PRISSMA – Plateforme de Recherche et d’Investissement pour la Sûreté et la Sécurité de la Mobilité Autonome

PRISSMA – Platform of Research and Investment for Dependability and Safety of Autonomous Mobility

ARBARETIER Emmanuel
AIRBUS PROTECT
emmanuel.arbaretier@airbus.com

GRUYER Dominique
Université Gustave Eiffel, Champ-sur-Marne
dominique.gruyer@univ-eiffel.fr

REGNIER Rémi
LNE
remi.regnier@lne.fr

JOUVE Pierre
STRMTG
pierre.jouve@developpement-durable.gouv.fr

PIPERNO Alain
UTAC
alain.piperno@utac.com

GAVA Cédric
SPHEREA
cedric.gava@spherea.fr

HADDAD Sammy
OPPIDA
samy.haddad@oppida.fr

JOINVILLE Thibaut
CEREMA
thibaut.joinville@cerema.fr

BERTAULT Véronique
RATP
veronique.bertault@ratp.fr

NOUACER Reda
CEA
reda.nouacer@cea.fr

BOHN Christophe
IRT-SYSTEMX
christophe.bohn@irt-systemx.fr

GUYONNAUD Marie Françoise
ATC France Paris2connect
mguyonnaud@smartuse.org

TAHIROU Youssoufa
VALEO
youssoufa.tahirou@valeo.com

DESOUSA FERNANDES Rafël
UTAC
rafael.desousafernandes@utac.com

LAUGIER Christian
INRIA
christian.laugier@inria.fr

Résumé

Le projet PRISSMA vise, en réponse à l'appel lancé par le Grand Défi IA du Conseil de l'Innovation et le Ministère de la Transition Ecologique et Solidaire, à proposer la définition et la conception d'une méthodologie, de procédures, d'outils, et de plateformes permettant l'évaluation, la validation, et potentiellement l'homologation de systèmes de systèmes et de systèmes à base d'IA impliqués dans les moyens et les services de mobilité automatisée ou autonome (MMAA). Regroupant une vingtaine de partenaire français, le projet PRISSMA se place particulièrement dans le cadre de la « démonstration de sécurité » des moyens de mobilité utilisant des techniques d'Intelligence Artificielle (pour l'évaluation et la validation de la sécurité et de la sûreté des MMAA).

Ce projet propose un référentiel d'activités structuré (spécification processus, définition et élaboration de domaines d'emplois, définitions des exigences, audits, mise en œuvre de campagnes de tests virtuels, contrôlés ou réels…) et des méthodologies adaptées dont la pertinence est illustrée au travers de leur application à des preuves de concept sur des expérimentations.

Enfin, PRISSMA élabore une méthode de démonstration globale de la sécurité d'un MMAA.

Mots clés *- Système de systèmes, IA, système de mobilité automatisée ou autonome, chaîne outillée, évaluation, validation, vérification, homologation, automatisation de la conduite, simulation, scénarios, capteurs, modélisation, réglementation, normalisation.*

Summary

This document describes PRISSMA project consisting in defining verification, evaluation, and validation methodologies, procedures, and protocols allowing to calculate both the tools/models and system of systems/ AI-based components levels of performance/quality/representativeness. This methodology has been applied on several POCs (Proof Of Concept) dealing with different kinds of automated shuttles circulating in different environments, as well in a 100% virtual model, as in a real or controlled environment. These POCs have demonstrated how a methodology funded on three pillars (virtual testing, test campaigns in controlled environment, test campaigns in real environment) may assess and validate Safety performances of an automated shuttle.

Special attention has been dedicated in this paper to virtual testing process with four specific POCs, and tool chain required to perform this simulation process. The whole set of different kinds of models necessary to produce for these simulations is also described and the way to validate them. Finally, a set of relevant and adapted metrics and KPI is proposed and presented, to quantify Safety of automated shuttle.

Another development is dedicated to AI (more precisely Machine Learning) qualification or certification process elaborated ²by EASA (European Agency for Safety in Aeronautic) from which PRISSMA has got some inspiration to describe how AI based application (like autopilot) may be qualified and validated before they are integrated in a whole system.

Key Words – *System Of System, AI, Automated or Autonomous Transportation System, Simulation Workbench, IVVQ, ADAS, simulation, scenarios, sensors, modeling, regulation*

I. INTRODUCTION

PRISSMA is a project financed by French state and more specifically the Ministry of Ecological Transition through DGITM and DGEC, consisting in developing methods, technics and tools for validation safety, reliability and security of road transport systems including AI modules, like automated or autonomous vehicles or buses and shuttles.
In this paper, we designate by ARTS the expression "Automated Road Transportation System".

Following work packages have been defined to structure work performed:

WP1: how to validate and qualify an AI brick or module, and to specify test to demonstrate its performances
WP2: how to perform virtual simulations to validate and demonstrate safety of an ARTS?
WP3: how to perform controlled test to validate and demonstrate safety of an Automated Road Transportation System?
WP4: how to perform real tests to validate and demonstrate safety of an Automated Road Transportation System?
WP5: how to demonstrate cyber security performances of an Automated Road Transportation System?
WP6: how to produce an IVVQ (Integration Validation Verification Qualification report for an ARTS integrating WP1-5 input?
WP7: how to perform a Life Cycle Management Process for an ARTS in terms of maintenance, diagnosis and correction?
WP8: how to integrate output of international community, eco systems and regulation framework?

Following workflow shows how these work packages depend the ones on the others and have generated complementary elements to a general methodological framework of safety demonstration for an ARTS consisting in Audit, Virtual Simulations, Controlled test campaigns on artificial runways, Real Test campaigns in the real environment.

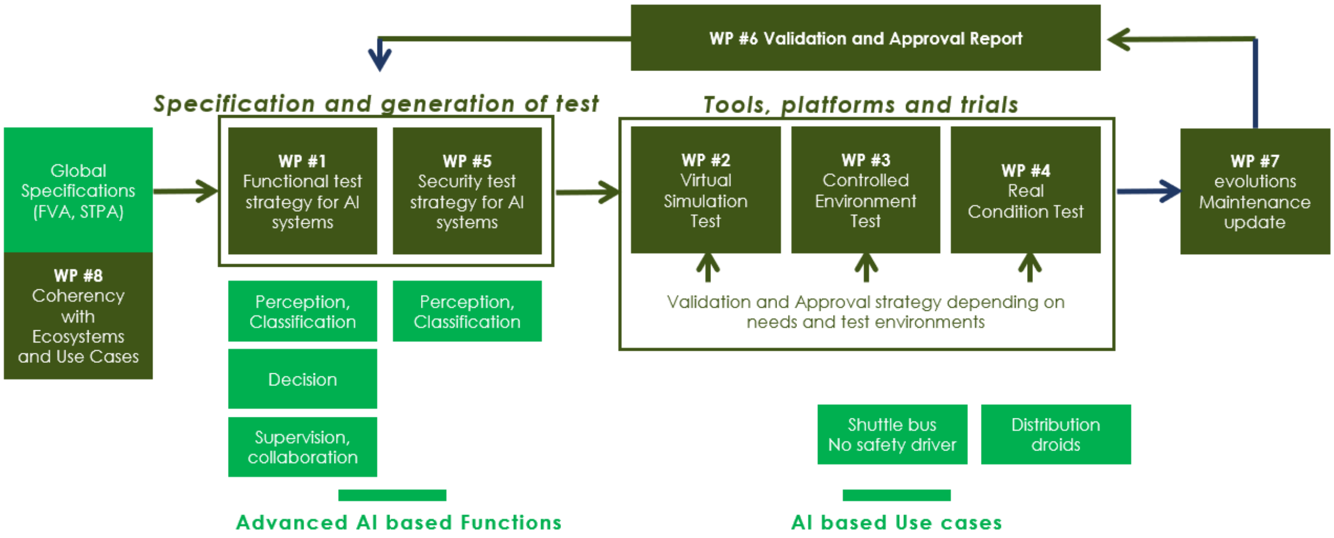


Figure 1: Structuration of PRISSMA project

In this paper we will introduce some highlight on specific points of this global framework which will be ideal to support Safety assurance during the whole Life Cycle of AI (Artificial Intelligence) including systems and especially ARTS (Automatic Road Transportation System).

