PRISSMA Project
Plateforme de Recherche et d'Investissement pour la Sûreté
et la Sécurité de la Mobilité Autonome
04/2021 - 04/2024

[L8.15] FINAL REPORT ON THE IMPACT OF AI IN SYSTEM ENGINEERING CHOICES

RAPPORT SUR L'IMPACT DE L'UTILISATION D'IA DANS LES CHOIX D'INGENIERIE SYSTEME

Main authors: Cédric Gava (SPHEREA)

Reviewer: Rafael de Sousa Fernandes (UTAC)

Keywords: AI, system engineering, SOTIF, performance evaluation

Abstract. This document presents the main results of the analysis of requirements and developments impacting the systems engineering process.

Constraints and the state of the art in AI knowledge can challenge conventional systems engineering processes. SOTIF principles represent the state of the art for vehicle commissioning and safety assessment of intended functionality, beyond the classical functional safety applied to well-known autonomous vehicle functions.

Nevertheless, objective assessment of AI system performance remains a challenge that must be met for the safe deployment of AI-based autonomous driving systems.

Résumé. Ce document présente les principaux résultats de l'analyse des exigences et des évolutions impactant le processus d'ingénierie systèmes.

Les contraintes et l'état de l'art des connaissances sur l'IA peuvent remettre en question les processus d'ingénierie des systèmes classiques. Les principes SOTIF constituent l'état de l'art pour la mise en service d'un véhicule et l'évaluation de la sécurité de la fonctionnalité prévue au-delà de la sécurité fonctionnelle classique appliquée aux fonctions bien connues du véhicule autonome.

Néanmoins, l'évaluation objective des performances du système IA reste un défi qui doit être relevé pour le déploiement en toute sécurité des systèmes de conduite autonome basés sur l'IA.

Table of Contents

1 Introduction

As stated in deliverable 8.13 of the PRISSMA project, the system's analysis of an automated driving system should be initiated as a system of systems in the context of a land transport system.

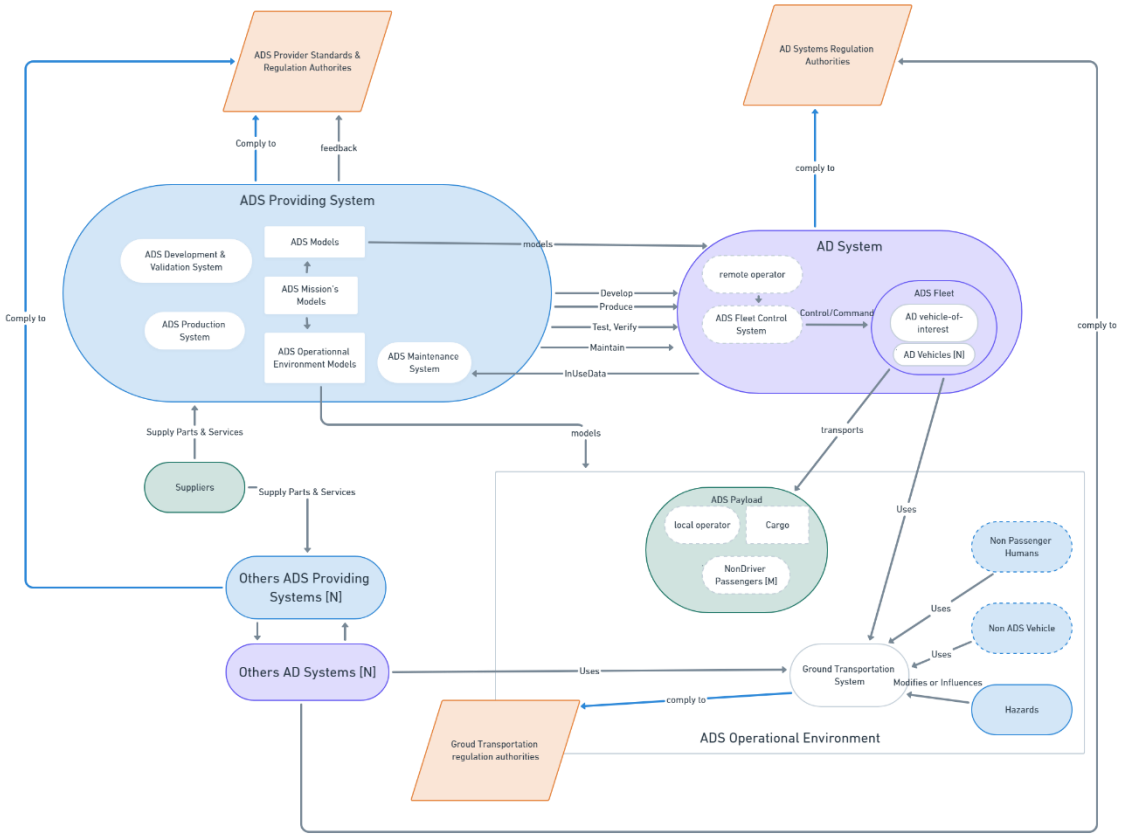

Figure 1 : AD system of systems example

2 Impact of AI on the different activities of System Engineering

2.1 Key issues regarding safety assessment for AI

In traditional system engineering, the safety insurance is based on the quality insurance principles: Plan-Do-Check-Adjust where it is possible to check that the results comply with the expectation, in iterative enhancement process. In safety critical systems, the generic process for safety insurance is comparable with the generic process detailed below [doi: 10.1109/ISSREW.2019.00091].

- Hazard Analysis: Identifying potential hazards associated with the system's usage.
- Safety Requirements: Establishing specific requirements to mitigate these hazards at system, software, and hardware levels.
- Risk Mitigation: Developing and implementing measures to reduce the identified risks.
- Verification: Demonstrating that the risk mitigation measures effectively reduce the risk to an acceptable level.
- Iteration: Repeating the process until the safety level is deemed acceptable.

The SOTIF (ISO 21448:2022) relies on the hypothesis that the vehicle functional safety has been demonstrated through the application of the ISO 26262:2018, which in turns relies on the generic quality insurance principle.

States-of-the-art in AI shows that such hypothesis does not apply in AI and, more specifically for supervised machine learning algorithms

- specificability: behaviors easy to train for with datasets are very difficult to specify using requirements (example is pedestrian detection. What a pedestrian is ? Does it means that people in a wheelchair are not included in this category?)
- hazard assessment impossible without specification: How to define risk mitigation requirements when functional requirements are not defined?
- risk mitigation verification is not possible: We cannot present irrefutable arguments demonstrating that these risk mitigation requirements are met (neither proof nor postulates that would demonstrate this coverage exist, due to issues of causality and non-linearity).
- achieved quality level is not quantifiable: it remains unclear when to stop this retraining process. Iterative improvement of this quality level is not possible
- isolation of defect: is almost impossible inside a neural network at the state of the art.
- quality assurance composition: demonstrating the system's quality assurance through the quality assurance of its AI components, similar to estimating system MTBF (Mean Time Between Failures) through the MTBF of internal components, is currently not possible at the state-of-the-art.

2.2 Cross-domain aspects

2.2.1 Needs and Requirements

The formalization of AI system definition assets is a key for transparent transition of those assets from human-readable paradigm to automatic processing.

Which raises the following question: should not well-defined ontologies be at the foundation of common representations of rules and scenarios of AD System in order to improve understandability between the different agents involved in the AD System of systems?

2.2.2 Traceability

Traceability is already identified as essential in the transportation industry, but this traceability shall also be enforced on the enabling systems and their associated development projects. The audits and qualification tests AI based AD system should encompass both the system, and it test enabling system: a flaw or default on an AD system should have traceability up to the configuration of the test systems (simulation, closed road or open road) thus including their complete configuration information.

In a broader aspect, the taxonomy of the different system engineering assets required for the SOTIF process and AI based systems should be established, in addition to the requirements:

- Scenario
- Operational Domain
- Triggering conditions of hazardous behavior
- Vehicle Dynamic
- Vehicle Land Safety Strategy

As usual in critical systems, all the assets shall be uniquely identified to be referenced.

Based on the identified AI functions-of-interest, AI activities-of-interest and operational domain the PRISSMA method shall verify that the ARTS supplier has:

- Identified the list of all the applicable regulations to the ARTS and ARTS chain of suppliers (the ARTS supplier and the suppliers of the ARTS subsystems, recursively) and extract the applicable regulations requirements from this list.
- Conducted safety assessment on those regulations requirements to demonstrate possible inconsistencies and risk on the impact of the whole set of applicable regulation requirements.
- Traced the selected requirements with the applicable regulations they are extracted from
- Setup a process to regularly identify any update in the applicable regulations requirements

Example: From the European ADS act (UE2022/1426), the regulation specifies the performance requirements of level 4 automation vehicle classified in the following 12 categories.

1. Dynamic Driving Task (DDT) under nominal traffic scenarios
2. DDT under critical traffic scenarios (emergency operation).
3. DDT at ODD boundaries
4. DDT under failure scenarios
5. Minimal risk maneuver (MRM) and Minimal risk Condition (MRC)
6. Human machine interaction for vehicles transporting vehicle occupants
7. Functional and operational safety
8. Cyber security and software updates
9. ADS data requirements and specific data elements for event data recorder for fully automated vehicles
10. Manual driving mode
11. Operating manual
12. Provisions for periodic roadworthiness tests

2.3 Technical process

The transition from pure simulation to closed road, and open road testing should consider the AD system's model lifecycle at the same time with AD system itself.

