system that analyses rails to detect the formation of cracks and suggests preventive maintenance action would for example fall into this category.

The issues associated with these two types of use are set out in sections 3.1 and 3.2.

In addition, both cases above will have the common issue of implementing essential feedback once the system has been authorised and commissioned. This reproducibility and explicability issue is covered in more detail in section 3.3.

3.1. Inference model derived from a machine learning algorithm directly involved in a safety action

If the inference model is directly involved in a safety action, the system comprising this model must have a safety level compatible with the seriousness of the accident or incident it covers. This means that the inference model must play no part in elevating the occurrence frequency of the considered hazard

situation above a given level. However, the demonstration that this safety level has been achieved will relate to the system as a whole and not just the inference model.

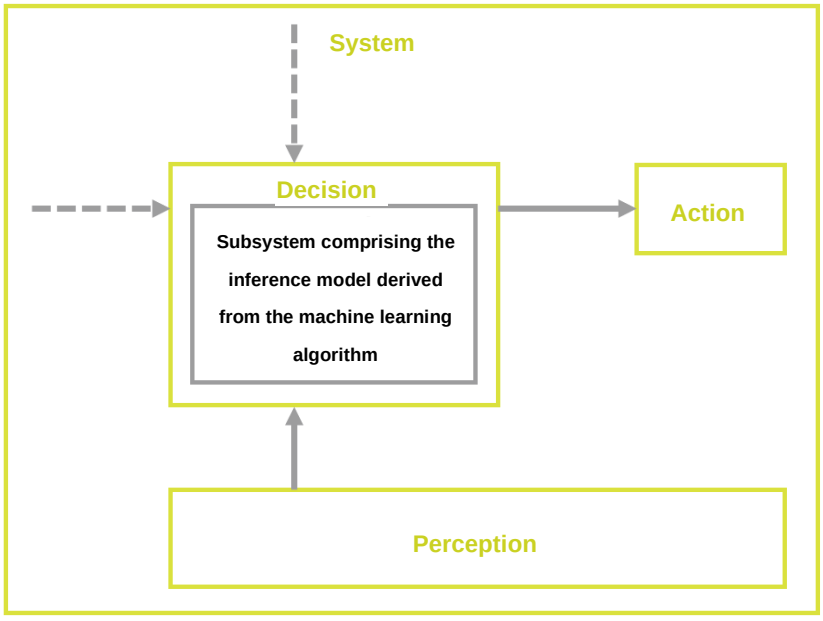
Broadly speaking, there are two main scenarios, each with its own challenges:

1. the subsystem comprising the inference model provides safety data essential to the control program's decision, and it is the only subsystem to provide this data;
2. the subsystem comprising the inference model analyses the data transmitted to it and makes the safety decision on its own.

Note: In the remainder of this note, we will use the term “accuracy” to qualify the output data from the machine learning algorithm. Output data will be considered accurate if i) it corresponds to what is expected, and ii) it is transmitted within a timeframe compatible with its use. To take a non-railway example, the output of an image classification algorithm will be considered accurate if, when presented with an image of a cat as input, the algorithm indicates within the given timeframe that the most likely class for the image is “cat”. This term has been chosen to avoid any confusion with the terms usually used in operational safety.

In the first case above, the data must be supplied with an adequate accuracy level. The subsystem comprising the inference model must therefore be analysed to assess the accuracy of the information transmitted. This accuracy level must be incorporated in the system’s safety demonstration (insofar as it furthers completion of the risk management process in accordance with Regulation (EU) No. 402/2013) for the frequency of hazard situations to be guaranteed. This safety demonstration will also take into account the other data on which the subsystem making the decision is based (especially in the case of sensor fusion)

In the second case above, the decision made by the subsystem comprising the inference model is a decision with a direct impact on safety, and this subsystem must therefore be analysed to assess the correctness of the decision made. As in the previous case, this accuracy level must be incorporated in the system’s safety demonstration for the frequency of hazard situations to be guaranteed.

