ReliablE in-Vehicle pErception and decisioN-making in complex environmenTal conditionS

Grant Agreement Number: 101069614

D6.1 Experimental Procedures and Evaluation Methods

Document Identification			
Status	Final	Due Date	31/05/2024
Version	1.0	Submission Date	15/07/2024
Related WP	WP6	Document Reference	D6.1
Related Deliverable(s)	N/A	Dissemination Level	PU
Lead Participant	ICCS	Document Type:	Other
Contributors	All WP6 partners	Lead Authors	Bill Roungas (ICCS)
		Reviewers	Elena Krikigianni (SEAB)
			Fabio Tango (CRF)

Funded by the European Union

This project has received funding under grant agreement No 101069614. It is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Document Information

Author(s)		
First Name	**Last Name**	**Partner**
Leonardo	Gonzalez	TECN
Dariu	Gavrila	TUD
Anastasia	Bolovinou	ICCS
Markos	Antonopoulos	ICCS
Manos	Gkigkilinis	ICCS
Kirsty	Aquilina	APTIV-DE
Massimiliano	Lenardi	HIT-FR
Alireza	Ahrabian	HIT-UK
Andras	Palffy	PERCIV

Document History			
Version	**Date**	**Modified by**	**Modification reason**
0.1	05/03/2024	ICCS	Empty document with section structure
0.2	14/04/2024	All	First input from partners
0.3	30/05/2024	All	Revised input from partners based also on the comments from the mid-term review
0.3	20/06/2024	ICCS	Merged contributions in a single document. First version ready for Review.
0.4	27/06/2024	SEAB	Completed deliverable review from SEAB
0.5	30/06/2024	CRF	Completed deliverable review from CRF
0.6	12/07/2024	ICCS	Merged reviews and including contributions from partners.
1.0	15/07/2024	ICCS	Final review, Plagiarism check & Submission

Quality Control		
Role	**Who (Partner short name)**	**Approval Date**
Deliverable leader	Bill Roungas (ICCS)	12/07/2024
Quality manager	Panagiotis Lytrivis (ICCS)	12/07/2024
Project Coordinator	Angelos Amditis (ICCS)	15/07/2024

Executive Summary

Work Package 6 (WP6) focuses on evaluating the EVENTS experiments, that were prepared and implemented in WP5. To accomplish this, dedicated methods and algorithms will be developed for the effective analysis and evaluation methodology of the use cases (as defined in WP2). Within WP6, the relevant scenarios and data will be collected, which will allow to generate the test-cases for the overall evaluation of the EVENTS experiments from the technical perspective, with regards to their efficiency and their capability to extend the related ODD.

The main focus of this document (Deliverable D6.1) is to describe in detail the testing plan and the evaluation procedures for each system developed in the project, taking into account SOTIF verification procedures. This is the outcome of T6.1 dealing with the evaluation methodology and the respective scope.

The final results of the evaluation will be reported in deliverable D6.2 "*Technical evaluation results*", due on M36.

Table of Contents

List of Tables

List of Figures

Abbreviations & Acronyms

Abbreviation / acronym	Description
ACC	Adaptive Cruise Control
AD(F)	Autonomous Driving (Function)
AI	Artificial Intelligence
AL	Alert Limit
AP	Average Precision
AV	Automated Vehicle
BP	Behavioural Planner
CA	Consortium Agreement
CAM	Cooperative Awareness Message
CAV	Connected Automated Vehicle
CCAV	Centre for Connected & Autonomous Vehicles
CNN	Convolutional Neural Network
CPM	Collective Perception Messages
CRAT-pred	Crystal Attention Prediction
CTRA	Constant Turn Rate and Acceleration
DDT	Dynamic Driving Task
DENM	Decentralized Environmental Notification Message
DM	Decision Making
DNN	Deep Neural Networks
DTC	Distance to Collision
EB	Elastic Band
EC	European Commission
EXPs	Experiments
FIS	Fuzzy Inference System
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FoV	Field of View

Abbreviation / acronym	Description
FTP	Fail-degraded Trajectory Planning
GA	Grant Agreement
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HD	High Definition
IMU	Inertial Measurement Unit
IR	Integrity Risk
IoU	Intersection over Union
ISO	International Organization for Standardization
I/O	Input(s) / Output(s)
Km/h	Kilometres per hour
KPI	Key Performance Indicator
LiDAR	Light Detection and Ranging
LSTM	Long Short-Term Memory
M x	Month x
mAP	Minimum Average Precision
MDP	Markov Decision Process
MiL	Machine-in-the-Loop
minADE(k)	Minimum Average Displacement Error (over top k scored trajectories)
minFDE	Minimum Final Displacement Error
ML	Machine Learning
MNIS	Multi Target Normalized Innovation Squared
MOP	Moving Object Prediction
MOT	Multi-Object Tracking
MPC	Model Predictive Control
MR	Miss Rate
MRM	Minimum Risk Manoeuvre
NIS	Normalized Innovation Squared
ODD	Operational Design Domain
PE	Position Error