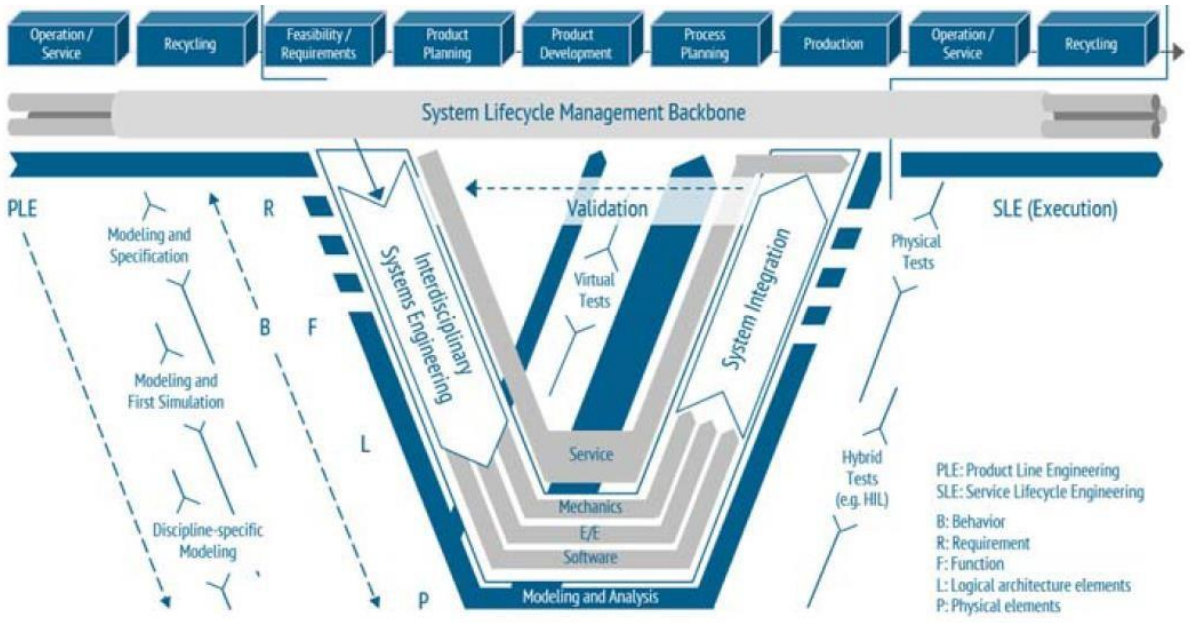

Figure 2 : MPVE V cycle

2.3.1 Business or Mission Analysis

The PRISSMA project itself is part of the Business analysis of deploying AD SoS at large scale (It is part of the deliverable 8.4 to link the validation framework to the economic efficiency).

2.3.2 Stakeholder Needs and Requirements Definition Process

The statement of all the stakeholders' needs and requirements is at the core of the system verification and system validation activities. The quality of the statements of the needs and requirements and the completeness of identification of the stakeholders of AD System of System will be a key factor for the success of the development of the framework for the safety of AI based AD systems.

When considering a test system as a system of interest, the best stakeholder needs definition is a complete test campaign definition of the system under test. The completeness and the quality of the definition of the tests carried out with the test system, along with the system under test requirements to be tested, is a key factor in the delivery of the test system that fit the need.

2.3.3 System Requirements Definition Process

This process is out of scope of as with the stakeholder needs and requirements definition process a good definition of system requirements is a key to the verification and validation of a dependable system capabilities, including its safety and securities performances.

The assertion of the quality of the system requirements, using precise rules like the writing guide [INCOSE-TP-2010-006-03 - Guide for writing requirements] is key to the success of the system.

2.3.4 Architecture Definition Process

The architecture of the AD system is out of scope of the PRISSMA project. The architecture under concern is the architecture of the test system: which components can be used to meet the test system requirements.

Based on the SOTIF process, it is required to elicit the triggering conditions of potential hazardous behaviors of the autonomous system. This elicitation has to be made during the process of defining an autonomous driving system (so before the process of defining the vehicles and therefore before the processes that define the technology of the sensors and AI components that will realize the autonomous driving system). The architecture process is the appropriate process to tackle this discovery:

- It is part of the definition of the autonomous driving system.
- But has to consider the technologies involved by its components which together are the solution to the need expressed for the autonomous driving system.
- The interfaces components (sensors) shall be extensively specified in this process, as any interface of any components of the system must be fully described during the architecture process to ensure successful integration.
- The inner components (AI) cannot be specified during this process, since they belong to the solution of the ADS components, and will be fully specified during the definition of these components. Instead, the constraints in the selection of such components shall be expressed during this step.

The technologies of AI and sensors shall be specified during the architecture process of the ADS to enable the elicitation of triggering conditions of the autonomous driving system

2.3.5 Design Definition Process

The complete set of requirements for the test system derived from the safety and security constraints of the system under test should have already been provided during the stakeholder needs and requirements definition process (see 6.2). Nevertheless, since the design definition brings to the architecture process additional constraints arose by the system elements, one particular component of the test system needs specific considerations in the scope of PRISSMA:
what is the impact of using AI based component in the test system?

2.3.6 System Analysis Process

This process gathers all the analysis required by the other processes. Always present in a system engineering project, there is nothing special for the PRISSMA project, unless to state that any required analysis shall be conducted to meet the project's mission's goals.

2.3.7 Integration Process

The integration of a given set of elements of a system needs some particular tests to be taken to assess the capabilities of this set regarding system's requirements. Depending on the system of interest, the scope of the tests to be taken should be analyzed properly:

* AD vehicle: the SOI is one vehicle. The integration of some components, including AI components, is a state-of-the art activity in system engineering: the context of these components is simulated. As mentioned in §7.5, the configuration management of the set of these system's element AND its test system should be done carefully.
* AD system of system: one AD system is a part of the AD system of systems. Even if this AD system comprising the AD vehicle fleet and possible remote supervision has been validated, the operation of the first vehicles in the Road is an integration for the AD system of system.

The transition from validation to operation of a given AD system should be ruled and audited by authoritative organizations.

2.3.8 Verification Process

The PRISSMA project focus on some particular verification activities: the tests that can be taken on an AD system (simulation test, closed road or open road tests).

Since the verification scope is to verify a system against its requirements, and not the stakeholder's needs it fulfills, the quality of the requirements will have a particularly strong impact on the seamless transition of the verification success to the validation success: if the system has good requirements, then a successful verification will enable a successful validation. But if the system requirements definition process has flaw, then the delivered system fulfilling its requirements but may not fill the stakeholders' needs.

2.3.9 Automated Road Transport Systems (ARTS) Qualification

The PRISSMA qualification relative to the safety demonstration of AI based ARTS relies on:

- The qualification of the 30AI activities-of-interest of the ARTS supplier
- The qualification of the applicable requirements and development artifact used for the evaluations
- The qualification of the 31AI components (324.2- AI component qualification requirements)
- The qualification of the ARTS 33AI functions-of-interest (344.3- ARTS AI functions-of-interest validation requirements)

As usual in critical system engineering, all the inputs of the process shall be individually qualified (the ODD, the pathway annotations, the scenario, the OEDR, etc, etc..)

The PRISSMA method shall evaluate and validate both 35AI functions-of-interest of ARTS system and AI components along with supplier's AI activities-of-interest.

Attention shall be paid on the characterizations of the functional domains of the AI functions of the ARTS. Considering that AI can be highly non-linear, the figure below summarizes the possible domains to be considered in these characterizations.

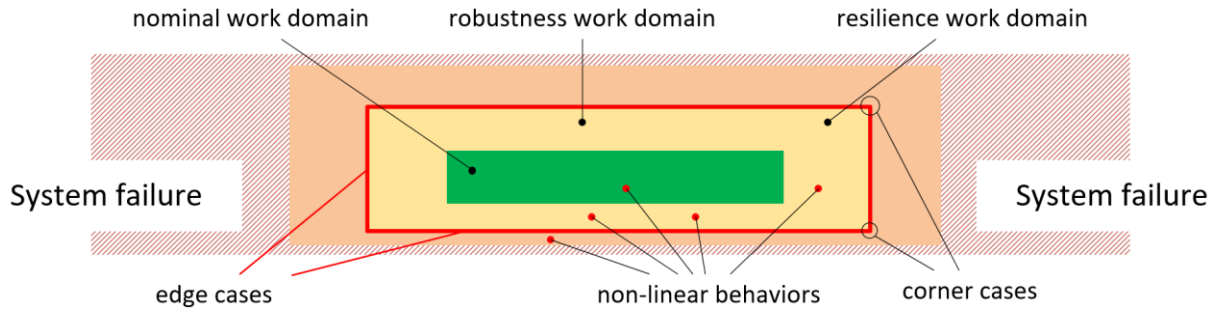

Figure 3: Nominal vs Robustness vs Resilience [Adapted from DEEL]

The PRISSMA method shall verify that the provider of the AI components has verified that the AI components meets the qualified performance, safety & security objectives during all the life cycle phases of the AI component (example: learning process, validation).

Including: Center/Tail distribution error minimization strategy, Convergence measurement, AI robustness and AI resilience.

In various types of AI models, a confidence index (or confidence score) can be used to indicate the probability that a given prediction or classification is correct. This is common in machine learning algorithms such as logistic regression, support vector machines (SVMs), or neural networks. The confidence index can help in decision-making by providing an estimate of the reliability of a prediction.

Confidence scores can be used in perception algorithms to estimate the likelihood that a detected object is of a certain type (e.g., a pedestrian, a bicycle, or another vehicle). Such scores can be used to inform decision-making in control algorithms.

The PRISSMA method shall verify that AI components-of-interest also provide a Confidence Index on their output and a confidence index justification document that details:

- What are the specifications of this confidence index (how is it computed, what is it's purpose)
- What are the expected performances of this index
- How to interpret the different values (range from very good to very poor confidence)

Example: Detection tracking of road markings can provide up to many confidence indexes 1st stage confidence: primitive extractors on the belonging of the pixel to a road marking 2nd stage confidence: occurrence of the tracking - the more the road marking has been positively followed, the better on the overall confidence

In the scope of the PRISSMA method, the main objective of the safety demonstration is to verify that the ARTS is at least as safe as human in equivalent situation (GAME principle). This demonstration shall therefore rely on an objective criteria that remains the same in all the situations. The proposal is to demonstrate that the ARTS has no accident resulting in severe or fatal injuries after being operated on a large enough distance or duration with appropriate justification of the coverage of the evaluation domain, which means the demonstration has been realized on many different situations. Ideally, the qualification should also enable to verify that the ARTS has no repeated incidents.