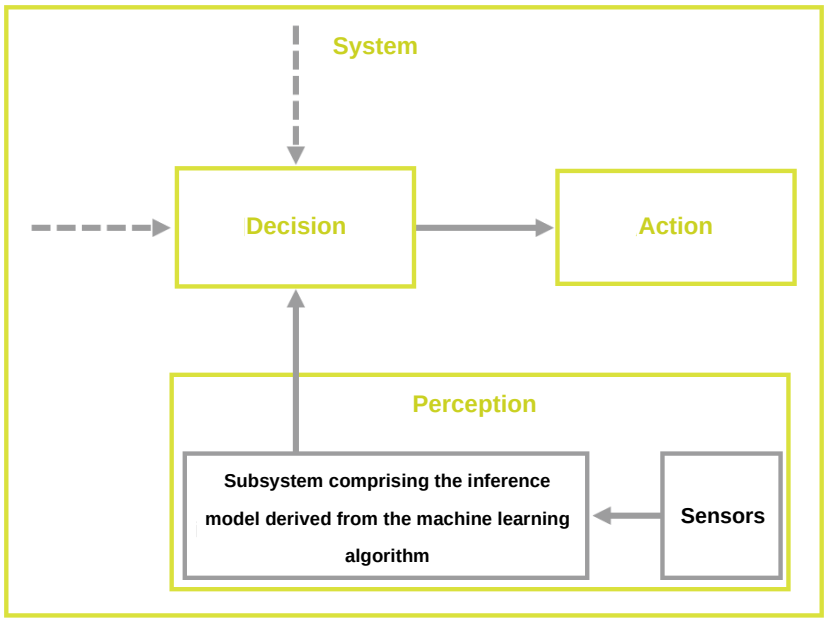


In both cases above, possible obsolescence of the training data must be taken into account. If the model parameters have been set with respect to a training database whose entries no longer match the data being input to the model, the training will no longer be valid. This could, for example, be the case when reading trackside signals if the signal gantries in the training database no longer match the gantries used on railway lines.

IN SUMMARY:

➔ **For these two cases, the challenge is to determine the accuracy level of the subsystem comprising the inference model so that it can be integrated into the system's full safety demonstration. This implies being able to determine, in the usage scope of the system, a maximum probability for each occurrence of erroneous output data liable to lead to a hazard situation. This accuracy level will be guaranteed as long as the training database remains valid.**

Note: The remainder of this note will not go into further detail on the second of these cases, whose specific features require more in-depth study. Determining the safety authorisation conditions for the inference models used in the first scenario seems to us to be an interesting first step.



3.2. Inference model derived from a machine learning algorithm assisting a human operator in decision making

In the case whereby the inference model provides data that will be analysed by a human operator, the issue is twofold in that it concerns both the subsystem comprising the inference model and the human operator. For this scenario, this document assumes that independently of the subsystem comprising the inference model, the human operator has the means to make a critical judgement on the data transmitted. However, it is considered that the human operator is unable to make a critical judgement in

the event of a false negative, i.e., data that should have been transmitted to the human operator but was not.

With regard to the subsystem comprising the inference model, for certain hazard situations, the objective could well be that the data it transmits can be understood and analysed critically by the human operator. At the very least, this would mean that the expected accuracy (not the guaranteed accuracy mentioned in section 3.1.) of the machine learning algorithm should be specified as well as its usage scope. Depending on the models used, the data transmitted may be accompanied by a reliability level or confidence interval. This also means that the human operator must be able to interpret the results provided (local explicability[1]).

As far as the human operator is concerned, in the aforementioned hazard situation scenarios, the aim is to enable critical judgement of the data transmitted by the subsystem comprising the inference model. This involves training the operator to understand the results of the machine learning algorithm. This could include training on the main operating principles and limitations of the algorithm itself, as well as training on the tools that can help interpret the data generated by the algorithm. This also implies that the operator be provided with other means to confirm or reject an analysis produced by the machine learning algorithm. Finally, that means implementing measures to remind the human operator that the system, although efficient, is not infallible.