Abbreviation / acronym	Description
PIs	Performance Indicators
PL	Protection Level
PP	Perception Platform
RADAR	RAdio Detecting And Ranging
RAIM	Receiver-Autonomous Integrity Monitoring
REQs	Requirements
RGB	Red, Green, Blue
RL	Reinforcement Learning
RQs	Research Questions
SA	Self-Assessment
SAE	Society of Automotive Engineers
SciL	Scenario-in-the-Loop
SMD	Safety-mode Decision
SiL	Software-in-the-Loop
SoA/SoTA	State of the art
SPaT message	Signal Phase and Timing message
SPECs	Specifications
TN	True Negative
TOR	Take Over Request
TP	Trajectory Planner
TPR	True Positive Rate
TSs	Target Scenarios
TTB	Time to Brake
TTC	Time to Collision
TTS	Time to Steer
UCs	Use Cases
V2V	Vehicle to Vehicle
ViL	Vehicle-in-the-Loop
VRU	Vulnerable Road User
WP	Work Package

1. Introduction

1.1Project aim

Driving is a challenging task. In our everyday life as drivers, we are facing unexpected situations we need to handle in a safe and efficient way. The same is valid for Connected and Automated Vehicles (CAVs), which also need to handle these situations, to a certain extent, depending on their automation level. The higher the automation level is, the higher the expectations for the system to cope with these situations are.

Today, CAVs are facing several challenges (e.g., perception in complex urban environments, Vulnerable Road Users (VRUs) detection, perception in adverse weather and low visibility conditions) that should be overcome to be able to drive through these events in a safe and reliable way.

Within our scope, and, to cover a wide area of scenarios, these kinds of events are clustered under three main use cases: a) Interaction with VRUs, b) Non-Standard and Unstructured Road Conditions and c) Low Visibility and Adverse Weather Conditions.

Our vision in EVENTS is to create a robust and self-resilient perception and decision-making system for AVs to manage different kind of "events" on the horizon. These events are due tothe dynamic changing road environment (VRUs, obstacles) and/or due to imperfect data (e.g., sensor noise and communication failures) and/or due to challenging ODD conditions (e.g. entering a road construction zone). The EVENTS AV should continue and operate safely within its ODD and near ODD limits due to its advanced perception and decision making as well as due to its self-assesment features. When the EVENTS system cannot handle the situation, an improved minimum risk manoeuvre should be put in place.

1.2 Deliverable scope and content of the Document

In the EVENTS project, WP6 is responsible primarily with the evaluation of the experiments, as those are defined in EVENTS Deliverable D2.1: User and system requirements for selected use cases [1], and to a lesser degree with the cost efficiency of sensor suites. The main focus of this document is to describe, in detail, the testing plan and the evaluation metrics for each experiment, by also taking into account functional safety concept as this has been identified and reported on system level, in EVENTS Deliverable D2.3: Vehicle System Hazard Analysis and Risk Assessment [3]. The content of this document is the outcome of T6.1, which deals with the evaluation methodology and the respective scope, and which will provide input for the subsequent tasks T6.2 & T6.3.

T6.2 is concerned with the data collection from the execution of the experiments and their preparation for the analysis. T6.3 is concerned with the technical evaluation of each experiment, using the methodology proposed in task T6.1 and the data collected in task T6.2. The results dealing with the performance of the different experiments, i.e., the outcome of tasks T6.2 & T6.3, will be documented in deliverable D6.2 (Technical evaluation results).

The document has the following structure:

- In Chapter 2, the methodology, including the high-level research questions and the testing and evaluation methods, is described.
- In Chapters 3 to 10, the individual evaluation setup of EXPs 1 to 8 respectively, are presented.
- Finally, in Chapter 11, the document is concluded and final remarks are reported.

2. Methodology

The evaluation objectives consider a) the individual modules within the perception & decision-making layers of each EVENTS vehicle system, b) each architectural layer separately, i.e., perception and self-assement layer and decision making and control layer as well as the integrated vehicle system as a whole. Therefore, different key performance indicators are proposed for each subsystem/system.

Based on the use cases and distinct experiments defined in WP2 [1] and the underlying architecture of each vehicle sub-system system in EVENTS described also in WP2 [2], D6.1 establishes the objectives (RQs) and the methods for the project evaluation, including the candidate testing environments and, when available, an indication of the most important functional scenarios [21] to be considered.

With respect to functional safety evaluation, the methodological process of the evaluation follows a traditional V-model, as shown in Figure 1, enhanced with Safety Of The Intended Function (SOTIF) aspects (as defined in the ISO 21448 [23]), as shown in Figure 2. This means that to a certain extent permitted by the project's timeplan, the evaluation runs in an iterative mode where results from simulations or test tracks are expected to feed back the modules' development work in WP3 and WP4. As shown in Figure 1, which depicts a simplified illustration of the V-Model process in ISO 26262, both component-level and system-level tests are considered by the ISO standard.