The evaluation domain is the space resulting from the combination of the following spaces:

- The ODD, including:

- The road infrastructure (Pathway) and the events that can reasonably occur on this pathway (both environmental conditions and actors events)
- ARTS capabilities limitations with regards to the possible events and environmental conditions (for example, if the ARTS cannot be operated safely at night, then night utilization is out of the ODD)
- ARTS functions and requirements
- OEDR (ARTS automatic driving requirement, being AI or not)
- Other system events (risks, failures, functional insufficiency, triggering conditions)

Hypothesis 1:

The vehicles of the ARTS must operate safely even if the other subsystems like the infrastructure or the supervision of the ARTS are dysfunctional. This hypothesis as an impact only on scenarios where one components of the infrastructure has a failure. For example, if connected traffic lights are used for the utilization of the ARTS, but the ARTS must remain safe when the traffic lights are dysfunctional.

Any vehicle with SAE 4 level is meant to operate safely inside it's ODD (including its pathway). As the infrastructure is a part of the ARTS, then the pathway is considered inside the system-of-interest.

The PRISSMA method shall verify that the ARTS provider has defined the safety objective metric in compliance with the state-of-the art and applicable regulations where the ARTS is operated (see 652.2- Qualification of applicable regulation requirements)

Note 1: This metric can be expressed in a distance without fatality (like 275 million fatality free miles) or in hour of travel without fatality (like 10^-7 fatality per hour).

The PRISSMA method shall verify that the ARTS supplier has performed enough of the necessary, sufficient and representative tests to demonstrate the safety of the ARTS on a sufficient distance (or duration) regarding the safety objective metric qualified announced.

Note 1: This requirement probably implies that huge amount of tests are done with simulation on qualified simulator

Note 2: Due to the statistical distribution justification requirement or in a broader way to increase safety demonstration, the total distance / duration covered by the test has high probability to overflow the initial safety objective metric.

Note 3: The overall safety demonstration relies also on the qualification of the inputs, safety analysis and risks mitigation strategies, all these points are covered in all the previous sections of this document).

Example 1: According to the National Highway Traffic Safety Administration (NHTSA) in the United States, in 2019, there were approximately 1.1 deaths per 100 million miles traveled (about 160 million kilometers). This equates to approximately 0.0000069 deaths per kilometer traveled. To demonstrate that fully autonomous vehicles have a fatality rate of 1.09 deaths per 100 million miles (a reliability of 99.9999989%) with a 95% confidence level, the vehicles would need to be driven 275 million fatality-free miles (440 million fatality-free km) [Sources: National Highway Traffic Safety Administration (NHTSA), "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" by N. Kalra and S. Paddock.]

Example 2: The example of acceptance criteria indicated in the footnote relies on a safety threshold (10-7 fatalities per hour of operation) based on the analysis of current EU road accidents aggregated data. Such threshold is suitable for the market introduction of ADS based on similar services as the ones which the aggregated data refers to; i.e. buses, coaches, trucks and cars. Therefore, a more suitable reference threshold could be derived specifically for each use-case, also considering the defined operational design domain (ODD) [UE ADS Act].

Example 3: The GAME principle (Globally At Least Equivalent) applies to Automated Road Transport Systems (ARTS). It aims to ensure that the overall safety level of an ARTS is at least equivalent to that of existing or comparable systems. The principle considers users, operating staff, and third parties. It allows for some flexibility by permitting a "system" approach to safety. The guide serves to formalize expectations and provide a framework for industry professionals. [Sources STRMTG GAME application guide].

Statistical distribution justification

The PRISSMA method shall verify that the ARTS supplier has provided the justification of the test run distribution within the 66evaluation domain.

Note 1: Unless required by an applicable regulation the use of scenario approach is a mean to give confidence in this justification

Note 2: One of the impact of this requirement could be the increase of the 69safety objective metrics (for example: cumulating 200 Million fatality free miles on highway, and 100 million fatality free miles on crossover , etc, etc...) even though the method for the allocation of global objective to different parts of the pathway is not available at the state-of-the-art.

2.3.10 Qualification of Simulator and Test sequencer

The PRISSMA method shall verify that the ARTS simulator, provided by the ARTS supplier, has the following properties:

- The whole simulator has been provided by a qualified process (specified, designed, evaluated, verified, validated by third party)
- The validation shall include a correlation/consistency justification campaign executed with the real vehicles of the ARTS on test tracks and (or) in operational pathway that indicates at least:
- usage of qualified correlation/consistency metrics between the simulated ARTS behavior and the real ARTS behavior (digital twin)
- % coverage of the tests passed on the real system versus the tests passed through simulation

Qualification of the Test Sequencer

Any evaluation made in the PRISSMA method and relying on tests ran with the simulator of the ARTS shall use qualified test sequencer that:

- Enable the PRISSMA evaluation to set it's own test campaign parameters and observe the results of the tests, independently from the one realized by the ARTS supplier
- Is different from the test engine used by the ARTS supplier or has been certified by an independent organization
  Rationale: to avoid any bias of the test campaign resulting from a possible bug in the test sequencer

In engineering, a simulator is a hardware or software tool used to replicate the behavior of a physical system for testing, analysis, and design purposes. It allows engineers to experiment with various scenarios without the need for real-world testing. [IEEE Standard 1076-2017].

The PRISSMA method shall verify that all the models and simulation of the ARTS have followed a qualification process, comparable with the "credibility assessment framework" [UE ADS Act] or any process demonstrating that the models & simulations used in the safety argumentation of the ARTS have been specified, verified, validated, documented, maintained.

The Simulation is the result of the utilization of the Simulator.

In science, a model is an intellectual or material construct designed to represent an aspect or process of reality, highlighting certain essential features while disregarding other details." -[Modeling and Simulation in Science and Mathematics Education (A. A. Berry, 2008]

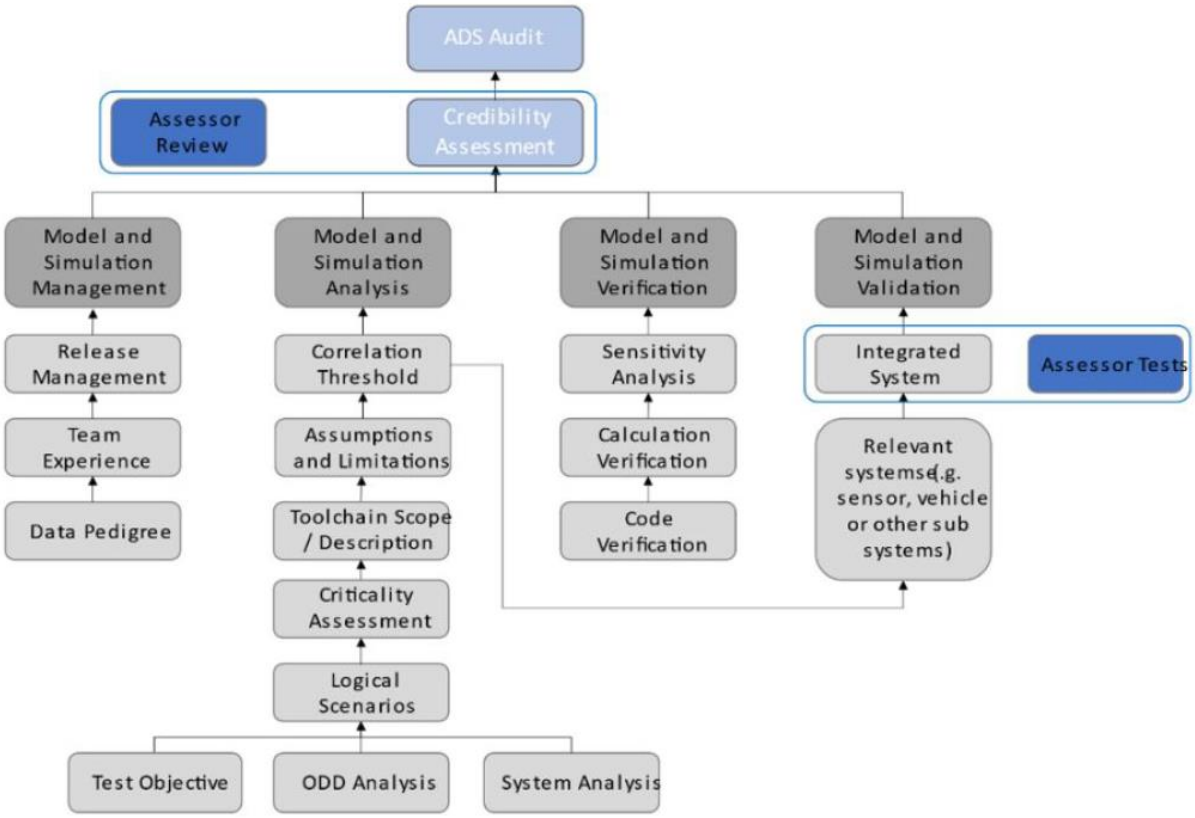

Figure 4: Graphical representation of the relationship between the components of the credibility assessment framework to assess the M&S [EU: ADS Act]

2.3.11 Transition Process

The transition of some systems during their lifecycle can arise some particular troubles. For an AD system, its transition from validation (maybe in closed road) to operation (possibly in open road) should be planned, along with the transition for the enabling systems that will guarantee its security. These questions are typically core questions of the PRISSMA project.

2.3.12 Validation Process

This process is key for the PRISSMA project. The good definition of stakeholder needs and requirements is a key factor of success in traditional engineering projects and will also be the case with AI AD system: what environment is enough for validation, what are all the

stakeholders involved for a particular AD system are typical questions relative to validation that should be addressed.

2.3.12.1 Impact of the transition to open road

When engineering a standalone system (not system-of-systems engineering), the complexity of the whole system is split into smaller parts which can be separately studied. When running such processes, the requirements of a given system's element is an input for the V cycle of this system element. This element can then be recursively divided into sub-elements, and the process is iterated until the level of parts to be analyzed fits the organizations constraints.

The system analysis of the parent system can contribute to realize the mission analysis of the system elements and, ideally, the requirements of the system elements are all defined during the definition of the parent system. In this case, the validation of the system's element can be assessed separately before the element is assembled into the parent system to be integrated.