IN SUMMARY:

➔ **In this case, for certain hazard situations, the challenge is for the human operator to be aware of the fallible nature of the subsystem comprising the inference model, and to be able to make a critical judgement on the data resulting from the inference model (because the algorithm's outputs are intelligible to him and because, in case of doubt, he has other tools at his disposal to confirm or refute the analysis of the inference model).**

3.3. Enabling feedback

Feedback, in the broadest sense of the term, which includes both the analysis of safety events and the checks carried out by the operator (RU or IM) or by EPSF, plays an important role in maintaining and improving the safety level of the rail system. The various players in the rail system analyse events for the purpose of learning from them, for different purposes and for different events: RUs and IMs for events that concern them directly, EPSF for events reported to it and with an aggregating role at national level, in particular to share feedback for the benefit of all, and the Land Transport Accident Investigation Bureau (BEA-TT) for the most serious events.

For this feedback to take place, these players need to be able to determine the causes of the various incidents and accidents. Each event must be analysed in depth to determine its root causes. The aim

[1] As regards intelligibility of output per Maël Pégny, Mohamed Issam Ibnouhsein. How transparent are machine learning algorithms? 2018. Hal-01791021

of identifying these root causes is to determine which safety barriers were ineffective and why, as well as any missing safety barriers.

Reproducibility will be necessary in order to be able to check whether the subsystem comprising a machine learning algorithm has failed or not. This means that it must be possible to retrieve the state the subsystem was in at the time of the event, and that the system's input data must be available so that the event can be replayed. On this second point, input data can be recorded at different times: at the output of the sensors (raw data), during any pre-processing, just before processing by the inference model. The point in time when this data is recorded must be considered and justified, especially in terms of the processing carried out on the data upstream of the inference model.

If the reproduction of the event concludes that the subsystem comprising the machine learning algorithm has failed, it will be necessary to be able to explain this failure, especially with regard to the inference model. This means that the people in charge of feedback (within the rail operators but also within the EPSF and the BEA-TT) will have to be able to understand the choice made by the inference model in the specific case of the event. The inference model is therefore expected to be at least locally explicable[2] . This local explicability for the people in charge of feedback may require specific tools.

In the context of a system comprising a machine learning algorithm, this need to be able to provide feedback raises the question of reproducibility on the one hand and explicability on the other.

IN SUMMARY:

- **Given the role of feedback in maintaining and improving the safety level of the railway system, especially through the analysis of precursor events, the reproducibility of safety events is an important issue. At this stage, reproducibility means that the learning process must be frozen when the inference model is brought into service, and that the algorithm's input data must be recorded over a sufficiently long period using robust “black box” type devices.**
- **Moreover, in order to implement feedback, the machine learning algorithms must be locally explicable, i.e., a result given by the inference model must be able to be explained by the people in charge of feedback (at the railway operator but also at EPSF and BEA-TT), which may require specific knowledge and tools.**

This section will be limited to the cases of an inference model directly involved in a safety action present in the perception subsystem (cf. section 3.1.) and to that of an inference model assisting a human operator (cf. section 3.2.). As already stated, the case of an inference model present in a decision subsystem will not be gone into in detail in this note.

[2] This concept is developed in particular in Maël Pégny, Mohamed Issam Ibnouhsein. How transparent are machine learning algorithms? 2018. Hal-01791021