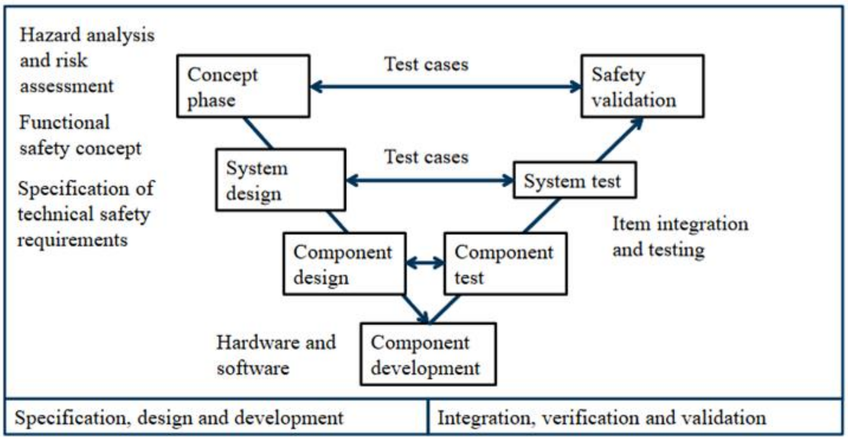

Figure 1: Traditional V-model [18]

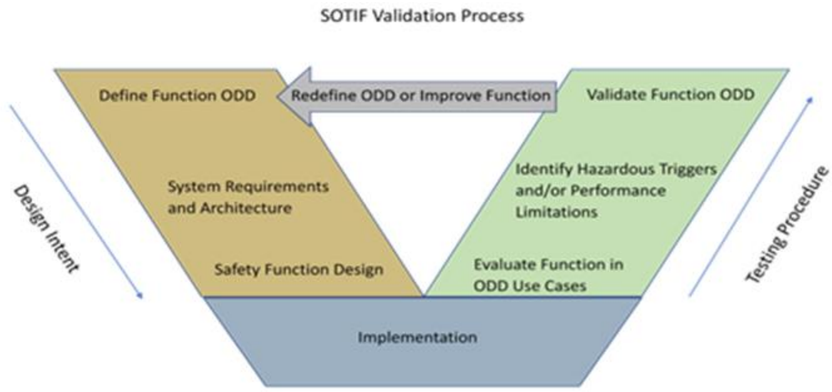

Figure 2: V-model enhanced with SOTIF aspects [20]

Based on the V-model enhanced with SOTIF aspects, two approaches for the system evaluation are adopted:

1. **Bottom-up**. In the Bottom-up approach, each EVENTS AD system under test is compared against an AD Baseline system or manual driving.
2. **Top-down**. In the Top-down approach, the EVENTS AD system is assessed against some acceptance criteria derived from the SOTIF analysis, in T5.3.

2.1 Input

As an input of this work and in order to define the EVENTS set of KPIs and candidate test scenarios and/or test cases, three different sources, two from within the EVENTS project and one from an external project, have been considered.

With regards to the KPIs that preceded T6.1:

- Sub-system KPIs that were proposed from the system requirements per experiment and can be found in deliverable D2.1 [1].
- Module specific KPIs that were proposed by WP3 & WP4 per experiment based on the SoA (Bottom-Up approach) and can be found in deliverables D3.1 [4] & D4.1 [5] respectively.
- Inputs from an external project: Hi-Drive [6] system-level KPIs.

For those experiments that are defined as end-to-end (the perception layer is considered as input to the decision-making layer), the decision-making KPIs are also the end-to-end/system KPIs.

With regards to the test cases (scenarios + acceptance criteria) in EVENTS, which is work that will follow T6.1, task T5.3 is considered relevant. Outcomes of T5.3 (see also Section 2.4) will be integrated in the subsequent tasks of this WP, in T6.2 (data campaigns) and in T6.3 (tests execution and reporting), in order to help all evaluation teams in defining their set of scenarios and if applicable, their test cases when end-to-end systems are tested. More specifically, outcomes to be utilized are:

- System and sub-systems safety acceptance criteria per experiment and per vehicle (Top-Down approach based on the EVENTS master architecture).
- Set of triggering conditions that lead to hazardous scenarios per experiment or a set of challenging functional scenarios per experiment (this will be based on T5.3 HARA analysis based on each experiments' architecture; this improves the reference work done in D2.3 based on the EVENTS generic architecture, as it considers the specificities of each experiment).

2.2 Research Questions

Each experiment intends to answer the following **research questions** (RQs):

1. To what extent does the proposed solution improve the perception of the AD compared to the baseline/SoA?
2. To what extent does the proposed solution improve the self-assessment of the AD compared to the baseline/SoA?
3. To what extent does the proposed solution improve the decision-making of the AD compared to the baseline/SoA?
4. If applicable, does the experiment improve robustness by fulfilling the minimum safety requirements on a system-level, defined as acceptance criteria by the SOTIF analysis?

2.3 Evaluation Areas & Testing Environments

The evaluation process will take place in three different areas:

1. **Nominal ODD**: PIs for assessing the perception and decision-making sub-system and the end-to-end system performance in nominal ODD conditions (input 1 in Section 2.1). Black-box or a white-box testing can be considered.
2. Assessing the perception, decision-making and end-to-end system reliability in **boundary ODD conditions**, which would include challenging scenarios (e.g., darting out pedestrian), challenging ODD conditions (e.g., rain) or long-tail (rare) scenarios, the latter of which assumes generated scenarios in simulation.
3. Argue about the AD system safety and robustness by applying safety acceptance criteria for selected end-to-end system PIs, provided by task T5.3, in known unsafe situations (e.g., darting out pedestrian), exploring, to the extent possible, situations on ODD boundary (i.e., unknown unsafe) and verify through small-scale real-world testing (handled by WP6).