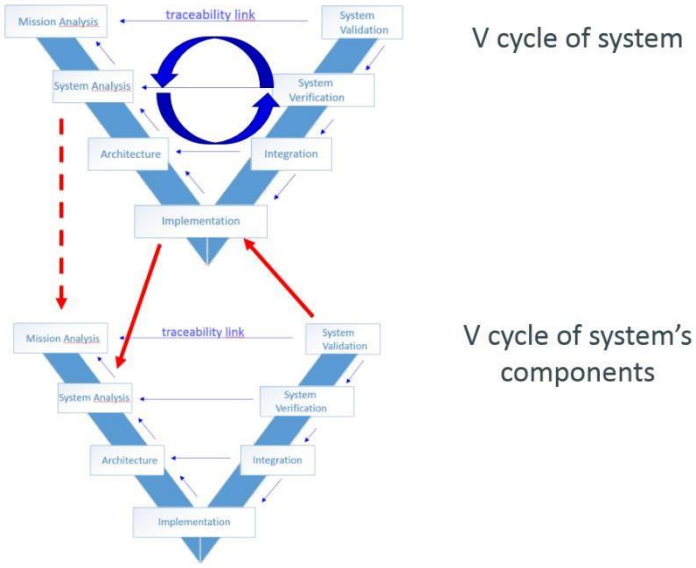

Figure 5 : V cycle integration between system and system's parts

When engineering a system of systems, some capabilities of the system of system can only arise because of the presence of such systems and cannot be easily split into smaller independent functions allocated to each subsystem.

In that case, it's highly probable that an AD system contributing to the ground transportation system of systems. It can only be fully validated after being integrated into the SoS. Since the SoS of ground transportation cannot be halted for integrating the ADS, it implies that the ADS can be completely secured after only a sufficient time of operation. The impact of the transitions process between the different V cycle is then to be highly considered.

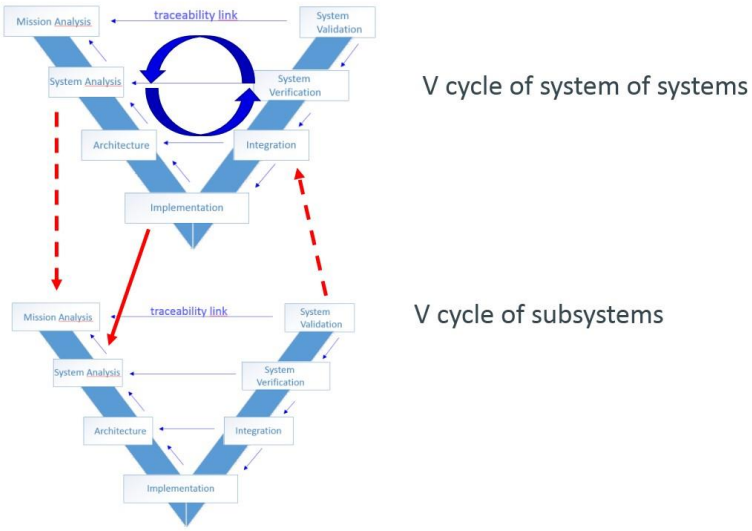

Figure 6 : V cycle integration between system of systems and subsystems

The previous statements lead to the following hypothesis on the process for validating an autonomous driving system:

- In the recurring V cycle of the SoS of ground transportation, a clear list of capabilities and requirements applicable to any ADS is made available to all the stakeholders involved in the design operation and maintenance of ADS.
- In the V cycle of the ADS, the system validation strategy should define the ADS functions that can only be validated by integrating the ADS system into the ground transportation SoS. Those validation strategies should include some KPI to monitor the progress of the validation other a period of time, and different phases of maturity where a human operator might have to be permanently in capability of taking over the vehicles to prevent accidents.
- The continuing recording and analysis of the data of the integrated ADS is mandatory to assess the correct behavior and safety of the ADS as a part of the ground transportation SoS. In other words, the completion of the validation of the ADS can only be assessed after monitoring the decrease some operational KPI (like amount of nearly accidents, or amount of security distance limit broken).
- The components of the ADS are all fully qualified before the ADS system is integrated into the ground transportation SoS:
    - Vehicles are homologated in closed roads
    - Supervision and infrastructure are verified
- The ADS system has been verified and certified in closed road before being integrated into the ground transportation system of systems.

2.3.12.2 Impact of the insertion of a new component in a system

The verification or the validation of a system relies on different techniques like peer review, tests or analysis. The verification and validation tests are driven by black box technique: the system is considered a black box, which can be influenced only by its inputs, and monitored by its output. The functional chains are chains of internal functions which can be observed by the interfaces of the system.

In the simple example below, two functional chains are enough for defining the expected behavior of the system, and thus define the scope of the functional verification and validation of the system.

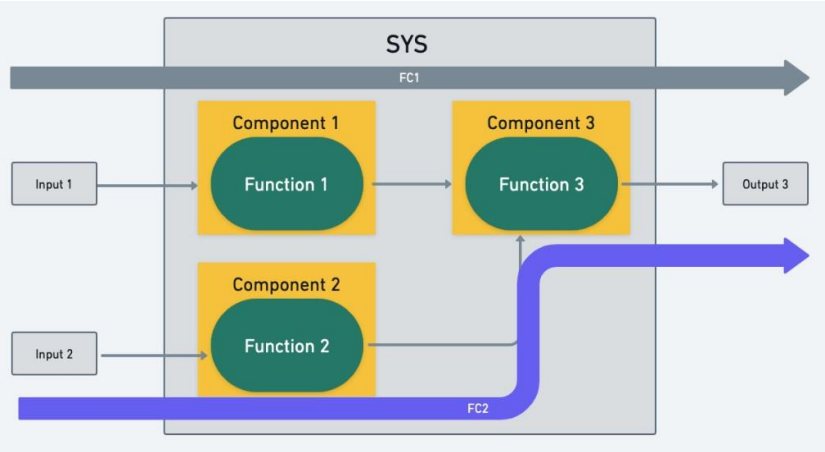

Figure 7 : Example system with functional chains

Even if the system is tested as a black box, the knowledge of its inner composition enables to infer some properties of its components. For example, an error detected on the functional chain FC2 implies that either Component 2 or Component 3 has defect. The testability analysis of such system would lead to test the functional chain FC1 to locate the defect: if the FC1 has no error, the defect is on the component 2 or in the connections of Component 2 with its neighbors.

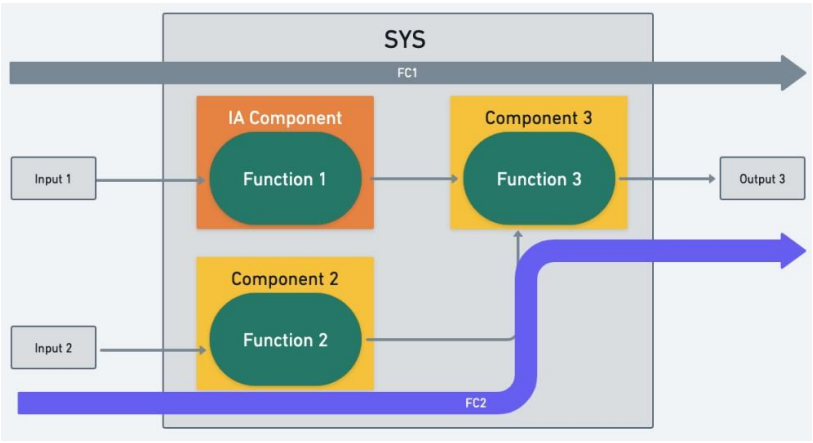

Figure 8 : Example of the introduction of AI component into A functional chains

2.3.13 The qualification strategy of PRISSMA method

The PRISSMA certification of an Automated Road Transport System (ARTS) is based on the successive qualification of its constituent AI components and functions, as well as the concepts and processes of its life cycle:

- Homologation of the vehicles
- Qualification of other system components (supervision, connected infrastructure)
- Qualification of the ARTS supplier process, whether this process involves the integration of existing components or includes, directly or indirectly, the complete development of each component

- Qualification of the ARTS operator and maintainer process, including it's safety management system (“Système de Gestion de la Sécurité" in French)

The homologation of the vehicle relies on:

- The qualification of this vehicle supplier's process
- The qualification of the AI components used in this vehicle

All qualification processes through the PRISSMA method are based on equally qualified inputs: requirements, performance and safety objectives, Operational Design Domain (ODD), Object and Event Detection and Response (OEDR), routes, scenarios, and metrics.

2.3.14 Maintenance Process

The maintenance process is addressed in section with the concept of Integrated Logistic Support which focus on the means and organizations required to increase reliability and availability of a particular system of interest.

The PRISSMA method shall verify that the ARTS supplier is able to demonstrate the safety of the ARTS when the triggering conditions which led to a hazardous behavior of the ARTS (accident or near-accident) are reproduced.

The PRISSMA method shall verify that the ARTS supplier is able to demonstrate the compliance of its process with the relevant qualified requirements from the PRISSMA baseline for this ARTS.

The PRISSMA method shall verify that the ARTS supplier implements maintenance and feedback activities achieving the following outcomes:

- Update the catalog of scenarios, including misuses, to be used for safety argumentation for the updates of the ARTS.
- Ensure the recording of pertinent vehicle data (sensor inputs, decisions) in order to provide feedback to the ARTS activities-of-interest in case of system's failure, accident or near-accident in order to fix the ARTS functions.
    - Note 1: The access to the recorded video by the local infrastructure to collect the potential hazardous behavior of an ARTS that has not detected near-accident or hazardous behavior should also be considered (to complete the set of data that can be used for post analysis).
    - Note 2: Sensors provided only AI-computed output, and not raw input, should be avoided (as this might hide the triggering condition recording)

Demonstrate rigorous configuration management practices for the update of the ARTS and AI components in addition to risks assessments and mitigations in the updates of AI components.

Note: This is implemented in WP7

2.4 Technical management process

2.4.1 Decision management process

In such new technological field as AI based AD system, any decision regarding the development and maintenance of the AD system should be aggregated correctly for feedback and problem solving.

This activity is generally not considered enough in a result-oriented implementation of this system engineering process, where the delivery of acknowledged work products matter more than rationale of past decisions.

In this new field of AI, tracking the alternatives and decisions in all the organizations involved in the ADS providing system will be a key for improvement in case of undesired behavior of the ADS system.

2.4.2 Risk management process

The risk management process is at the core of safety and security and is emphasized in many working groups and projects about AI based systems.