4. Expected description

4.1. The system (vehicles or fixed installations)

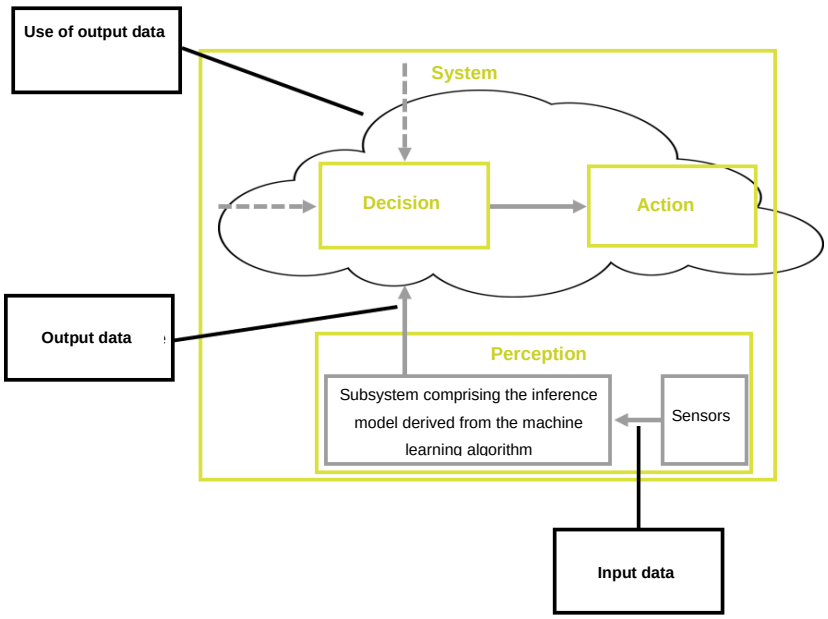
The authorisation logic applying in the railway sector, which can concern either vehicles or fixed installations, was explained at the beginning of this note. There will therefore be no authorisation specifically for equipment comprising a learning algorithm. Authorisation will cover the entire “vehicle” system or the entire “fixed installation” system. It should be noted for the purposes of this note that the term “system” refers to the full scope of the authorisation (vehicle or fixed installation). For the record, authorisation also takes into account the safe integration of this system into the rail system per the meaning in Annex I of Directive (EU) 2016/797 of the European Parliament and of the Council of 11th May, 2016 on the *interoperability of the rail system within the European Union*.

The description of this "vehicle" or "fixed installation" system is therefore a basic necessity. It should give an overview of how the system works and describe how the subsystem comprising the inference model fits into the system and how it contributes to execution of the system's functions. It must also stipulate the conditions for operating and maintaining the system, with special attention to organisational and human factors.

In the scope of the system description, the subsystem comprising the inference model can be considered a “black box”. The aim of this system description with respect to the subsystem comprising the inference model is to identify the following:

- the input data for the subsystem comprising the inference model
- the output data for the subsystem comprising the inference model
- the way this output data is used by the system in question and, where applicable, by the human operator in fulfilling the expected functions.

The diagram below shows the elements concerned.



To illustrate this, we shall consider the simplified case of a subsystem used to detect and recognise any obstacles that might be located on the track within the running gauge. In this example, the input data to the obstacle detection subsystem comprising the inference model is in the form of images from a camera filming in front of the train. For its output, the subsystem transmits its presumption of what category is in front of the train: no obstacle or presence of a human, animal, tree, rock, or smoke. Based on the likelihood of this information, the train will react as follows (assuming all other parameters remain unchanged):

- no obstacle: no change to the traction setting
- human, animal, tree, rock: initiate emergency braking
- smoke (due to a fire close to the tracks which could spread to the train if stopped alongside): traction set to reach the maximum permitted speed.

In this example, we can see that it's not the accuracy level of the entire output data from the subsystem comprising the inference model that matters to the safety demonstration, but the specific accuracy of its ability to detect there is no obstacle and the accuracy of its ability to detect that the obstacle in front of it is smoke. These accuracy levels will be incorporated into the safety demonstration to ensure that the risk of collision and the risk of fire are covered.

Furthermore, this operational description of the system goes hand in hand with its scope of use in nominal and degraded modes, which will stipulate the usage limits of the system and therefore of the subsystem comprising the inference model. This scope of use will notably indicate the maximum operating speed, the limiting daylight conditions (night/heavy sunlight), the limiting weather conditions (snow, fog, etc.) and any constraints exported to the operator and/or maintainer. The safety demonstration must guarantee that the system is not used outside its scope of use.