Because validation of such AI based systems mainly depends on Virtual Simulation Test campaigns, we will rather develop how WP2 has been performed, but also make a quick overview of WP1 contribution which shows how AI applications may be validated and qualified separately, before they are integrated in a Main System (Automated Shuttle in this paper), and global validation is operated then on the whole system.

II. IMPORTANCE OF VIRTUAL TESTING

Testing in real or controlled environment is unavoidable, above all to provide "ground reality" for virtual testing. However Virtual testing is introduced to reduce the burden of physical tests and effectively provides evidence on the AI performance across the operational domain of an Automated & Autonomous Road Transport System (ARTS) (Connected Autonomous Vehicle). Virtual testing, evaluation, validation, and certification enter a specific design plan adapted from the V-cycle, which is the reference to present the design life cycle of a product such as an ADAS (Advanced Driver Assistance System) or an Automated Driving System (ADS) (Advanced Driving System) as shown in Figure 2. The validation stream is always related to the specification stream, meaning that validation plans are designed concerning the specifications. However, specifying and validating complex systems of systems such as a Connected Autonomous Vehicle (CAV) is a challenging process. To operate validation plans showing a suitable level of safety and reliability with an acceptable time and budget, virtual method tests from MIL (Model-In-TheLoop) to VIL (Vehicle-In-The-Loop) now complement physical testing: closed site tests and open road tests.

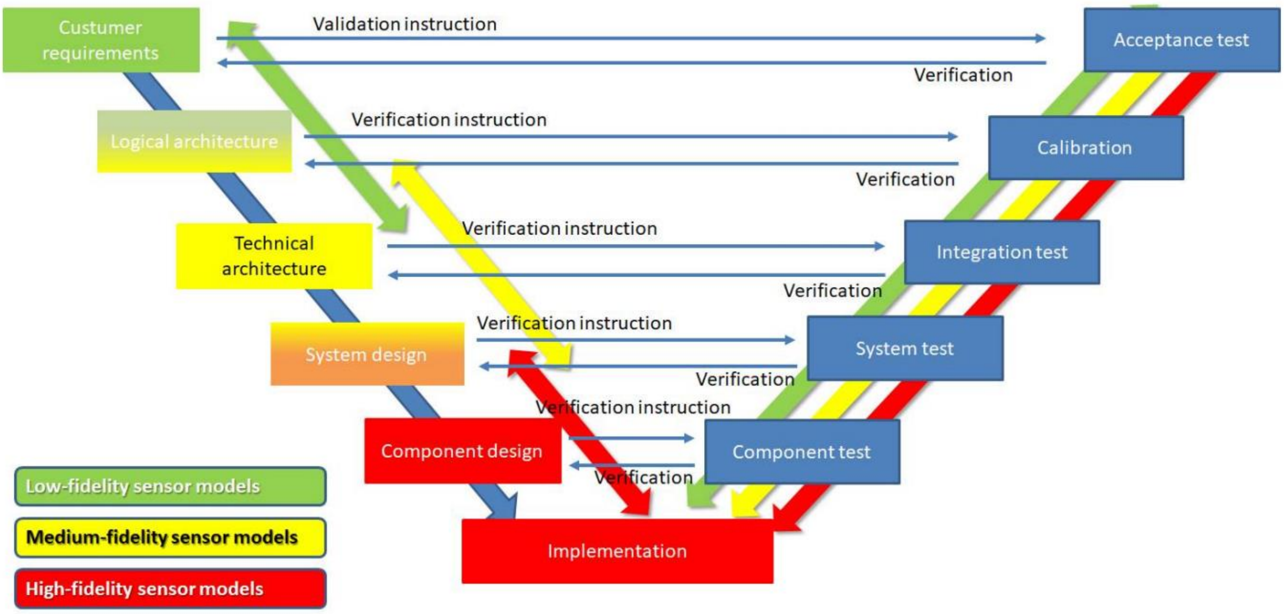


Figure 2: V-cycle for virtual prototyping, test, evaluation, and validation

This approach addresses these aspects with 2 points of view:

• The first one is focused on the Verification/Validation of virtual test means for their qualification and performance assessment. Virtual test means involve tools as well as models.

This work has been done in WP2. More accurately, this evaluation has gone through the definition of the criteria and metrics allowing both the verification and the evaluation of the simulation platforms as well as the qualification of the level of realism, the relevance and the validity of the ground truths established for AI modules.

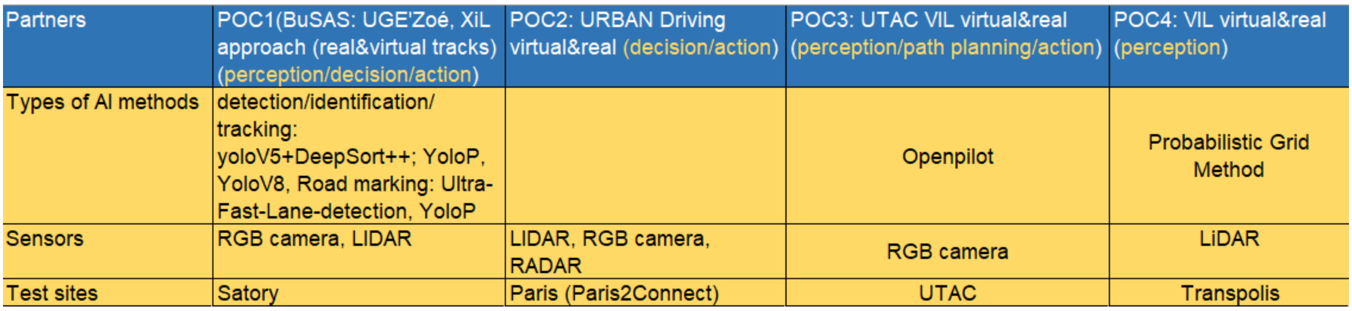
For this, this work package has proposed to define a set of indicators validating the representativeness of the simulation for the evaluation of ARTS. Concerning the formal approaches, a verification of their properties on the models has been proposed. This has involved the generation of conformance tests, and the validity of the simulation. These actions and relevance of the proposed methodology, procedures, and protocols have been tested with the 4 Proof of Concept defined and presented in WP2.

• The second one proposes procedures and protocols for the evaluation and validation of AI-based systems under test and using simulation tools (verified, validated, and respecting the requirements established at the beginning of the program). This part corresponding to aims to propose procedures and protocols for the evaluation and validation of critical AI-based systems and subsystems from specific evaluation environments described and characterized in the project to be representative. /

Tools, technics and models used in these procedures and protocols are themselves submitted to a validation process particularly on the level of realism, the relevance of simulation tools, and the quality of ground truths in terms of evaluation references. In addition, this task addresses best practice procedures and implementation of experimentation plans using simulation and co-simulation platforms. In order to enable the implementation of the evaluation and validation stages, this project has proposed definitions, criteria and evaluation metrics for AI-based systems based on previously mentioned collaboration with external AI ecosystem.