2.4.3 Configuration management process

Attention should be paid to the configuration management of the configuration management process should be carefully applied to enabling systems and components:

* Simulation models and verification activities: asserting the validity of the results of a particular verification campaign made on a particular set of simulation models interaction (what is the validity of test campaign conduced on the AD vehicle's model in version x, interacting with environmental model version y and mission's model version z? When one configuration of one of these models is updated, what can be stated about the results of previous test campaign?
* Identifying the baselines of system of system elements: the AD System along with its test system should be identified jointly when taking tests

2.4.4 Information management process

Human and AI based systems can be viewed as agents interacting with their environment. The use of ontologies has been widely used to gather concepts and definitions into reasonable structures both for human and computer.

In addition, some information regarding the security of operating AD systems should be transmitted by a given AD systems operator to disseminate potential knowledge across the different stakeholders of the AD systems, including the organizations developing and maintaining those AD systems.

2.4.5 Measurement process

The assertion of the security of AI based AS system will probably involve some KPI evaluating the quality of the tests taken on AD systems. Should the statement of these KPI be part of the PRISSMA project?

2.4.6 Quality assurance process

The quality assurance process is key to the success of AD systems, but goes far beyond the scope of [1].

2.4.7 Agreement process

As noted in the introduction note of this process, asserting the security for the supply chain is key in the security of the system as a whole. The establishment and monitoring of the agreement with the supplier shall include the security, including cyber security, of the whole AD system and its associated ADS providing system.

3 Impact of the SOTIF principles on the system engineering activities

This paragraph synthesizes the ISO/DIS 21448 based on the presentation [1].

3.1 Functional Safety

As stated by the ISO26262 standard, the functional safety is the "Absence of unreasonable risk due to hazards caused by malfunctioning behavior of Electrical/Electronic systems."

Such methodological approach relies on two pillars:

1- A systematic identification of the possible threats and the actions on the design to mitigate their risk of occurrence and gravity at an acceptable level.
2- The requirement of a peer review by stakeholders different from the design team

The functional safety focuses on the system failures and the ways to mitigate the impact of these failures on the safety of the system.

3.2 Safety of the Intended Functionality (SOTIF)

As stated by the ISO21448 standard, the safety of the intended functionality is defined by the "Absence of unreasonable risk due to hazards resulting from functional insufficiencies of the intend functionality or by reasonably foreseeable misuse by persons."

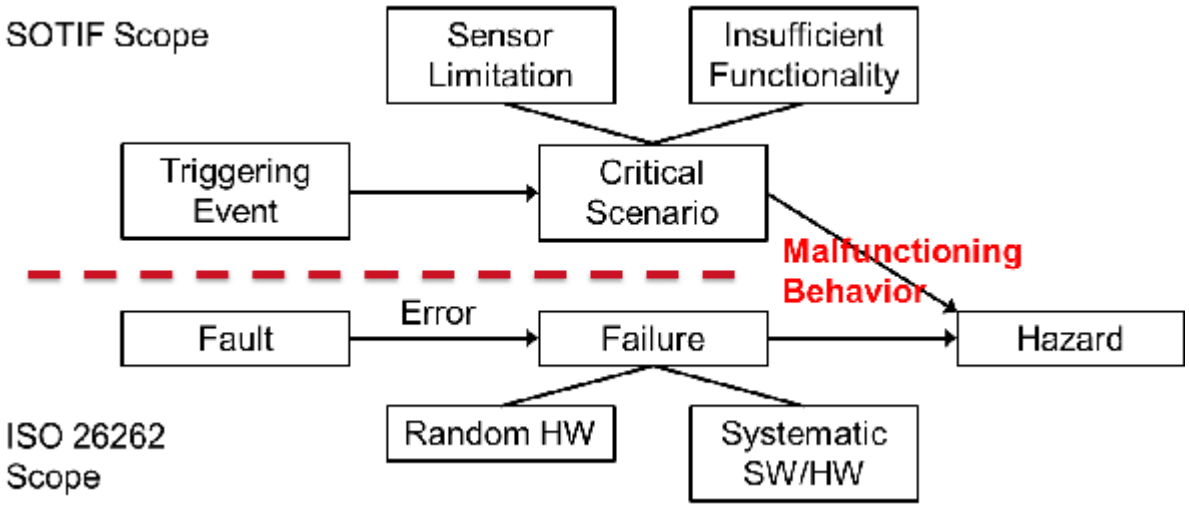

Figure 9: SOTIF vs ISO 26262

The scope of this standard are the Road vehicles, and particularly, the innovative functions of driving relying on sensors and complex algorithms that can demonstrate hazardous behavior without any system failure. The main objective of the SOTIF principles is to ensure a certain level of safety when no system failure has occurred, and as such is a complementary activity with the functional safety.

Source	Cause of hazardous events	Within scope of
System	E/E System failures	ISO 26262
	Insufficiencies of specification, performance limitations or insufficient situational awareness, with or without reasonably foreseeable misuse	ISO/DIS 21448
	Incorrect and inadequate Human-Machine Interface (HMI) design (inappropriate user situational awareness, e.g. user confusion, user overload, user inattentiveness)	ISO/DIS 21448 European Statement of Principles on human-machine interface [1]
	system technologies EXAMPLE: Eye damage from a laser sensor	Specific standards
External factor	Reasonably foreseeable misuse	ISO/DIS 21448
	Attack exploiting vehicle security vulnerabilities	ISO/SAE 21434 or SAE J3061
	Impact from active infrastructure and/or vehicle to vehicle communication, and external systems	ISO/DIS 21448 ISO 20077; ISO 26262
	Impact from vehicle surroundings (e.g. other users, passive infrastructure, weather, Electro-Magnetic Interference)	ISO/DIS 21448 ISO 26262

Figure 10: Scope of ISO/DIS 21448 versus other safety standards

The ISO/DIS 21448 prescribes a process to accept the commissioning of a system:

1- Evaluate by analysis: to select the functions where the SOTIF principles apply and to identify the risks and the acceptance criteria
2- Evaluate known hazardous scenario: assess if the V&V strategy demonstrate risk of hazard is small enough
3- Evaluate unknown hazardous scenario: explore unknown scenarios and evaluate if the risk of hazard is small enough

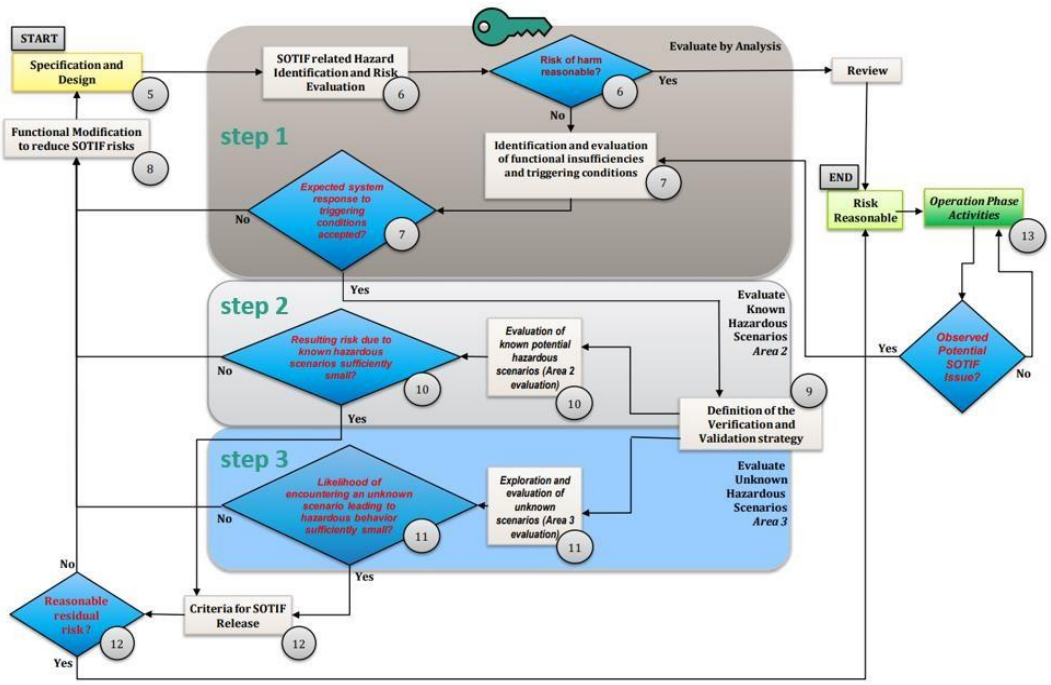

Figure 11: Overview of the SOTIF assessment process

4 Impact of AI on the different activities of Performance System Engineering

4.1 Performance targeted in System Engineering

The system engineering process addressed in this document is very much targeted into Performance Assurance.

There are two categories of performances:

- Functional performances: like the maximum speed limit of a vehicle or kilometers autonomy in standard environmental conditions for example
- Non-functional performances: like Safety, Cyber security, Operational Reliability or Availability, Sustainability…

One of major issues of System Engineering process is to:

- Specify qualitatively and quantitatively those performances at the beginning of the process
- Allocate those performances in a top-down process through the different breakdown levels of the system, from the highest top level up to the elementary decomposition entity level
- Integrate in a IVVQ activity mode and on a bottom-up cycle manner, from the lowest decomposition level up to the highest level which is the end user level - Demonstrate full compliance of the system

The performance, safety & security objectives come from three different sources:

- The qualified regulation requirements
- The ARTS supplier
- The PRISSMA method itself

Only qualified objectives, qualified KPI, qualified metrics applicable to 14AI function-of-interest or 15AI activities-of-interest can be used in the PRISSMA method. The ones coming from the applicable regulation requirements are, by essence, qualified to be used in the PRISSMA method. the other shall follow a qualification process defined in the following sections.

4.1.1 AI lifecycle data qualification requirements

Data used for a critical system is essential for ensuring the accuracy, reliability, and safety of the system. It plays a critical role in the design, development, testing, and maintenance of critical systems and should be selected and qualified carefully to ensure that it meets the specific requirements of the system.