The evaluation of the EVENTS system and sub-systems will occur in three distinct environments:

In-Lab and/or in simulation

1. Real-world data replay in lab environment (more suitable for perception modules).
2. Model-in-the-Loop (MiL), in which the SuT would run in a simulation environment.

Real-world

1. Operation in the field, in which the systems would run in a demo vehicle either in a test-track or, if the vehicle is permitted to do so, in a public road (suitable for partners with end-to-end AV systems).
2. Operation in the field, test-track and/or open road, with the system under test in shadow-mode, in which the SuT would run in a demo vehicle, still no AD function is controlling the vehicle. (suitable for partners with no AD demo vehicle).

Mixed virtual-real world (hybrid)

Scenario-in-the-Loop (SciL) and Vehicle-in-the-Loop (ViL), in which said systems would run in a hybrid environment interconnecting a demo vehicle and a simulation environment and where real-time exchange of information between the two is required.

All EVENTS experiments will evaluate their perception system and four of them will evaluate their decision-making and end-to-end system. The data used for the evaluation will be gathered from three sources: publicly available datasets, data recorded from simulations or hybrid setups and data recording from the EVENTS demo vehicles in test tracks or open roads. A summary of the evaluation plan per experiment and partner can be found in Chapter 11.

2.4 Safety Of The Intended Functionality (SOTIF)

Task 5.3 aims to systematically identify potential hazards related to the SOTIF. A SOTIF hazard can arise from both internal and external factors when there is no fault in the system. Internal factors to the system include performance limitations, insufficient situational awareness or foreseeable misuse by the user (the latter is not tackled in EVENTS). External factors include triggering conditions such as challenging weather conditions.

Based on ISO 21448 [23], SOTIF activities are implemented in three phases, the two of the three are also addressed here in EVENTS evaluation process as follows:

1. Design Phase: Identify SOTIF hazards (critical scenarios) and then derive performance requirements on perception layer, decision making layer and end-to-end system. This work will be performed in EVENTS T5.3.

2. Verification phase: This includes technical reviews to verify all derived performance requirements are met, the known (unsafe) scenarios are covered, i.e., pass-fail criteria are met and the system behaves as expected. This is done by considering test cases with a high coverage of relevant hazardous scenarios in small scale test track experiments and injection of potential triggering events via in the loop testing (e.g., SiL/MiL) of selected SOTIF relevant scenarios. This phase will be carried out in WP6, T6.2 and T6.3 per experiment, taking into account the pass-fail criteria proposed in T5.3. In this latter task, pass-fail criteria definition will consider the proposed subsystems' PIs from this deliverable and if needed it will propose new PIs.

3. Validation phase: The validation phase requires a large amount of data collection mileage so it is not applicable to EVENTS project but is required for automotive industry production systems.

Since T6.2 work runs in parallel with WP3 and WP4 for a few months, it is possible that the results from T6.2 with respect to SOTIF related assessment, will feed the refinement of the WP3 and WP4 modules in an iterative process. During evaluation, a valid statistical claim requires that the system under test be stable, and thus any iteration on the function design will require repeating any verification tests that may have been affected by the change.

2.5Template used to gather input from all teams

In order to define a thorough evaluation plan, each module/AD subsystem developer has described their module/subsystem evaluation plan. In order to have consistency between the partners' input, a template was used to gather all the information relevant to the evaluation plan. The template required from partners the following information:

- Target evaluation objective (performance, reliability, safety),
- Evaluation scenarios,
- Evaluation PIs,
- Datasets (train, test), and

- The baseline showing against which system they are comparing.

The abovementioned description is presented in Sections 3 to 10, starting from the results of WP2, reported in deliverable D2.1 “User and system requirements for selected use cases” [1].

3. EXP1

EXP1 is about safe, comfortable and time-efficient automated driving in complex urban environment while interacting with VRUs (e.g. pedestrians, cyclists). The environment perception, road user motion prediction, motion planning and vehicle control will be demonstrated in a single integrated system on-board TUD's own vehicle prototype. The experiment consists of the ego-vehicle driving on a two-lane road (i.e., one lane on each side) whereas several VRUs might (or might not) move into the vehicle's path (e.g., crossing, walk longitudinally, swerve), possibly from behind occlusions (e.g., parked vehicles). The question is whether to decelerate, accelerate or steer away. The experiment is led by TU Delft (TUD).

In EXP1, TUD's evaluation plan includes work on their perception layer, on their decision-making layer with the perception layer abstracted and on their decision-making layer with the perception layer present (end-to-end experiment).

3.1 Perception Layer

With regards to the perception layer, TUD's evaluation objective is to assess the perception performance in real-world driving scenarios, based on two datasets recorded in the USA and Europe, the nuScenes and Zenseact.