In the case of a machine learning algorithm assisting a human operator, the system description should also include the interaction between the system and the human operator. This description should enable the details provided to the human operator to be comprehensively mapped and qualified, so that the operator can understand and take a critical look at the data transmitted by the system, coupled with a "confidence" level. It must also enable an understanding of the conditions under which the human operator will be required to interact with the system in nominal and degraded operating situations. Organisational and human factors in particular must be taken into account.

4.2. Subsystem comprising the inference model

Once the system has been described and the role of the subsystem comprising the inference model defined, the factors used to determine the accuracy level of the output data from this subsystem must be justified and detailed.

At this stage, the following points are identified as requiring special attention, given their impact on the accuracy level of the output data of the subsystem comprising the inference model:

- the inference model's architecture
- the learning phase
- the inference model validation phase
- the hardware configuration used for the validation and operational phase.

The inference model's architecture must be described, along with the reasons that led to its choice. The purpose of this description is twofold: on the one hand, to provide arguments explaining how the chosen architecture is suited to the function that the inference model must fulfil, and on the other, to provide the traceability required for feedback purposes.

The way in which learning is conducted will be described. The purpose of this description is twofold: on the one hand, to indicate what has been done to achieve optimum learning for the function in question (for example, the minimum of the cost function used), and on the other, to demonstrate that the learning data is representative with respect to the intended scope of use.

The process for validating the inference model will be described. Justification must be provided for the following three points in particular:

- how does the algorithm design process contribute to its validation?
- how has the database used for validation been produced, especially in relation to the training database, and is it guaranteed representative of all the real-life situations encountered?
- what metric is being used for the assessment and for what reasons?

Management of the learning expiry period must be addressed. A description is required of the measures to be implemented to ensure the learning is still valid in relation to the real situations encountered.

The hardware configuration that will be used to run the subsystem comprising the inference model, along with its conditions of use, will be described for the purpose of demonstrating that the calculations performed are those expected, that the calculation time is compatible with the subsystem's use, and that calculation errors are controlled. It should be noted that the hardware configuration when using the machine learning subsystem may differ from that in the learning phase. It must however be the same configuration as used for the validation tests.

In relation to all the above, there shall be a description of the project's internal double checking process and the process of double checking by an independent third party, in accordance with Regulation (EU) No 402/2013.

5. Requirements and issues to be addressed

In view of the points raised in this note, systems incorporating automatic learning algorithms cannot be authorised unless certain requirements are met. Some requirements already seem achievable, while others probably require further research and development.

Special attention shall be paid to ensuring the training database in representative. It must relate to the scope of use as specified in the system description.

At this stage, the requirements identified and the actions to be taken in relation to some of these requirements are described below. These requirements should enable the following three general necessities to be met:

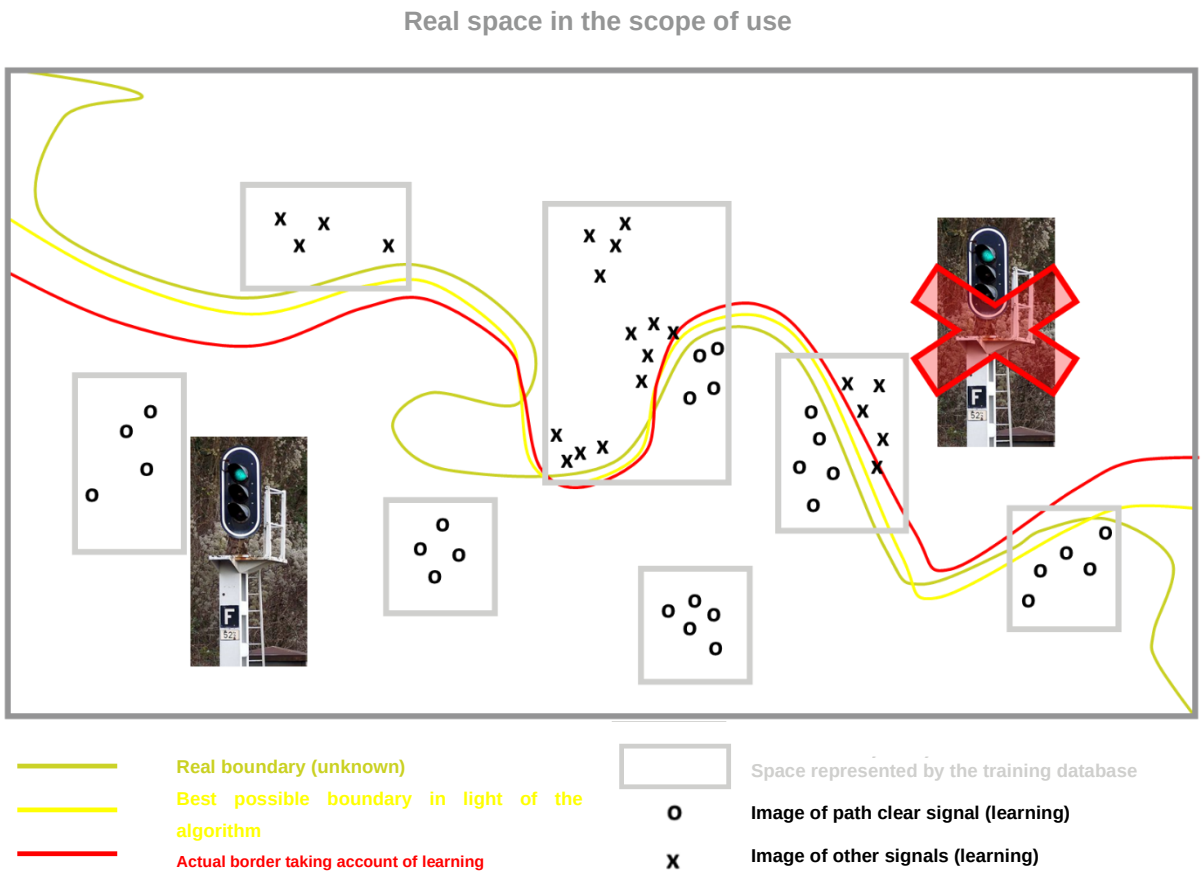
1. the subsystem comprising the inference model must be certifiable so that it can be incorporated in the system's safety demonstration (this includes the operation and maintenance phases to ensure the ongoing safety level is maintained)
2. the subsystem comprising the inference model must be auditable
3. once commissioned, the action of the subsystem comprising the inference model must be reproducible.

5.1. The subsystem comprising the inference model must be certifiable so that it can be incorporated in the system's safety demonstration

Requirement R1: It must be possible to determine and demonstrate an accuracy level for each output of a subsystem comprising an inference model in a given scope of use.

To meet this requirement, it seems clear we must distinguish between two main sources of error:
- the first source of error is specific to machine learning algorithms and is discussed below;
- the second source of error concerns calculation errors due to a hardware error and is common to all software. Hardware errors are not discussed in detail below but are the subject of questions to be addressed.

The diagram below shows a breakdown of errors specific to inference models derived from machine learning algorithms. This breakdown is theoretical as some curves are not known. It deals with the example of an algorithm for detecting a clear-track railway signal.



An erroneous output result from the subsystem comprising the inference model could therefore be due to one of the following:

- the choice of an inference model that does not completely hug the curve forming the boundary between all the “path clear” signals and all those that are not “path clear”. This is represented on the diagram as the difference between the green and yellow curves;
- a training database that is not representative of all the situations that may be encountered (the real space of situations), which means the subsystem cannot be accurate in certain situations;
- the implementation of a learning phase that does not allow the best possible curve to be attained, given the structure of the inference model chosen. This especially applies to deep neural networks, for which the cost function is not convex and therefore only a local minimum could be reached during training.

These last two points account for the difference between the yellow curve and the red curve.