AI lifecycle data refers to the comprehensive data set used throughout the lifecycle of an Artificial Intelligence system (design, maintenance, utilization):

- Maps: Topographical maps and ground truth data from both simulated and real-world environments.
- Databases: Involves databases for learning, testing, and validation, supporting supervised or semi-supervised machine learning approaches. This data may serve as a hypothesis or a reference.
- Accident Statistics: Databases of accidents and statistics serve as a foundation for learning and improvement.
- Reference Data: Data used as reference for comparison.

To qualify AI lifeycle data, the PRISSMA method shall verify that the supplier of the data has followed a process to qualify the data, with at least the following steps:

- AI lifecycle data specification: The reference data constraints and properties must be specified, in compliance with any regulation or norm relative to the aspects reference by the data. In particular, if the reference needs to evolve in a controlled manner over time, with a specific objective, it is in the specification of this reference that this should first be described. Example: a test base must evolve regularly to prevent the AI developer from knowing the test base, and thus the references associated with this test base will need to follow this evolution.
- AI lifecycle data selection or definition: The reference data must be extracted from a verified source or created to meet the specified objectives.
- AI lifecycle data validation: The validation activities shall ensure that the chosen characteristics are suitable for the intended purpose. The selection of the persons realizing the validation must be justified (in particularly they should be qualified experts different from the persons who have specified the metric).
- AI lifecycle data verification: The reference data must be verified to ensure that it performs consistently and reliably in routine use. This may involve analyzing a set of control samples or using proficiency testing programs to assess performance. The selection of the people realizing the verification should also be justified.
- AI lifecycle data monitoring and maintenance: Finally, the reference data must be regularly monitored and maintained

The PRISSMA method shall verify that the AI lifecycle data supplier have at least specified, verified and justified the AI lifecycle data has the following properties in the operational domain of the system of interest regarding the qualified performance, safety & security objectives applicable to this system-of-interest

- Suitability: The AI lifecycle data should be appropriate for the intended use and meet the specific requirements of the critical system.
    - Note: Suitability has been preferred over representativity - The relevance of the data depends on the ODD application. This could mean representativity if our goal is to train a calibrated model predicting balanced statistics. It can also mean exhaustively when we aim to over-represent rare phenomena and classes in order to improve the ability to detect them.
    - Note: In contrary, for a road map, the suitability is a synonym of representatitivy, because all the features (road signs, road marking, etc..) shall be present in the road map.

- Accuracy: The AI lifecycle data should be accurate and have a known level of uncertainty. The accuracy of the reference data should be justified regarding the performances of the ARTS and the applicable regulations.
    - Note: For the position of a micro vehicle, a map with a 10 m precision is insufficient. If road map is used, it icludes accurate road map road features (road marking positioning and type, width of lane, number of lane, curvature, …)
- Acquisition Repeatability: The AI lifecycle data acquisition should produce consistent results when used repeatedly under the same conditions.
- Acquisition Reproductibility: The AI lifecycle data acquisition should produce consistent results when used by different operators/annotators or in different laboratories.
- Traceability: The AI lifecycle data should be traceable to a recognized standard or calibration process, or appropriate observation which ensures that the data is reliable and trustworthy.
- Stability: The AI lifecycle data should remain stable over time if no changes to the data is required, without significant changes in its properties, such as composition or physical characteristics.
- Robustness: The AI lifecycle data shall remain stable over time en in case of occurrence of expected disturbances, critical or hazardous events.
- Completeness & identification of missing data: The AI lifecycle data shall be complete with regard to the objects and space represented by these data.
    - Note: The resilience is not a property of the reference, but a property of the system using the reference (independantly of the nature of the system, it can be a technical system like an ARTS or a process like the AI training process).

The PRISSMA method shall verify that the AI lifecycle data supplier have at least specified, verified and justified the the AI lifecycle data has the following properties in the operational domain of the system of interest regarding the qualified performance, safety & security objectives applicable to this system-of-interest

Suitability: The AI lifecycle data should be appropriate for the intended use and meet the specific requirements of the critical system.

Note: Suitability has been preferred over representativity - The relevance of the data depends on the ODD application. This could mean representativity if our goal is to train a calibrated model predicting balanced statistics. It can also mean exhaustively when we aim to over-represent rare phenomena and classes in order to improve the ability to detect them.

Note: In contrary, for a road map, the suitability is a synonym of representatitivy, because all the features (road signs, road marking, etc..) shall be present in the road map.

Accuracy: The AI lifecycle data should be accurate and have a known level of uncertainty. The accuracy of the reference data should be justified regarding the performances of the ARTS and the applicable regulations.

Note: For the position of a micro vehicle, a map with a 10 m precision is insufficient. If road map is used, it icludes accurate road map road features (road marking positioning and type, width of lane, number of lane, curvature, …)

Acquisition Repeatability: The AI lifecycle data acquisition should produce consistent results when used repeatedly under the same conditions.

Acquisition Reproductibility: The AI lifecycle data acquisition should produce consistent results when used by different operators/annotators or in different laboratories.

Traceability: The AI lifecycle data should be traceable to a recognized standard or calibration process, or appropriate observation which ensures that the data is reliable and trustworthy.
Stability: The AI lifecycle data should remain stable over time if no changes to the data is required, without significant changes in its properties, such as composition or physical characteristics.
Robustness: The AI lifecycle data shall remain stable over time en in case of occurrence of expected disturbances, critical or hazardous events.
Completeness & identification of missing data: The AI lifecycle data shall be complete with regard to the objects and space represented by these data.

Note: The resilience is not a property of the reference, but a property of the system using the reference (independantly of the nature of the system, it can be a technical system like an ARTS or a process like the AI training process).

The PRISSMA method shall verify that human annotations used for qualified CCR reference data have followed a qualification process to asses the following properties of the annotation:

Accuracy of annotations: an expert supervise the annotation
Repeatability of annotations: qualification intra annotator
Reproductibility of annotations: qualification inter annotator
Traceability: record of the identity of the annotator

Note: Example of human annotation qualification process:

Define the annotation guide that the annotators will follow to limit human cognitive bias and justify how to control the influence factors and qualify this annotation guide
Selection of annotators
Training of annotators
Qualification of annotators by semantic and syntactic verification on a first sample of each annotator separately in order to eliminate or correct defects
Possibly a second qualification phase on a second sample according to the results of the first qualification phase 6) Global annotation of the database
Global annotation of the database
Inter- and intra-annotator qualification on the complete database
Adjustment of annotations following the qualification of the complete database (deletion of uncertain annotations...

Note: Syntax verification involves scrutinizing the 'form' of the data to ensure it is computationally tractable, meaning it is coherent and uniform. Semantic verification, which can be of various types, includes manual checking of annotations made on a random sample for each annotation source. The objective is to ensure these are in alignment with the guidelines provided in the annotation guide. Intra-annotator verification aims to confirm an annotator's consistency in their annotation work over time. Inter-annotator verification, on the other hand, seeks to ensure that annotations among different annotators are mutually coherent.

4.1.2 Interpretability

The PRISSMA method shall verify that the AI supplier's has provided the interoperability justification documentation that demonstrate the interpretability of the AI ouputs by domains experts, including the justification of any function related to the supply of information required for the interpretation of the outputs (like a logging sytem or additional information about the output).

4.2 Performance Engineering Process on systems non including AI

The ARP 4754 standard provides a full framework for safety performance follow up from the very early phases, where failure configurations are listed, and all catastrophic situations (aircraft collision or clash) are allocated at 10^{-9} occurrence probability per hour target.

From these assumptions, fault trees are built or generated if a Model Based Safety Analysis approach is performed, and minimal cut of all these Fault Trees are calculated, providing the combinations of hardware failures or software errors sufficient to produce the different dreaded events.

Then analytical relations exist between probability of the event and failure distribution law parameters of the hardware components:

That's why for each dreaded events a dependency function exists:

Probability= $f(law_1, law_2, \ldots, law_n)$ where law_i is the probability of failure not detected of equipment "i".

This top down allocation process produces an allocation results on each equipment in terms of inherent reliability allocated and testability parameters as well, which can be noted:

$Law_i = f_i^{-1}(Top\_event\_probability)$

The same way, logical relation does exist between Boolean equation of each event and the way software error causes "se_j" are combined with hardware failure modes "fm_i": $Top\_event = boolean\_equation(fm_1, fm_2, \ldots, fm_i; se_1, se_2, \ldots, se_j)$

And regarding software errors, composition algorithms exist to allocate "Development Assurance Level", that is to say, DAL levels on every piece of software.

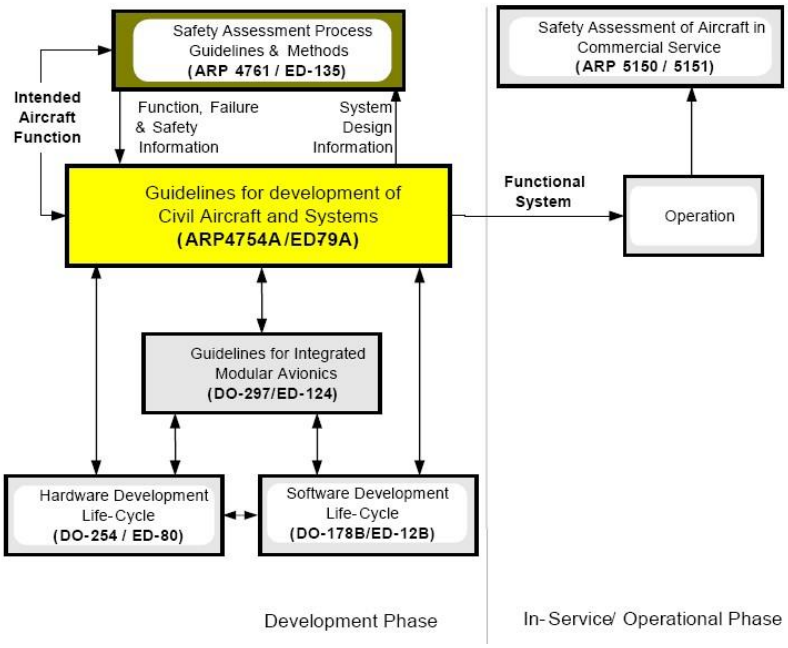

Figure 12: Framework of aeronautics standards

The frameworks of aeronautics standards provide top-down allocation and bottom-up verification performance process on Hardware and Software contributors to hazardous events.