For measuring the perception performance, TUD will use two versions of the average precision (AP) metric, i.e., the KITTI [10] and nuScenes [11] definitions. Average precision measures the area under the precision-recall curve, and it is the main metric used for evaluating 2D and 3D object detection methods. The score is a value between 0 and 1, where 1 corresponds to a perfect detector (recall of 100% with a precision of also 100%). The main difference between the two versions (KITTI and nuScenes) is that the KITTI version is Intersection over Union-based, while the nuScenes version is distance-based. This means that the KITTI version measures how well the predicted bounding boxes overlap with the ground truth bounding box, while the nuScenes version measures how close the predicted bounding boxes are with respect to the ground truth bounding boxes.

The results will be compared against two baselines 1. the standard detection model of Autoware, which is trained on nuScenes, and 2. the model trained on the recent Zenseact dataset [12]. Both baselines are obtained using full supervised training, i.e., using labelled data to train the detector. The nuScenes dataset has a LiDAR with only 32-vertical beams, while the Zenseact dataset has a LiDAR with 128-vertical beams. TUD's vehicle also has a LiDAR with 128-vertical beams. The training data for baseline 1 comes from Boston/Singapore, and the data for baseline 2 was recorded in Europe. As a result, for both baselines, there is a domain shift (data distribution of training data is different from the test environment), and there are (small) differences in the sensor characteristics. However, that is the standard for relying on detectors trained with supervised learning.

3.2 Decision-making Layer

TUD will evaluate the decision-making layer with the perception layer both abstracted and present.

3.2.1 Perception Layer Abstracted

With the perception layer abstracted, TUD's evaluation objective is to assess the motion planner with decision-making reaching the goal as fast as possible while keeping sufficient distance from obstacles. The motion planner will be validated in the Autoware planning simulator that is extended with virtual pedestrians. The evaluation scenarios will include the ego-vehicle driving on a two-lane road involving VRUs entering its path; pedestrians will be spawned in random start locations with random directions on a straight section of the road.

KPIs will be collected over a large number of simulations to compare their statistics. The performance of motion planner with decision-making will be measured by the PIs shown in Table 1.

Table 1: Evaluation PIs of the decision-making layer of EXP1

Evaluation PIs	**Description**
Minimum Distance to Obstacles [m]	It consists of the minimum distance between the vehicle and any of the obstacles (each modelled as a set of the discs) minus their respective radius. When zero, it indicates that a collision occurred.
Number of collisions	It validates whether any collisions occur in the simulation.
Task Duration [s]	It is the time taken for the vehicle to reach its goal. Faster motion planning leads to a lower task duration.
Computation time [ms]	It should remain below the allocated planning time (typically 50-100ms).

The developed planner will be compared against the Autoware planning stack, which consists of a behavioral layer and a local planner. Both planning layers consist of several different planners that can be activated in different situations based on a behavior tree. For obstacle avoidance, a rule-based behavioral layer generates a rough trajectory [13]. An Elastic Band (EB) local planner refines the steering to evade obstacles without changing the velocity [14]. The main difference between the baseline and the proposed planner is that the latter generates and optimizes multiple high-level passing behaviors. This allows it to explore multiple locally optimal trajectories and to identify the best. In contrast, the baseline only optimizes the steering angle of a single trajectory.

3.2.2 Perception Layer Present

With the perception layer present, TUD's evaluation objective is to assess the end-to-end performance on a real track, including real-world uncertainties and false positive detections. The evaluation scenarios will include the ego-vehicle driving at low speed (10-15 km/h) on the test track with multiple VRUs. The pedestrians and cyclists would be at a safe distance from the ego-vehicle. The EuroNCAP-certified motion-synchronized dummy of vulnerable road users dummy (adult or cyclist) would be crossing the vehicle path with a high collision risk.

Since the perception module affects the motion planner's performance, the KPIs would be the same for the decision-making layer with the perception layer abstracted and present. The default Autoware perception module based on the CenterPoint detector [15] will be used for perception; it is trained on nuScenes and internal datasets (not publicly available). The default Autoware planning stack, including a behavioural layer and a local planner, will be used as baseline.

4. EXP2

EXP2 incorporates perception augmentation via safe integration of collective perception (CP) info, predictive planning for the control of the platooning in an urban environment (T4.1), management of the platooning behavior (T4.2) and design of a safe operational model for when an attached vehicle is in the platoon (T4.3). AV control takes advantage of augmented perception (inside and outside CAVs' FOV) offered by fusion of cooperative awareness messages (CAM) and collective perception messages (CPM) (T3.4 and T3.5) shared by other road users and platoon members. EXP2 is led by Tecnalia (TECN) and ICCS also participates.

EXP2 includes work on the perception and the decision-making layer, so the evaluation plan for these two modules is presented below.

4.1 Perception/Self-assessment Layer

The evaluation of the collective perception module designed for tasks T3.4 & T3.5 is planned in two setups, along different levels of sophistication of the core algorithm. Additionally, a study on the sensitivity of the algorithm with respect to the localization errors will be also carried out.

Specifically, the two setups consist in:

1. CARLA simulations

    - A basic scenario of 3 to 5 agents moving in a roundabout area will be devised. A special consideration is given in the generation of the environment elements in CARLA in order to create occlusions from the perspective of the sensor equipped connected vehicles (e.g., static obstacles or vegetation). Additionally, sensor noise level can be customized from within CARLA environment.