For the proposed simulation environments, an identification of the optimum scope of use (including dysfunctional through the injection of failures, etc.) has been carried out.

These actions have been tested on the four WP2 POCs defined and presented in following figure. The proposed evaluation and validation process has taken into account metrics and KPI (for example on performance, security...) adapted to the level of complexity linked to the applications (AI-based systems, system of systems, communication and cyber security systems) and targeted objectives



Partners	POC1(BuSAS: UGE'Zoé, XiL approach (real&virtual tracks) (perception/decision/action)	POC2: URBAN Driving virtual&real (decision/action)	POC3: UTAC VIL virtual&real (perception/path planning/action)	POC4: VIL virtual&real (perception)
Types of AI methods	detection/identification/ tracking: yoloV5+DeepSort++; YoloP, YoloV8, Road marking: Ultra-Fast-Lane-detection, YoloP		Openpilot	Probabilistic Grid Method
Sensors	RGB camera, LIDAR	LIDAR, RGB camera, RADAR	RGB camera	LiDAR
Test sites	Satory	Paris (Paris2Connect)	UTAC	Transpolis

Figure 3: Table summarizing the four WP2 POCs implemented in PRISSMA project and used to test the PRISSMA's verification/evaluation/verification methodology.

III. VALIDATION BY SIMULATION

When it comes to simulation-based assessment, the first thing to do is to define the tools required for the simulation platform. In the PRISSMA project, this important task has been addressed by WP2. PRISSMA has recognized that the choice of implementation of this platform is mainly based on the preliminary choices of requirements, ODD (Operation Design Domain), use cases and scenarios to be addressed. From there, you can choose the different tools, models, engines and platforms you need. The final stage consists in choosing the evaluation and validation tools, both for the system under evaluation and for the sub-study platform.

9

ODD defines the operating conditions under which a vehicle's automated driving systems can be safely engaged: it can address among others infrastructures, environmental and atmospheric conditions, circulation trajectories, traffic context, etc.

A. Software tools supporting vehicle simulation

Tools constituting this kind of simulation platform are submitted to different requirements:

- Some of the metrics used in a camera simulation model may be: Effective Focal Length, distortion, Effective Focal Length Color Fidelity and Sensitivity and consistency, Noise (SNR, Dynamic Range, Tonal Range) or White Balance Accuracy et Color Contrast…

- Some other metrics may be relative to the image generation using the road infrastructure and the environment conditions (weather, light). This second sub part validation is a critical stage for not only the verification and validation of the camera model and its capability to generate realistic data, but for the simulation engine and its capabilities to generate the effects allowing to generate a realistic rendering.

For a couple of years, a significant number of synthetic datasets have emerged to overcome the lack of real data with the objective of creating increasingly large and realistic synthetic datasets, which is crucial for AI-based algorithms. But to what extent can we say that the computer-generated images are faithful to reality? Then it is essential to propose some metrics to quantify the level of fidelity of these kinds of images. Plenty of metrics already exist to quantify image quality but none to quantify the fidelity of synthetic images.
The term fidelity is preferred to realism due to its subjective nature, making quantification a complex task. Additionally, a comprehensive conceptual framework of fidelity has already been proposed in our project.

In the PRISSMA methodology for simulation verification and validation, the fidelity can refer to the extent of similarity between some selected features in the virtual environment and their corresponding reference features in the real environment. Therefore, high-fidelity simulations can correspond to a faithful representation of the real environment's features, while low-fidelity simulations may correspond to a simpler representation of these features in comparison to the real world.

The figure below illustrates the comprehensive diagram of the proposed verification method, which consists of introducing a set of metrics to quantify the fidelity of synthetic images. Learning and statistic-based approaches are used to exploit the image information such as textures and others that are relevant to real images.

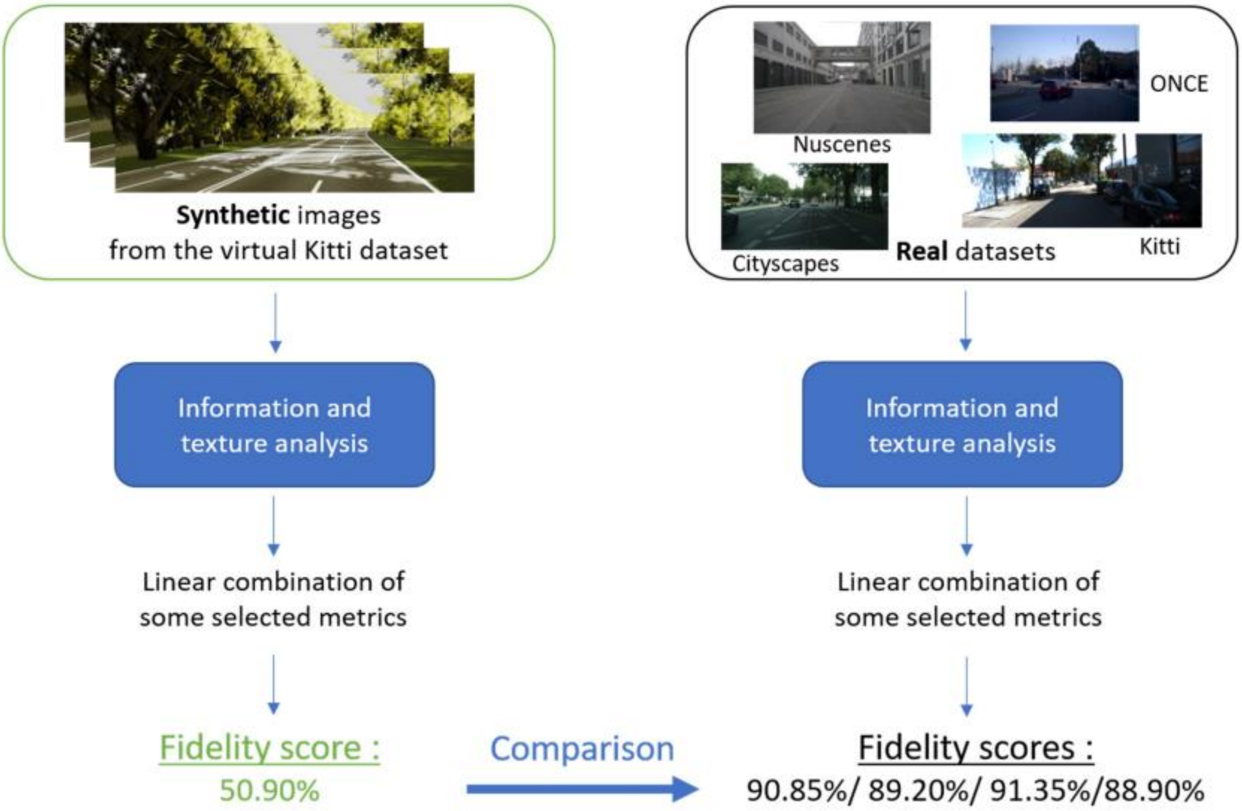


Figure 4: Diagram of the proposed verification method to assess the fidelity level of a virtual image coming from a virtual camera

A linear combination of these metrics allows to computes the fidelity scores for the synthetic datasets. The same process is applied to the real datasets to provide reference scores, giving an indication of the fidelity level of synthetic images. This method has shown promising results, but it needs to be extended to various types of scenes. The experiments were conducted using data from urban areas under clear daytime conditions. The next step is to apply this method to different scenes and under adverse weather conditions. The image information will vary significantly depending on scenarios, like under foggy conditions.