In view of these factors, this first requirement gives rise to the following questions:

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

Question Q1.2: How can the inference model be designed in such a way as to demonstrate that it is suited to the desired function?

Question Q1.3: How can learning be carried out in such a way as to demonstrate that optimum learning has been attained? Is this demonstration necessary for certification?

Question Q1.4: How can the accuracy of the subsystem comprising the inference model be validated? In particular, how many tests should be carried out to ensure statistical significance with respect to the safety objective to be achieved (generalised error compared with the error actually observed during the tests)?

Question Q1.5: Do programming languages have an impact on the safety level of the inference model derived from the machine learning algorithm? If so, are the languages currently in use sufficiently robust?

Question Q1.6: Should the hardware configuration used for testing and operating the subsystem be secure (e.g., with a 2 out of 3 architecture)?

Question Q1.7: Can the relevance of the use of the subsystem be monitored during its operation?

***

Requirement R2: To ensure that this accuracy level is maintained over time for each output of a subsystem comprising an artificial intelligence algorithm, a monitoring process must be implemented.

Over time, the accuracy of the inference model could decrease, either because the quality of the transmitted data changes (ageing sensors, new sensors with different sensitivity, changes in the sensor's "technical" environment, e.g., if the colour of the windscreen in front of a camera) changes, or because the '"real space" changes beyond the algorithm's generalisation capabilities (for example, the introduction of a new railway signal light gantry).

Considering these points, this second requirement raises the following questions:

Question Q2.1: How should the technical environment upstream of the inference model be monitored? Is the principle used in qualifying a change to the rail system both necessary and sufficient?

Question Q2.2: Is it possible to continually monitor or manage changes in the "real space"?

Question Q2.3: Is there need for a periodic review of the certification of the subsystem comprising the inference model?

5.2. Once commissioned, the action of the subsystem comprising the inference model must be reproducible

Requirements R3: In the event of an incident or accident, it must be possible to reproduce the action of the subsystem comprising the inference model.

The main objective of this requirement is to be able to reproduce what the subsystem comprising the inference model did so that we can determine whether it was responsible for the occurrence of a hazard situation, which means determining whether or not the data from this subsystem was adequate. If the output data was indeed inadequate, the subsystem's auditability will help us understand the cause of the error.

To be able to reproduce an action by the subsystem comprising the inference model, following an event, it seems that two conditions must be met:
i) to be able to use the subsystem in the same state as it was in at the time of the event
ii) to know the input data for the subsystem at the time of the event.

Considering these points, this second requirement raises the following questions:

Question Q3.1: Is it necessary and sufficient to freeze the learning process before commissioning in order to meet this requirement?

Question Q3.2: What relevant data should be recorded upstream of processing by the inference model in order to be able to reproduce an event? (input data for the inference model, raw data from sensors, partly pre-processed data, etc.)

Question Q3.3: As with the JRU (*Juridical Recording Unit*), how can these relevant data/decisions be securely stored?

5.3. The subsystem comprising the inference model must be auditable

Requirement R4: Human operators must have the necessary skills and tools to take a critical look at the data transmitted by the subsystem comprising the inference model. This requirement must be applied to the analysis of an event as well as to a subsystem that assists a human operator's decision making.

In order to comply with this requirement, it appears that three distinct cases must be considered:
- the case whereby an event is being analysed by a specialist feedback team or a specialist maintenance team. The timeframe for this analysis is not immediate and certain aspects of the analysis may require further investigation, which could be outsourced. This case also covers analyses conducted as part of the operator's ongoing monitoring or EPSF's audits
- the case whereby an accident or incident is analysed immediately in order to determine whether the subsystem comprising the inference model is involved and could pose a serious imminent risk, which would mean suspending the use of all identical subsystems
- the case whereby a subsystem is assisting a human operator during operation of the rail system.