Let's notice that railway (ISO 61508) and automotive framework (ISO 26262) have provided similar safety performance top-down allocation and bottom-up verification process.

4.3 AI bricks performance management in W cycle

So far, no standard RAMS performance (Reliability, Availability, Maintainability, and Safety) can naturally be defined for an AI software.

From an academic point of view, the following performances are applicable and are subject to current R&T projects working about computation theories, methods and technics:

- Relevancy performance: False Positive percentage (FP), False Negative percentage (FN) and combinations of these rates

- Steadiness: ability of the brick to behave continuously depending on the input variations without incoherent behavior

- Resilience: ability to keep a correct behavior when input are unsteady or slightly beyond applicability domain regarding the input

- Explainability: ability to explain and justify logically behavior of the brick depending on the input

- Interpretability: ability to understand and interpret behavior of the brick depending on the input with a human point of view

- Coverage Rate: percentage of use cases "well covered" given a framework of reference scenarios…

  For many of these KPIs, very sophisticated mathematical methods may be used such as Topological Data Analysis, Abstract Interpretation, Adversarial Attacks…

	Condition (as determined by "gold standard")		
	Condition positive	Condition negative	
Test outcome positive	True positive	False positive (Type I error)	Positive predictive value = Σ True positive/ Σ Test outcome positive
Test outcome negative	False negative (Type II error)	True negative	Negative predictive value = Σ True negative/ Σ Test outcome negative
	Sensitivity = Σ True positive/ Σ Condition positive	Specificity = Σ True negative/ Σ Condition negative	

Figure 13: Confusion matrix

Let's notice that System Engineering process of AI softwares rather refers to a W shaped process than a V shaped process as is illustrated by the following figure:

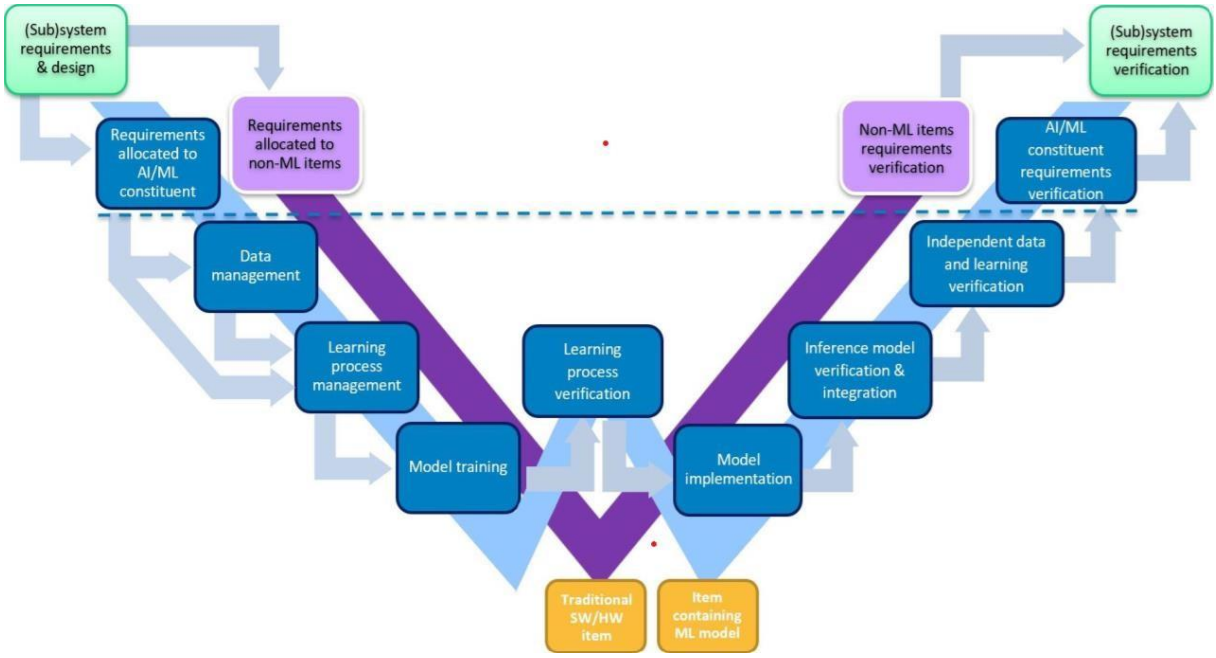

Figure 14: Development Process cycle of AI applications in APSYS Digital Transformation Office

In Top-Down front end analysis phases, Validation & Verification Tools have to assess completeness and representativeness of Data Set.

In the meanwhile, methods and tools for quantification and generalization guarantees have to be provided concerning Machine Learning and Deep Learning applications, if AI softwares under analysis refer to these technics.

In the Bottom-Up phase, Methods and Tools for the verification of ML algorithm and model robustness and stability have to be deployed.

4.4 Semantic Gap between System / SoS level and AI brick

ISO 26262 framework illustrates the approach that can be adopted to integrate safety analysis in system engineering process:

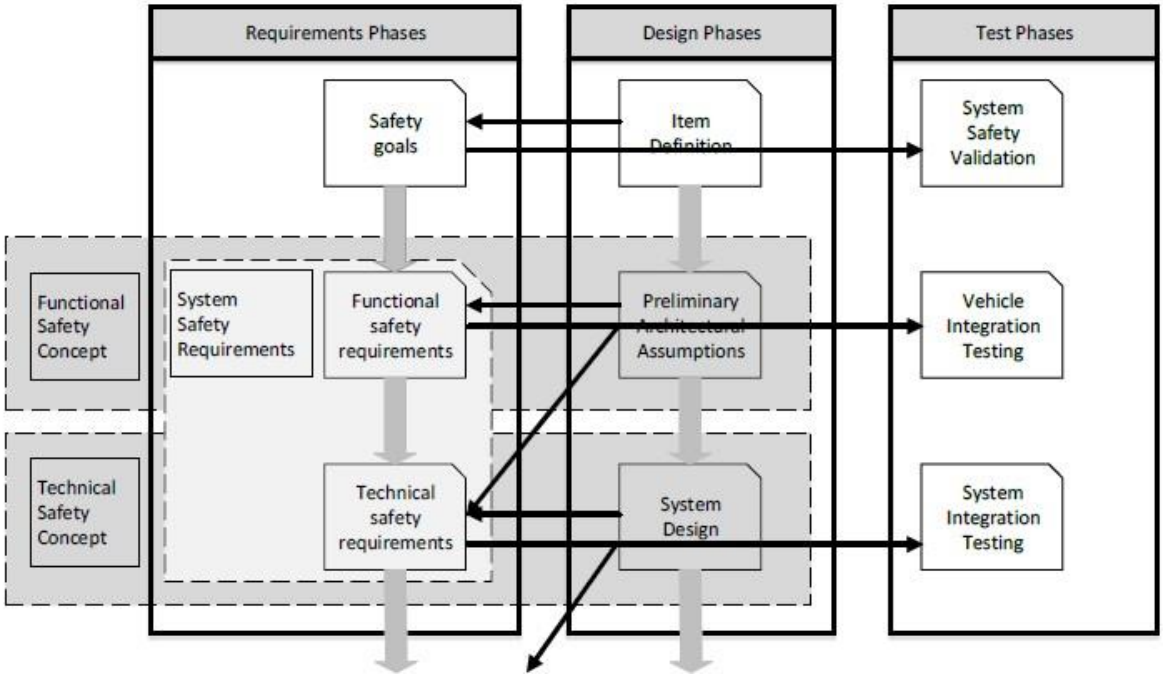

Figure 15: Safety engineering development process in automotive industry

In this framework, one has to allocated SG « Safety Goal » in FSR « Functional et System Safety Requirements », and then in « Technical Safety Requirements », by highlighting « Functional Safety Concept », but also « Technical Safety Concept ».

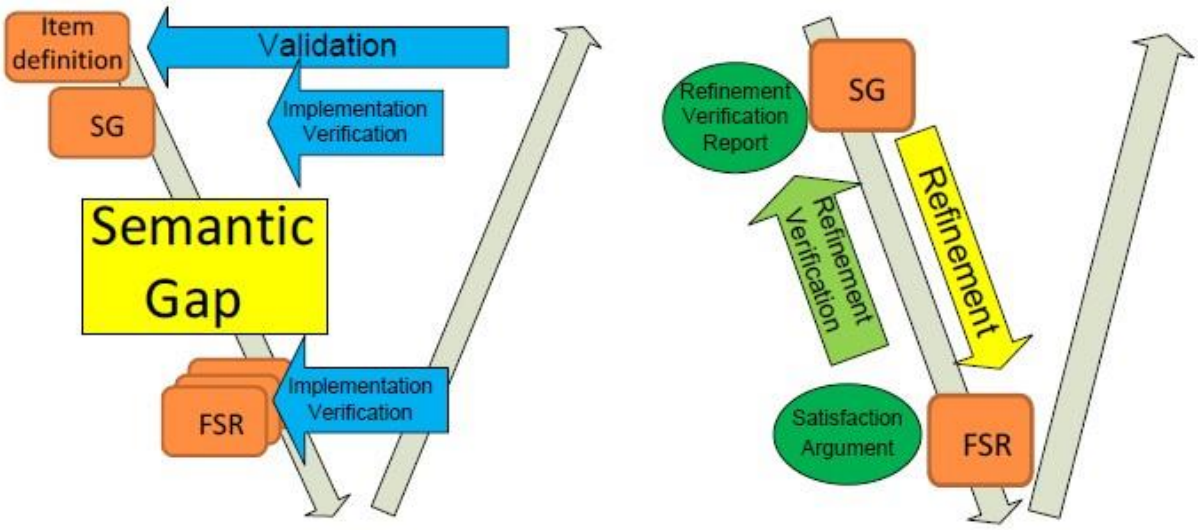

Figure 16: Approach Breakdown due to Semantic Change

In the refinement approach switching from Safety Goal to Functional Safety Concept, one observes what Rolf Johansson calls a semantic gap.