    - In at least one scenario variation one of the agents is assumed not connected and without a perception stack.

    - Each of the other agents (2-4 agents, added incrementally, thus gradually increasing complexity) is equipped with the following perception stack:

        - GPS and IMU providing measurements on the position and orientation of the vehicle.

        - Front RGB camera.

        - Lidar.

At each moment of the simulation, both the CARLA ground truth plus the perception data from each sensor will be saved, along with the transformation projecting Lidar points on the RGB camera image at each time instant. These transformations will be used to fuse Lidar and camera data, and provide an estimation of the position of each perceived object relative to the ego vehicle. Although this will not provide position and orientation of perceived vehicles per se, it will be utilized to extract the occupancy state of the grid cells within the ego vehicle's FoV. The collective perception algorithm in increasing levels of sophistication (see below) as presented so far will be used to fuse the individual perception data coming from each connected vehicle.

2. The hybrid setup of experiment 2

In this case, a json file simulating a Collective Perception Message will be available from each connected agent at each time instant of the simulation. The json file will provide the following information:

- Ego FoV angle. Each CCAV's sensor suite implies a FoV angle for the particular CCAV. For example, a single front camera implies that the CCAV's FoV angle is equal to the FoV of the camera.
- Ego state information. Information of the ego state of the CCAV, specifically:
    - Ego Vehicle Position coordinates in *x,y*
    - Ego Vehicle Speed vector *vx, vy*
    - Ego Vehicle Heading (yaw angle)
- Observed objects information. Information concerning each one of the objects perceived by the CCAV, specifically:
    - Position coordinates in *x,y*
    - Speed vector *vx, vy*
    - Heading (yaw angle)

The CPM messages will be used to calculate the FoV of each vehicle and calculate the resulting probabilistic occupancy grid.

In both the above setups, the resulting probabilistic occupancy grid will be compared with the occupancy grid constructed directly from the ground truth. This will be carried out, both from the ego-CAV perspective and the bird-eye-view scene perspective, by calculating the metrics shown in Table 2. PIs will be analyzed per scenario over time.

Table 2: Evaluation PIs of the perception/self-assessment layer of EXP2

Evaluation PIs	Description
Global IoU of occupied areas [%]	The Intersection over Union (IoU) between areas estimated as occupied (free) by the output of the collective perception module and areas actually occupied (free) according to the ground truth. The binary estimated occupancy state of each cell will be derived from the estimated probabilistic state by specified probability thresholds. IoU will provide a metric of similarity between the sets of estimated as occupied (free) and actually occupied (free) grid cells, ranging between 0 and 1; 0 indicating disjoint sets, 1 indicating equal sets.
Local IoU of occupied areas [%]	The Intersection over Union (IoU) between areas estimated as occupied (free) by the output of the ego-vehicle perception and areas actually occupied (free) according to the ground truth.
Precision/Recall metrics	The estimated probabilistic occupancy grid can be seen as a binary classification of each grid cell to "occupied" vs "free" via specified threshold probabilities. The resulting confusion matrix and precision/recall metrics will provide a more detailed evaluation of the algorithm estimations.

An even finer analysis will consider the aforementioned IoU and precision/recall metrics but specifically for the boundaries of occupied regions.

Within the above setups, three different levels of sophistication for the core collective perception algorithm will be evaluated. Specifically:

a. A simple version, taking into account only the occupied/free state of each grid cell as perceived by each agent, without taking into account the individual perception models and their respective probabilities.

b. A more sophisticated version, carried out by Bayesian fusion of the perception data of each agent, by taking into account the individual perception models and their respective probabilities. At each time step, the prior occupancy probability for each cell of the grid will be set to 0.5.

c. This version will be the same like the second version, except that at each time step, the prior occupancy probability for each cell of the grid will be set equal to the respective probability of the posterior of the previous time step. For sufficiently high sampling rates, this is expected to provide a more smooth and reliable estimation of the subsequent posterior with better handling of instantaneous "ghost (or missed) objects" noise.

The overall approach is expected to be sensitive to localization errors of the involved agents. To address this, a study on the sensitivity of the algorithm output in terms of the metrics described above with respect to the localization errors will be also carried out.

4.2 Decision-making Layer (perception layer present)

With regards to the decision-making layer of EXP2, the work involves behavioural and motion planning, and fail-safe control.

4.2.1 Behavioural and motion planning

The target evaluation objective with regards to the behavioural and motion planning is to integrate the perception inputs, to generate an appropriate plan and geometry to follow while maintaining a safety distance, and to reduce the risk situations to a minimum (based on TTC<1.5s). In more detail, the evaluation plan would:

- Provide an efficient plan for all automated participants in the scene at high level (maintain lane, change lane, enter roundabout, etc.), measured in successful completion of simulated and hybrid scenarios, and time to finish the manoeuvre.
- Improve, in terms of smoothness and safety, the trajectories for automated vehicles, which integrate possible conflicts, and adapt appropriately. These will be measured in lateral accelerations and other comfort variables.
- Target to avoid risk situations in speed planning, measured in TTC and DTC.