In the first case, the team assigned to understanding why the subsystem comprising the inference model has produced a given result could have several skills among its members, including one in machine learning algorithms. It will also have the time to call on external skills and tools. In the second case, the team in charge of the analysis must be able to determine in a short space of time whether or not the algorithm is to blame, or at the very least, whether there is any suspicion of its involvement. In the third case, each operator using the subsystem comprising an inference model must be able to make critical judgements on the output data transmitted by the subsystem in a short space of time.

In view of the above factors, in each of these three cases, this third requirement gives rise to the following questions:

Question Q4.1: What training should be given to the analysis teams and what training to the operators to enable them to interpret in any given case (local explicability) the output of the subsystem comprising the inference model, and therefore make a critical judgement on this output?

Question Q4.2: Is it conceivable that the subsystem comprising the inference model could autonomously and reliably assess whether it is being used in its nominal operating conditions (for example, with sufficient daylight or with an excessive level of fog)?

Question Q4.3: Are reliable tools that are independent of the subsystem comprising the inference model necessary to enable this critical judgement to be made? If so, what useful tools are available for each case?

6. Acknowledgements

In drawing up this report, the working group was able to exchange views with the following bodies:

- The Heudiasyc laboratory of the UTC (Université Technologique de Compiègne), in conjunction with the CNRS, and in particular Sébastien Destercke and Mohamed Sallak;
- The certification mission of the DEEL project *(Dependable and Explainable Machine Learning*, www.deel.ai) and in particular Hugues Bonnin, Florence De Grancey, Sébastien Gerchinovitz, Franck Mamalet, the SNCF SA research department, the CIM (Centre d'ingénierie du matériel) of SNCF Voyageurs, and not forgetting the Numalis company and in particular Cyril Cappi, Cédrick Lelionnais and Arnaud Ioualalen.

The members of the working group heartily thank them for the time they were prepared to offer.

The working group would also like to thank the Anavid consortium and Elghazel Conseil for their update on the use of artificial intelligence algorithms in various industrial sectors.

7. Bibliography

Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press, Cambridge. http://www.deeplearningbook.org

Hervé Delseny, Christophe Gabreau, Adrien Gauffriau, Bernard Beaudouin, Ludovic Ponsolle, Lucian Alecu, Hugues Bonnin, Brice Beltran, Didier Duchel, Jean-Brice Ginestet, Alexandre Hervieu, Ghilaine Martinez, Sylvain Pasquet, Kevin Delmas, Claire Pagetti, Jean-Marc Gabriel, Camille Chapdelaine, Sylvaine Picard, Mathieu Damour, Cyril Cappi, Laurent Gardès, Florence De Grancey, Eric Jenn, Baptiste Lefevre, Gregory Flandin, Sébastien Gerchinovitz, Franck Mamalet, Alexandre Albore (2021). *White Paper on Machine Learning in Certified Systems.* DEEL project. arXiv:2103.10529.

Jason Jo, Yoshua Bengio (2017). *Measuring the tendency of CNNs to Learn Surface Statistical Regularities.* arXiv:1711.11561.

Maël Pégny, Issam Ibnouhsein (2018). *Quelle transparence pour les algorithmes d'apprentissage machine ?* (how transparent are machine learning algorithms?) hal-01877760

Sitou Afanou, Cédrick Lelionnais (2021). *Les systèmes de « deep learning » pour l'embarqué ferroviaire* (deep learning systems for onboard rail systems). In RGCF January, 2021

Lucian Alecu, Hugues Bonnin, Thomas Fel, Laurent Gardes, Sébastien Gerchinovitz, and al. Can we reconcile safety objectives with machine learning performances? ERTS 2022, June, 2022, TOULOUSE, France. ⟨hal-03765471⟩

LNE. Standard for the certification of processes for AI - Design, development, evaluation and maintenance in operational conditions. Revision No. 2.0 - 12/07/2021

EASA - EASA Concept Paper: First usable guidance for level 1 machine learning applications. December, 2021, issue 01.

Établissement public de sécurité ferroviaire (*French railway safety authority*)
60, rue de la Vallée – CS 11758 – 80017 AMIENS Cedex 1, France