Generating a score is essential to determine whether a virtual data set will be sufficiently representative and true to reality to be used in learning, assessment, and validation procedures.

To illustrate, Figure 5 shows an example of the simulation environment set up for PRISSMA evaluation process:

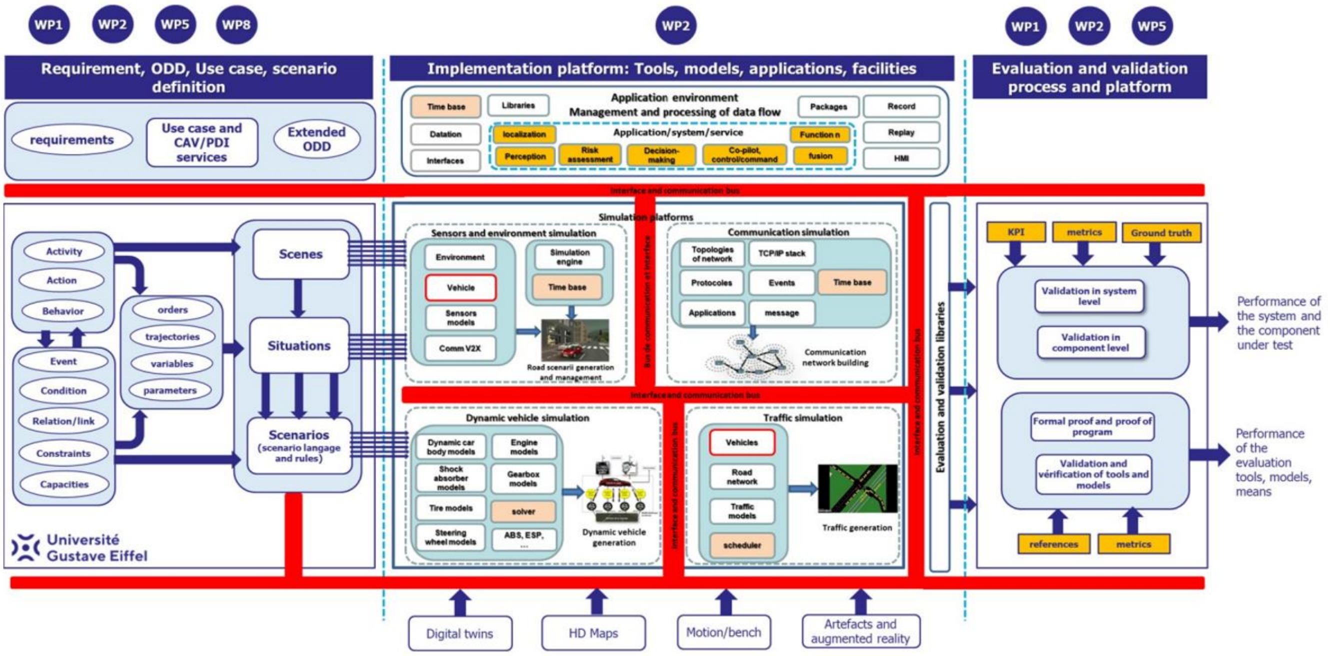


Figure 5: Global view of the simulation environment for evaluation process with its systems, functions, and component

Previous figure highlights three fundamental pillars of simulation process:

- on the left part: preparation of all necessary input for simulation preparation
- on the central part: tool chain activation
- on the right part: post processing of the simulation

For information, dependency on other Work Packages has been represented on the top of the diagram (purple circles), as well as the references of WP2 tasks concerned by the phase (orange squares).

B. Models required to perform vehicle simulation

In following, we enumerate most kinds of models and components commonly used in vehicle simulations:

- Vehicle dynamics model:

– Multi-body dynamics model: represents the mechanical systems of the vehicle, including suspension, chassis, steering, tires, and drivetrain. It simulates the movement, forces, and interactions between these components based on Newtonian physics principles.

– Powertrain model: simulates the engine, transmission, and other powertrain components, accounting for torque, gear ratios, fuel consumption, and performance characteristics.

- Environment Model:

– Terrain model: represents the road surface and terrain features, including elevation changes, curvature, friction characteristics, and surface irregularities affecting vehicle dynamics.

– Weather and environmental conditions model: simulates weather conditions (such as rain, snow, fog), lighting, visibility, and other environmental factors that influence vehicle performance and handling.

- Sensor Models:

– LiDAR model: simulates Light Detection and Ranging sensors that use laser pulses to measure distances and create 3D point clouds, commonly used for perception in autonomous driving simulations.

– Radar model: emulates Radar sensors that use radio waves for object detection and speed measurement.

– Camera model: simulates cameras for visual perception and computer vision tasks, including object recognition, lane detection, and traffic sign recognition. Of course, the models can be adapted to suit all different camera technologies (cyclop, infrared, RGB, fisheye, event-based camera...).

- Navigation models:

- GPS model: refers to a simulated representation of a Global Positioning System (GPS). A GPS model in a vehicle simulation includes the following aspects: satellite constellation simulation, signal propagation and reception, position calculation algorithms, error modeling, accuracy and uncertainty Estimation and integration with vehicle dynamics.

- INS model: inertial Navigation System model. An Inertial Navigation System (INS) is a navigation aid that uses a computer, motion sensors (accelerometers), and rotation sensors (gyroscopes) to continuously calculate the position, orientation, and velocity of a moving object without external references such as GPS. This model includes inertial sensors models (accelerometers measure linear accelerations, while gyroscopes detect angular velocities) and can include also integration algorithms, error characteristics and calibration procedures

- Odometer model: an odometer model refers to a simulated representation of an odometer, the instrument used to measure the distance travelled by a vehicle.

•Sensors deployed on the infrastructure: as part of a system of systems, where the infrastructure plays an important role, we need to add models for all the sensors and remote equipment that communicate with the vehicle. In this context, V2X communication must also be modelled: V2X represents all mechanisms supporting information exchange between automated vehicle under consideration and all other vehicles

• Control Systems Model: represents the electronic control units (ECUs) and control algorithms responsible for vehicle stability, traction control, anti-lock braking systems (ABS), and other advanced driver assistance systems (ADAS). • Driver Behavior Model: simulates human drivers' behavior, including decision-making, reaction times, and driving styles, which influences vehicle operation and response in the simulation.

• Traffic Model: simulates other vehicles, pedestrians, and entities interacting with the simulated vehicle. It includes models for vehicle movement, traffic patterns, and interactions with the environment. • Simulation Framework: Provides the infrastructure to integrate and manage different models, components, and simulations in a cohesive environment. This includes simulation rendering and physics engines.

• User Interface and Visualization Tools: interfaces for users to interact with the simulation, visualize data, and analyze results.

• Data Analysis, Validation and Calibrations Tools: software tools used to analyze simulation results, compare against real-world data, and validate the accuracy and reliability of the simulation models.