The architecture has a global planner that generates a path that goes on the center of the lane, based on the map available. After that, a local planner generates the reference trajectory for the controller using the state of the vehicle and the global path. The benchmarking of the trajectory would be as follows:

- Bezier curve: We use two ways of generating the control points needed for 5th order curves.
    - Equidistant points: We need 6 points to define 5th order bezier curve. Three of them are the vehicle position and two points ahead. The other three are composed by the final point of a segment of the global path (the center of the lane) and two points above it following a tangent line to the path. Distance between consecutive points will be defined in three different ways: i. Static, ii. Ponderation of the speed and of the lateral acceleration, and iii. Sum of ponderated speed and acceleration.

    - Optimization of curvature: The method used in the Autoware stack that forms an optimization problem that minimizes the square of the derivate of the curvature.

- Splines: Using a segment of the global path (the center of the lane) we are going to use 3rd, 4th and 5th order splines with weighted smoothing.

- Model based optimization: Just like in the Model based Predictive Control we use a kinematic model with lane boundary constrains to generate a curve by minimizing the position error between the trajectory and the center path.

The evaluation scenarios would be using Carla framework to test against current baseline. The evaluations PIs are shown in Table 3.

Table 3: Evaluation PIs for the behavioural and motion planning module of EXP2

Evaluation PIs	**Description**
Max lateral acceleration [m/s^2]	The max lateral acceleration of the ego vehicle (Lateral behaviour)
Average lateral acceleration [m/s^2]	The average lateral acceleration of the ego vehicle (Lateral behaviour)
Average lateral error to reference trajectory	The average lateral error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Average angular error to reference trajectory	The average angular error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Average lateral error to center line	The average lateral error to the center of the line (Lateral and Heading Error)
Average angular error to center line	The average angular error to the center of the line (Lateral and Heading Error)
Max lateral error to reference trajectory	The maximum lateral error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Max angular error to reference trajectory	The maximum angular error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Max lateral error to center line	The maximum lateral error to the center of the line (Lateral and Heading Error)
Max angular error to center line	The maximum angular error to the center of the line (Lateral and Heading Error)
Average difference of reference trajectory to center line [m]	The average difference between the reference trajectory generated by the motion planning module and the center of the line (Geometry)
Max difference of reference trajectory to center line [m]	The maximum difference between the reference trajectory generated by the motion planning module and the center of the line (Geometry)

Normalized max curvature [m]	The normalized maximum curvature of the reference trajectory generated by the motion planning module (Geometry)
Normalized average curvature [m]	The normalized average curvature of the reference trajectory generated by the motion planning module (Geometry)
Normalized max curvature difference to center line [m]	The normalized maximum curvature difference between the reference trajectory generated by the motion planning module and the center of the line (Geometry)
Normalized average curvature difference to center line [m]	The normalized average curvature difference between the trajectory generated by the motion planning module and the center of the line (Geometry)
Average computation time [ms]	The average computation time
Max computation time[ms]	The maximum computation time
Success rate [%]	The last vehicle exits the roundabout and there have not been any collisions (Decision Making)

Comparisons in decision-making and behavioural solutions are complex since its performance is heavily linked to the architecture in place. Nonetheless, EXP2 will focus on comparing the motion planning module, with the KPIs outlined above (Lateral Behaviour, Geometry, and Lateral and Heading Error) with the current SoA. Part of the challenge of the task lies in the comprehensive approach taken into the efforts to integrate several modules in WP4, and closely tied with different modules from WP3.

In simulations the performance of the vehicle and geometry will also be closely linked with the selection of the controller in place, so this is a variable to control.

4.2.2 Fail-safe Control

The objective is to evaluate a controller solution tightly integrated with motion planning task, with an additional layer for possible failure scenarios, and specifically test for an improved coverage for Model Predictive Control (MPC) failure to converge and the controller performance.

The evaluation scenarios would be using Carla framework to test against current baseline without fail-safe controller. A failure will be induced directly into the architecture, triggering the behaviour to change controllers, in different stages of the experiment run. The evaluations PIs are shown in Table 4.

Table 4: Evaluation PIs for the fail-safe module of EXP2

Evaluation PIs	**Description**

Maximum Control Error in normal mode	The maximum control error when the ego vehicle is in normal mode
Maximum Control Error in mode transition	The maximum control error when the ego vehicle transitions from normal to degraded mode
Average Control Error in normal mode	The average control error when the ego vehicle is in normal mode
Average Control Error in mode transition (normal to degraded)	The average control error when the ego vehicle transitions from normal to degraded mode
Time to stable conditions after non-converging MPC [ms]	The time required to establish stable conditions after non-converging MPC

Tecnalia has previously worked on failure-tolerant architectures, with a focus on the decision stage [17]. The scope of the work was the execution of a fallback manoeuvre after the event of a GNSS failure. In this previous work, the fallback mechanism was the compute of a degraded localization and the generation of a safe trajectory. The motion control module was based on an MPC that executed the planned trajectory. In this project, the aim is to address the additional issue of an MPC failing to converge.