Indeed, what is meaningful or relevant for an autonomous vehicle, is not to satisfy with a list of « Functional Safety Requirement » spread out in an enumerative way in something like a checklist; but it consists in showing a relevant behavior in dynamic scenarios, in interaction with many actors, which are much better demonstrated on a virtual way or hybrid way with a simulation platform or a digital twin for example…

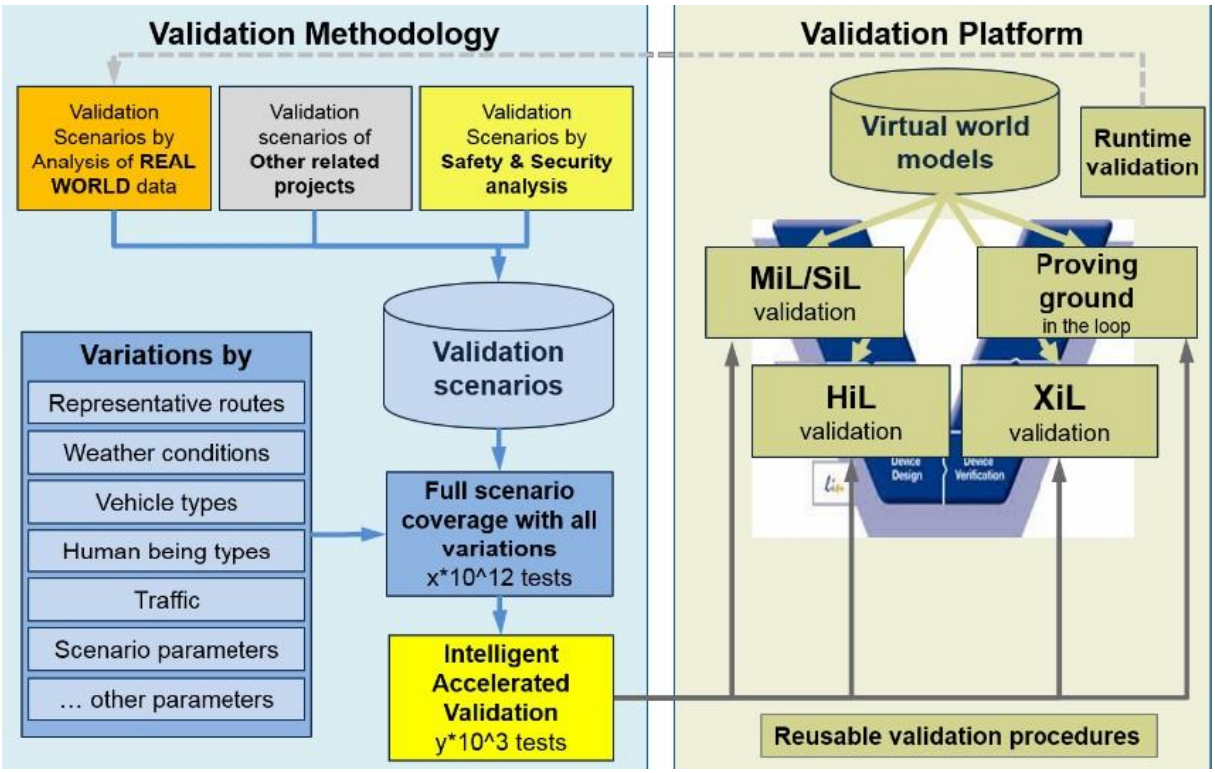

Figure 17: Approach Breakdown

4.5 An attempt to solve this gap…

An hybrid approach is proposed by French Department of Defense (DGA) for systems integrating of AI applications.

A PHA (Preliminary Hazard Analysis) and HA (Hazard Analysis) is performed on the System of Interest to identify every possible contribution of AI bricks on Hazard production.

Depending on the level of severity of the Hazard, and the controllability of the situation by the end user, different value thresholds of performance metrics (cf §3.3) can be envisaged to warranty low likelihood of a possible occurrence of an AI misbehavior causing the hazard:

Exigences / Catégorie	Métrique/cas 1		…	Métrique/cas N	
	Jeu de données de test	Jeu de données de qualification		Jeu de données de test	Jeu de données de qualification
Non critique	$> S^1_1$	$> S^1_1$ - n%	…	$> S^N_1$	$> S^N_1$ - n%
Peu critique	$> S^1_2$	$> S^1_2$ - n%	…	$> S^N_2$	$> S^N_2$ - n%
Semi-critique	$> S^1_3$	$> S^1_3$ - n%	…	$> S^N_3$	$> S^N_3$ - n%
Critique	$> S^1_4$	$> S^1_4$ - n%	…	$> S^N_4$	$> S^N_4$ - n%
Très Critique	$> S^1_5$	$> S^1_5$ - n%	…	$> S^N_5$	$> S^N_5$ - n%

Figure 18: PHA matrix

Remark: criticality of the hazard is a synthesis between the severity of the hazardous situation and the controllability of the consequence appearance, and 5 discrete levels are proposed by the French DOD standard.

Each set of columns is associated with a possible metric when it is considered as applicable and meaningful for the AI brick, and threshold values have to be provided by R&T projects of the State Of The Art.

4.6 Towards an incremental engineering process

Multiple occurrences of iteration loops in the operation of development cycle for AI application strongly encapsulates different points of view enhanced by diversity of metrics able to be assessed.

Data Set framework has cautiously to be traced, optimized and targeted to relevant use cases, with an attention played to coverage of ODD or functional scopes where the software has to be qualified:

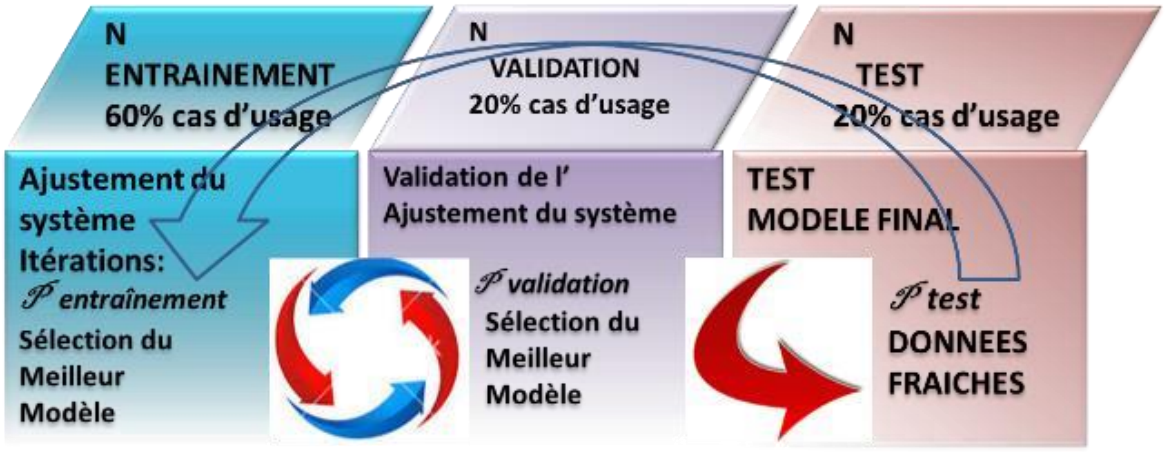

Figure 19: Cross Validation process of Data Set

At the end of the day, multiple loop iterations will have to be processed to reach expected performances without degradation, bias nor over fitting:

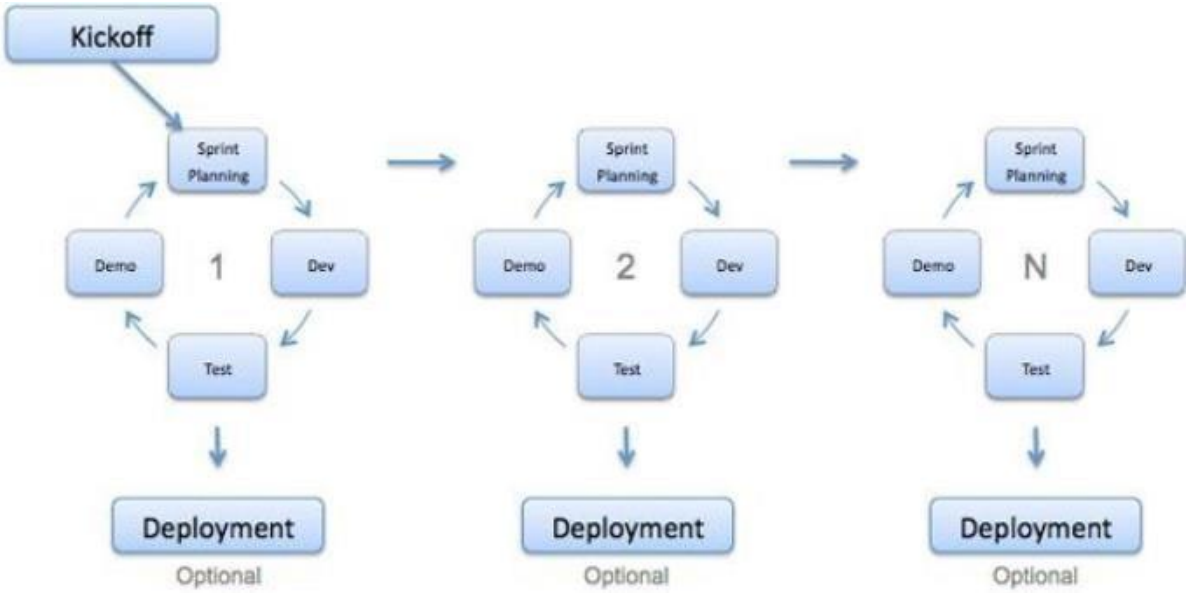

Figure 20: Illustration of the iteration process

5 Conclusion

This document has introduced some of the main aspects of the AI impacting the system engineering process and the major hypothesis regarding these impacts.

Some paradox may still be present when applying the classical system engineering process to AI based system, noticeably on the verification & validation activities. The SOTIF principles is the state of the art for the commissioning of a vehicle and assessing the safety of the intended functionality beyond the classical functional safety applied to well-known functions of autonomous vehicle.

Still the objective evaluation of the performance of the IA system is still a challenge that needs to be addressed for the safe deployment of IA based autonomous driving systems.

6 References

[1] Presentation of the SOTIF (ISO/DIS 21449) for the PRISSMA Project by C. Bohn