The generator of scenarios is crucial in building the framework, generating necessary configurations, and selecting algorithms for evaluation. It is responsible for generating configurations of evaluation scenarios based on Operational Design Domain (ODD) and Object and Event Detection and Response (OEDR). It also selects candidates of AI algorithms for the framework according to specific objectives, then evaluates and validates them based on a representative real-world dataset. Moreover, the generator component generates the configuration of the ground truth for the executor based on the selected algorithms, ensuring the accuracy and reliability of the evaluation process.

C. Integration of simulation activities into Verification, Evaluation and Validation process

The evaluation objectives of an AI-powered system in ADS are derived from an analysis of the system and its operating environment: this means a production of a Functional Analysis and Physical Work Breakdown Structure as well which will be integrated in the simulation platform, taking into account modeling languages proposed by the platform.

Initial performance objectives to assign encompass multiple levels:
- At the system level, the overall performance and quality of the AI system are evaluated in simulated environments: this can be expressed in terms of targeted number of accident per 1000 hours;
- The components/functionalities level focuses on evaluating specific functions and algorithms necessary to meet the expected functionalities of the system;
- Additionally, the scenarios level evaluates the system's capabilities within a defined ODD, including safe driving in different scenarios under varying conditions like non-optimal weather, traffic, and lighting. Categorizing the evaluation objectives into these levels facilitates a comprehensive evaluation of the system's performance, safety, and areas for improvement, offering valuable insights into its capabilities and limitations.

Scenario Definition required by simulation process, involves the conceptualization and specification of following fundamental elements:
- Scene contains the overall environment where the scene takes place, including:
  – Dynamic elements which are objects capable of movement or state changes, such as vehicles, pedestrians, or cyclists;
  – Static elements, which are stationary objects in the scene, such as road infrastructure or buildings;
  – Environment factors, which refers to the surrounding conditions, such as weather or lighting, which can influence the behavior of dynamic elements.
- Event represents incidents or occurrences that unfold during the scenario. These events can be pre-defined or dynamically generated and contribute to the scenario's progression. They include stimuli, triggers, or changes in the environment or state change of other objects (outside ego), shaping the sequence of actions and reactions within the scenario.
- Action pertains to the response or behavior exhibited by the ego object in the scenario. It demonstrates how the ego object in the scene reacts to events or encountered conditions. Actions may include acceleration, braking, or changes in the direction of the ego vehicle.
- Criteria refers to the specific conditions or standards required for the simulation scenario to be deemed complete or successful. These criteria could include factors such as reaching a particular time limit, accomplishing predefined objectives, meeting specific performance metrics, satisfying safety requirements, or any other relevant measures that define the desired conclusion of the scenario.

D. Verification, Evaluation, Validation scenario configuration

This part involves the implementation and customization of a scenario based on the definition. This process focuses on the detailed setup and arrangement of specific elements, conditions, and variables within the scenario. An effective scenario configuration should be done within the defined boundaries of ODD and OEDR (Object and Event Detection and Response). ODD contains the specific operating conditions and environments within which ADS is intended to function safely and effectively.

By considering the ODD in the scenario configuration, the scenarios accurately reflect the real-world conditions that the system is designed to encounter. This involves defining geographic boundaries, traffic conditions, and factors that influence the system's operational limits, thus ensuring the scenario's relevance and accuracy. OEDR focuses on the system's ability to detect and respond to specific objects and events within its operational environment. When configuring scenarios, it is imperative to define the types of objects the system should detect. Furthermore, the scenario should include events that the system should recognize and respond to, such as sudden lane changes, emergency braking, or any other relevant mapping. By incorporating these elements, the scenario enables the evaluation and improvement of the system's perception and response capabilities. By aligning scenario configuration with the ODD and OEDR, the resulting simulations accurately represent the operating boundary and allow for a comprehensive evaluation of ADS.

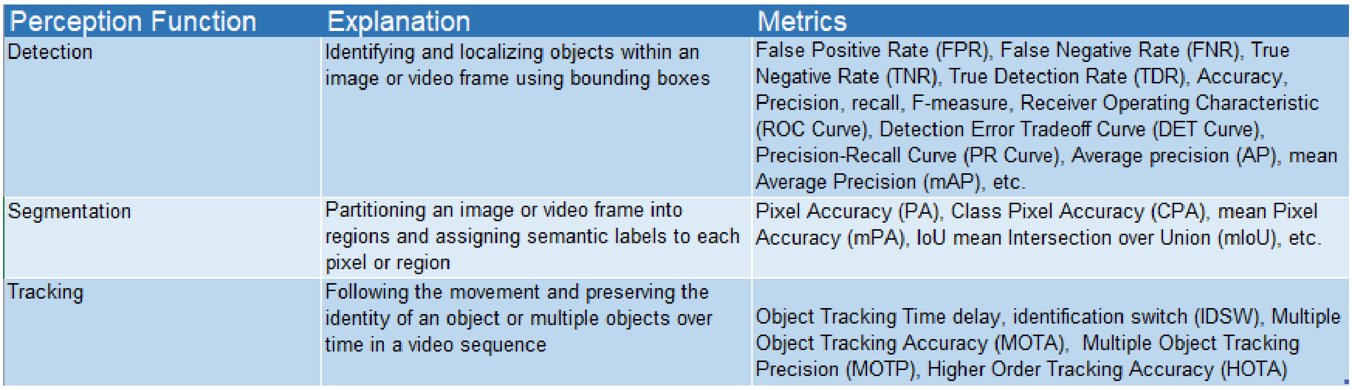
In order to evaluate the high-level quality of AI-powered system in ADS, expressed in terms of KPI values displayed in figure 6, such as a visual perception system, it is necessary to implement a full mobility service and propose relevant and representative scenarios involving an exhaustive set of conditions/configurations/situations allowing for quantification of the performances and the quality of the service.

Precise values of quantitative values or thresholds of metrics / KPIs, corresponding to what could be called "high quality of AI powered systems", are still under discussion. No value in absolute can yet been displayed.

The metrics (in the case of visual perception system) can refer to a set of specific Key Performance Indicators (KPIs):

- Risk specific: Longitudinal and lateral distance, Time to collision (TTC), Time Exposed Time-to-Collision (TET), Deceleration Rate to Avoid a Crash (DRAC), etc.
- Task (detection/tracking) specific: Success rate, Loss, Distance, etc.
- Time specific: Frequency, Time to detect/track, False alarm frequency.

An exhaustive list of these criteria can be found in the PRISSMA output deliverables and is summarized through Figure 6:



Perception Function	Explanation	Metrics
Detection	Identifying and localizing objects within an image or video frame using bounding boxes	False Positive Rate (FPR), False Negative Rate (FNR), True Negative Rate (TNR), True Detection Rate (TDR), Accuracy, Precision, recall, F-measure, Receiver Operating Characteristic (ROC Curve), Detection Error Tradeoff Curve (DET Curve), Precision-Recall Curve (PR Curve), Average precision (AP), mean Average Precision (mAP), etc.
Segmentation	Partitioning an image or video frame into regions and assigning semantic labels to each pixel or region	Pixel Accuracy (PA), Class Pixel Accuracy (CPA), mean Pixel Accuracy (mPA), IoU mean Intersection over Union (mIoU), etc.
Tracking	Following the movement and preserving the identity of an object or multiple objects over time in a video sequence	Object Tracking Time delay, identification switch (IDSW), Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Higher Order Tracking Accuracy (HOTA)

Figure 6: Example of metrics and KPIs for simulation post processing of STRA

IV. VALIDATION OF IA BASED APPLICATIONS

Before validating the whole vehicle, one has to consider that embedded AI based devices on board the vehicle have been validated independently.

A specific question to handle is thus the certification or qualification of the simulation tool for AI based application: are these tools independent of AI? For the purpose of this development, it has been supposed they are themselves containing AI in order to be closer to the application they are evaluating. Therefore, it seems like we are stuck in a repeat loop. How an AI based application, using AI based simulation tools, can be qualified? First, all the best practices used to develop the main AI application remain applicable for the simulation device. Secondly, the simulation must go through a rigorous safety assessment process that takes into account the severity and the frequency of consequences on the application Operational Design Domains (ODD). In the following part, it is assumed that the simulation tools are AI based. Hence, their development cycle must follow the adapted software engineering cycle. All the components including AI do follow the "W" approach rather than the classical "V" approach (see. Figure 7). The following approach is issued from development in aeronautics concerning AI systems qualification and certification of systems including AI-based software modules. Following figure is issued from EASA concept Paper: First usable guidance for Level 1 machine learning applications ('assistance to human') suggests separating the AI based subsystem from the classical components. The classical components go through the normal V&V process while the AI based element follows the W shaped cycle (steps in blue). Note: EASA designate European Union Aviation Safety Agency that is a certification organism.

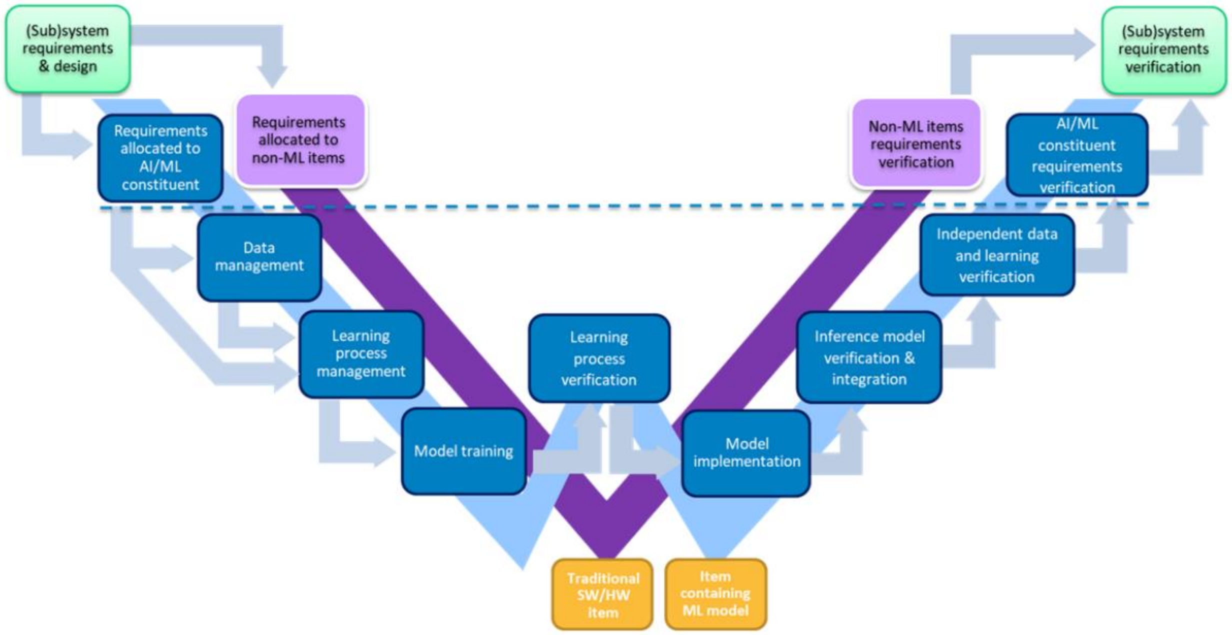


Figure 7: Global view of learning assurance W-shaped process, non-AI/ML content V-cycle process.

This process remains iterative as shown in the figure 8 below. Learning process verification affects requirements, data management, learning process management and model training for the main application that is the automated or autonomous road transport system in the context of PRISSMA (see figure 8).

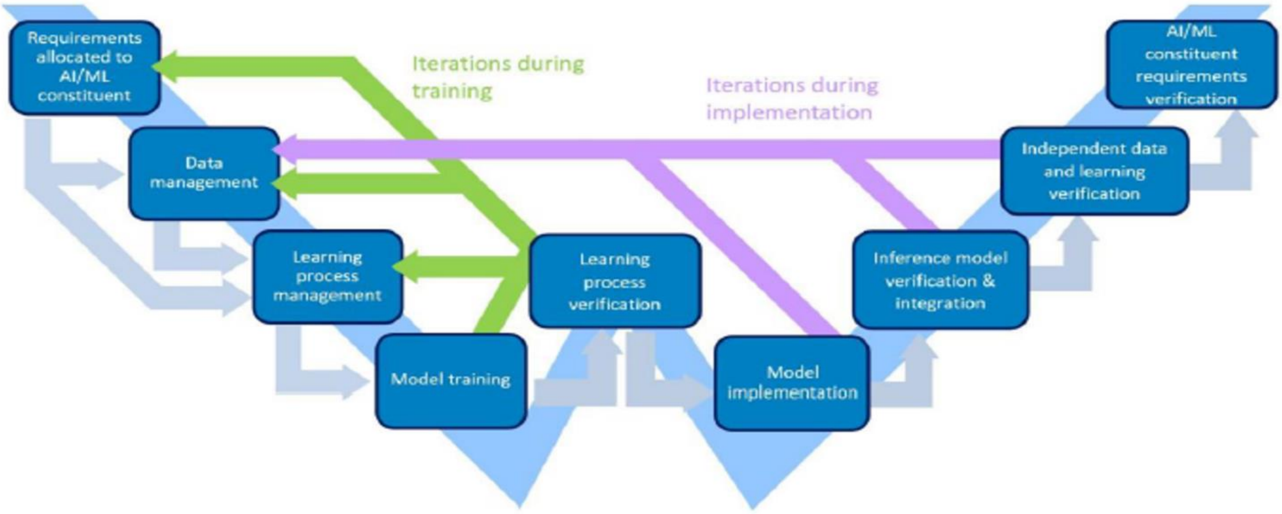


Figure 8: Iterative nature of the learning assurance process.

Tackling specifically the simulation tool, it may be located within the Learning process verification box. However, in order to perform this specific task it has to fulfil its own W cycle too and simulate accurately the main application behavior.

In the Aeronautics domain, European Union Aviation Safety Agency (EASA) published MLEAP deliverable Phase 2 - Interim Public Report on Machine Learning Application Approval (MLEAP). Currently it is the most recent work dealing with validation and qualification of machine learning for transportation domains whatever aeronautics, automotive, railway, etc. The focuses are:

- Data completeness and representativity, with handling of the simulator

• Model development, through the handling of the generalization properties (related to the Learning Process Management and Independent data and learning verification steps in the W cycle)

• Model evaluation, in particular in terms of robustness and stability (related to the Learning Process verification, Interference model verification and integration and Independent data and learning verification steps in the W cycle)

Following figure 9 displays iterative process involved in AI application iterative improvement process: If wrong classification process have been operated during the development of this AI application, following process may bring a mitigation for this classification default;

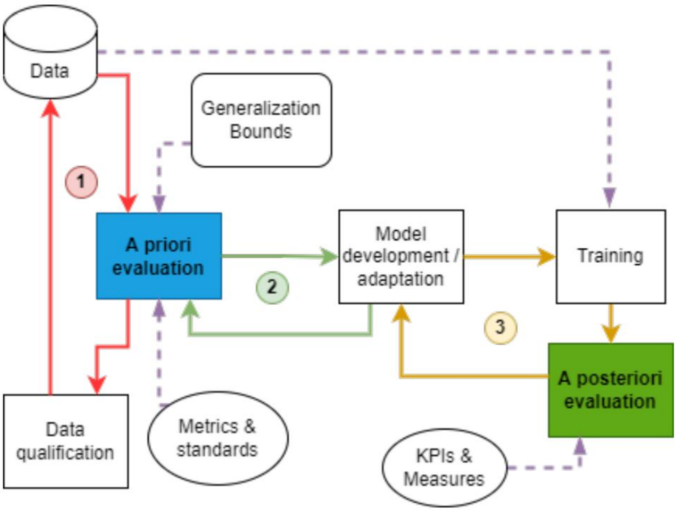


Figure 9: Model verification and updating process. 1: A priori evaluation; 2: model development/adaptation; 3: training; 4: implementation and embedding

In case of ARTS the simulation tools shall take in account the nominal behavior but also degraded behavior of AI based systems like sensors. Worst cases, edge’s cases, corner cases, rare cases must be “pushed to the limit” to be modelled correctly.

The simulator for ARTS AI based application allows measuring the quality of training of these two applications. The evaluation for ARTS AI based application needs to guarantee that the application AI modules are stable and robust. PRISSMA reports present two methods based on performance measures based on empirical data and validation of explicit properties to verify. PRISSMA report provides access to different methods of evaluation: desired generalizing ability of ML/DL (Machine Learning / Deep Learning) through the Random labelling, data corruption and finally through evaluation of ML approaches :

- Random labelling consists of tagging the data with the wrong labels for example labelling a dog picture as an air plane while keeping a set of data that is correctly labelled. Then to run the learning algorithm in parallel to compare the results of a model trained with natural data vs randomized data. “The hypothesis is that if it turns out to be the same in both cases, it cannot even distinguish learning from natural data (where generalization is possible) from learning on randomized data (where no generalization is possible).”
- Data corruption can also be used to compare the behavior of a model with natural data vs partially corrupted data, shuffled pixel data, random pixel data and compare learning process and performance evolution: this can be useful to test robustness of algorithms but also to analyze cyber security issues.
- Data integrity and bias: Since data is collected by humans, it may reflect a bias that can remain undetected if the focus of testing is solely on performance. Behavioral tests can help detect the bias. ML can present failure modes due to performance bias failures, robustness failures or model input/output failures. These failures should be taken in account and to ensure correct evaluation.
- Machine Learning: different characteristics can be evaluated separately. ML correctness, robustness and fairness can be evaluated using tools such as DeepXplore and Themis. The core of ML/DL module can be tested on tools like Tensor Flow and Scikit-learn. Finally, the workflow and application scenarios can be evaluated separately. ML can be also tested through adversarial attacks where the aim is to confuse the model to train and assess if it is robust against such situations.

Finally, validation process of a critical Automated Driving System is based on three main steps :

• Test Run Preparation phase: This phase includes the definition of the objectives, the specification and the selection of the scenarios to be tested from the Database and the assignment of each test case to a specific test facility or defined simulation environment.

• Test run execution: Based on the specificity of each testing toolchain, execute the test per description on the concrete scenario (implemented test cases) given by the step before.

• Test Results Compilation: This task consists in extracting the results from the test execution and applying post-processing (like the creation of reference) and metrics and KPI to analyze the results. A final stage consists in compiling all test results into a unique document that is then distributed to required stakeholders (homologation body, auditing internal body, consumer testing if applicable etc...). A separate process here consists in assessing the adequacy of the testing method (i.e. Simulation, Open Road and Proving Ground) to the purpose of the test itself.

V. LIFE CYCLE MANAGEMENT

Performing Life Cycle Management of a system integration IA based modules is a challenge, considering specific properties of IA technologies. This is obviously the case for automatic shuttles, and most important feature of Life Cycle Management requirements has to cover relevant feedback and corrective action when an unacceptable operational situation has been experienced in the operational cycle of the system.

When this is the case, one has first to identify the single cause or multiple causes of this unacceptable behavior, and then to setup proper corrections : when operating an ARTS, big data is collected in real time around the automated busses and in the environment where it operated, and some criteria may highlight accident events or “near miss” situations where almost accidental situation has been reached; this is an automated process where a recurrent surveillance service of a remote maintenance center is involved and where different kind of expertise are required to qualify and understand context and origin factors of this behavior; the team is composed of design and maintenance engineers, data scientists as well as AI experts.

A. Diagnosis of AI based software

Different kinds of corrections have to be envisaged, depending on the nature of the causes diagnosed:

- If one cause is a failure mode of a hardware component or module, a proper corrective maintenance task can be enforced, in accordance and compliancy with the maintenance policy of the system: this failure mode refers to an identified Line Replaceable Unit which can be exchanged on site, or on another maintenance level, regarding the maintenance concept
- If one of the possible causes is a non AI software error, a cause analysis has to be applied to the software: it can be a specification error, or a coding error, and in both cases update of the software may be in question, as well as to find out why in the development process this error has been let unknown
- If one of the possible causes is an AI based software error, a cause analysis has to be applied to the software; after this cause analysis, correction(s) of the software must be proposed, and impact analysis of this (these) correction(s) have to be applied; besides a diagnosis has to be applied to the development process and framework which has let this error unknown.

This task may be a tricky task taking into account inherent properties of AI technologies and scientific domain.

These contributions should be qualified in the real world, as trustworthiness of models supporting simulations remains currently partial: replicability and repeatability of the unacceptable situation to which AI component has contributed would be decisive about the fact to qualify the irrelevant behavior and internal diagnosis.

B. Correction of AI based software

To find proper correction of AI based software able to reestablish convenient and acceptable behavior of the whole system in the use case addressed originally, one has to conduct a deep survey to identify part of the software to correct and precise elements to change, update or remove.

For example, if AI software is based on Neural Networks, one has to find out what layer (s) of the networks to modify, and what value of weights to modify and readjust to obtain correction of the global behavior of the top-level system in addressed use case. Preferentially AI developers familiar with Neural Network and if possible initial designer / developer of this brick should be involved in this correction. Contrary to non-AI diagnosis tools, there is not a large panel of relevant methodologies and tools to diagnose AI bricks and systems.

The learning models of the AI bricks of the autonomous driving system require diagnosis when failure cases are encountered during the operation of the autonomous vehicle. These learning models have to follow an elaborate testing and certification process to avoid accidents. This process is time consuming and can take up to 6 months to 1 year for each update. However, we expect that customers will always encounter failures that are underrepresented in the training data and not taken into account in the test data or due to missing features in the learning model.

Thus, an important issue facing autonomous vehicle operators is the maintenance of the autonomous driving system software of AI bricks between major software updates, in order to fix the driving behavior of the autonomous module on the encountered failure cases or to add the requested missing functionalities of the model without the need to validate the whole system from the beginning. We believe that the diagnosis and maintainability of learning models are important challenges for the success of autonomous shuttles. The maintainability of autonomous driving systems must correct the failures of the learning models without changing the driving behavior over all the kilometers that have been successfully driven before.

C. Non regression demonstration

Corrective action on a faulty software has to remove a faulty behavior, but at the same time, one has to be sure that it does not produce additional misbehavior on other use cases, which were not failing before. This a tricky issue which is not yet wholly covered by the state of the art but in which alternative solutions are proposed. Basic ones could refer to Impact Analysis operated simply by simulation and impact assessment on different families of use cases tested. Others more sophisticated refer to Topological Data Analysis, Abstract Interpretation or Adversarial Attacks, with many variations in the way they can be applied.

VI. CONCLUSION

In conclusion, PRISSMA project has enlarged spectrum of validation methodologies and testing scenarios to STRA systems. The evaluation protocol presented herein serves as a structured framework designed to validate the simulation framework, providing an approach to integrating simulation into a homologation process that is a real breaking point from conventional procedures. The inclusion of use cases through four Proof of Concepts demonstrates the practical application of the evaluation protocol but also highlights the adaptability and versatility of the simulation protocol across various scenarios. Each POC exemplifies the protocol's effectiveness in assessing the AI's performance, ensuring its robustness, safety, and reliability under diverse applications and uses. Furthermore, the proposal of validation conducted for each component of the simulation framework underscores the rigorous testing and validation procedures employed. The validation outcomes serve as a proof to the framework's capability to accurately simulate real-world scenarios, replicating complexities and nuances encountered during open road testing, XIL experiments and track testing in the other WP. Moving forward, this outcome serves as a springboard for continued refinement, optimization, and expansion of the evaluation protocol and simulation framework. The collaborative efforts involved in its development has reflected the commitment of all partners to ensure the safety, efficiency, and advancement of autonomous vehicles, fostering innovation while upholding stringent standards of quality and compliance.

However, one has to recognize that following difficulties have been faced in this project:

- Combinatory management and coverage proof of the Operational Domain (OD) through scenario approach
- Difficulty to assure non regression or continuous improvement in case of correction of AI applications
- Capability of generating representative critical scenarios and to manage “black swans”
- Identification of validation thresholds (which value is sufficient, on which stopping criteria?)
- Possible assurance process on “black boxes” of AI algorithms and cyber security fences
- Automatic detection of a possible deviation from the Operational Design Domain
- Lack of Maturity of available technologies during the project
- Time very limited for trials in real environment, post processing of situations met, with a very big volume and diversity
- Difficult statistical interpretation of tests in controlled environment and appearing a few times

However, specific following outcomes have been derived from this project:

- First of all production of a validation process framework of transportation system integrating embedded AI applications and taking into account specificity of AI
- Many KPI have been specified, described and put in practice in the 4 Proof Of Concept, taking specificity of AI
- A whole framework of requirements has been allocated at different levels of the system (sensors, sub systems, automated shuttle), but also applying to the different tools of the simulation platform
- Criteria on Data Validation and Management especially concerning Ground truth have been formulated
- Automatic Record of events during system operation, detection, feedback of unexpected situations, criteria for Data Management process have been specified

Futures perspectives obviously have to be highlighted:

- Work about a more deeper exploration of Operational Design Domain by better managing degraded situations and interactions of automated shuttle under consideration with other actors of traffic context,
- Work about better update of environment through configuration management of digital models and digital twins for example urban environmental settings,
- Continue to invest in real time simulation workbenches with high performance, integrating efficient tools from the different point of view of simulation requirements and capabilities, including among others coherent time sampling, interoperability, physical realism, optical / electromagnetic phenomena simulation capability,
- Foster mutualization of best practices in terms of technics and methodologies for producing these models and optimize their demonstration capability in coherency with controlled test campaigns and test as well in real environment,
- Encourage development of tools for easy production of digital twins and building of framework to share digital twins of urban and peri-urban geographical locations
- Last but not least contribute to integrate and introduce results of PRISSMA in regulation being currently written through contribution to national and international Working groups

References

[1] F. F. Philip Koopman, How Many Operational Design Domains, Objects, and Events?, Carnegie Mellon University, 2019.

[2] PRISSMA, Deliverable 8.11, OPERATIONAL DESIGN DOMAIN ,2023

[3] ISO, ISO 34503, Road Vehicles - Test scenarios for automated driving systems - Specification for operational design domain, 2023

[4] A. David, "The levels of automation – The Operational Design Domain", https://www.linkedin.com/pulse/levels-automationoperational-design-domain-part-2-n-david-apelt-1d/, 2019

[5] EU, 32022R1426, Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles, 2022.

[6] SAE, SAE J 3016, surface vehicle recommended practice, 2021.

[7] A. David" The levels of automation – Object and Event Detection and Response (OEDR) and the Game of Thrones", http://www.linkedin.com/pulse/levels-automation-object-event- [L6.5] WP6 INTEGRATION OF IA SPECIFICATIONS INTO DESIGN STANDARDS 58 detection-response-oedr-david-apelt/, 2019.

[8] Damien Pichereau, Le déploiement européen du véhicule autonome, 2021.

[9] UN/ECE, UN regulation 131, Uniform provisions concerning the approval of motor vehicles with regard to the Advanced Emergency Braking Systems (AEBS), 2014

[10] UN/ECE, UN regulation 152, Uniform provisions concerning the approval of motor vehicles with regard to the Advanced Emergency Braking System (AEBS) for M1 and N1 vehicles, 2020

[11] ISO, ISO16290, Space systems - Definition of the Technology Readiness Levels (TRLs) and their criteria of assessment, 2013

[12] United Nations, UN Regulation No. 160, Event Data Recorder, 2021.

[13] United Nations, UN Regulation No. 157, Automated Lane Keeping Systems (ALKS), 2021.

[14] UN/ECE, UN regulation 94, Uniform provisions concerning the approval of vehicles with regard to the protection of the occupants in the event of a frontal collision, 2012

[15] UN/ECE, Addendum 94 - UN regulation 95, Uniform provisions concerning the approval of vehicles with regard to the protection of the occupants in the event of a lateral collision, 2021

[16] UN/ECE, UN regulation 137, Uniform provisions concerning the approval of passenger cars in the event of a frontal collision with focus on the restraint system, 2020

[17] UN/ECE, Addendum 9, UN Regulation 10, Uniform provisions concerning the approval of vehicles with regard to electromagnetic compatibility, 2014

[18] UN/ECE, UN Regulation 121, Uniform provisions concerning the approval of vehicles with regard to the location and identification of hand controls, tell-tales and indicators, 2016

[19] PRISSMA, deliverable 1.5, TESTS AND AUDIT REQUIREMENTS, 2023

[20] PRISSMA, deliverable 6.3, STATE OF THE ART, 2024

[21] ISO, ISO19206-5, Test devices for target vehicles, vulnerable road users and other objects, for assessment of active safety functions - Part 5: Requirements for Powered Two-Wheeler targets, 2018 [22] PRISSMA, Deliverable 4.3, REAL CONDITION TESTS METHODS: INFRASTRUCTURE ANALYSES, PATHWAY SELECTION CRITERIA, 2023 [23] PRISSMA, Deliverable 2.7, METHODOLOGY, PROCEDURES AND PROTOCOLS FOR EVALUATION OF APPLICATIONS AND VIRTUAL TEST FACILITIES

[22] ISO, ISO 16505, Road vehicles - Ergonomic and performance aspects of [L6.5] WP6 INTEGRATION OF IA SPECIFICATIONS INTO DESIGN STANDARDS 59 “Camera Monitor Systems - Requirements and test procedures”