5. EXP3 (Perception only)

EXP3 is concerned with safe automated driving in a complex urban environment with occlusion, to demonstrate the integration of reliability assessment outputs of environment state estimation (onboard self-assessment methods) and V2X data into an onboard perception system. The experiment will be conducted both in a virtual and a real environment. The former will be simulation-based, and it will be primarily concerned with developing a self-assessment layer for the perception data (T3.5) along with complementary V2X data (T3.4). The latter will be realized in UULM's vehicle, with safety drivers/marshals to account for the prototypical status of the developed system, and in UULM's V2X infrastructure pilot site, where the automated ego vehicle will face objects and (artificial) error/degradation in one of the sensors/V2X. EXP3 is led by the University of ULM (UULM).

In EXP3, UULM's plan is to evaluate the self-assessment (SA) module of the tracking. The target evaluation objective is the detection of certain errors and assumptions violations injected into the tracking algorithm. These are among other:

- Increased number of clutter detections (Violations of the clutter assumptions in tracking).
- Increased number of missed detections (Violations of the missed detections modelling in tracking).
- Increased measurement noise of the detections (Violations of the measurement noise modelling).
- Ambiguous data association decision in the tracking algorithm (caused by ambiguous data association situations.

The evaluation scenarios pertain to both simulation and real-world, as follows:

- Simulation: Monte-Carlo Runs with injected errors from above.
- Real-World: Analyse KITTI sequences towards tracking assumption violations (especially ambiguous data association situations).

The evaluations PIs are shown in Table 5.

Table 5: Evaluation PIs of the self-assessment module of EXP3

Evaluation PIs	Description
Reliability	Error detection rate obtained by Monte-Carlo simulation runs
Performance [ms]	Time from error is injected to error is detected in the SA module

The KITTI dataset is used for analysing the real-world data effects for the SA module (e.g., ambiguous data association situations).

With regards to a baseline, there are currently no other comprehensive SA modules for tracking algorithms to compare in the State-of-the-Art literature. However, the SA module can be compared to one specific consistency test typically used in tracking, which is the normalized innovation squared (NIS) and a multi-target variant the multi-target NIS (MNIS). In simulations with available ground truth, the SA module can also be compared with ground truth evaluation metrics as the GOSPA [19] or the RMSE error of the track estimates.

6. EXP4

EXP4 is an end-to-end experiment starting with the precise vehicle localization, by defining a semantic representation of the environment (T3.2), and the motion prediction of dynamic objects in the scene (T3.3). The localization of the ego-vehicle will be further enhanced by using V2X information (CAM, CPM and SPAT messages, optional, if available), thus increasing the reliability of its position in case of a failure or sensor blockage (T3.4). Particularly in the context of roadworks, unmarked lanes and narrow roads, the ego-vehicle performs a self-assessment by deciding whether to trust its perception system (T3.5). EXP4 is led by Hitachi (HIT-FR & HIT-UK) and Tecnalia (TECN), CRF and WMG also participate.

EXP4's evaluation plan includes work on both the perception layer and decision-making layer.

6.1 Perception Layer

With regards to the perception layer, the target evaluation objective is the performance of the perception module, tested in a scenario in which a two-lane road has only one lane available due to roadwork bollard in the middle of the road. The datasets used for the evaluation involve both public datasets (Zod, Coco) as well as collected data from public roads by HIT.

The evaluation PIs along with their respective baselines are shown in Table 6.

Table 6: Evaluation PIs of the perception layer of EXP4

Evaluation PIs	**Description**
Mean average precision	Camera based 2D object detection (Zod & Coco datasets)
IoU of drivable road segmentation	The intersection of union of a drivable road segmentation extracted from an HD map
Distance error	Lane boundary estimation based on roadwork bollard (Own baseline defined from the collected data)

6.2 Decision-making Layer (perception layer present)

With regards to the decision-making layer of EXP4, the work involves behavioral and motion planning, and fail-safe control. The methodology used for the evaluation is almost identical to the one described in EXP2 but for the ease of readability of the document, it is provided again in the following sections.

6.2.1 Behavioural and motion planning

The target evaluation objective with regards to the behavioural and motion planning is to integrate the perception inputs, to generate and appropriate plan and geometry to follow while maintaining a safety distance, and to reduce the risk situations to a minimum (based on TTC<1.5s). In more detail, the evaluation plan would:

- Provide an efficient plan for all automated participants in the scene at high level (maintain lane, change lane, enter roundabout, etc), measured in successful completion of simulated and hybrid scenarios, and time to finish the manoeuvre.
- Improve, in terms of smoothness and safety, the trajectories for automated vehicles, which integrate possible conflicts, and adapt appropriately. These will be measured in lateral accelerations and other comfort variables.
- Target to avoid risk situations in speed planning, measured in time to collision (TTC) and distance to collision (DTC).

The evaluations PIs are shown in Table 7.

Table 7: Evaluation PIs for the behavioural and motion planning module of EXP4

Evaluation PIs	Description
Max lateral acceleration [m/s^2]	The max lateral acceleration of the ego vehicle (Lateral behaviour)
Average lateral acceleration [m/s^2]	The average lateral acceleration of the ego vehicle (Lateral behaviour)
Average lateral error to reference trajectory	The average lateral error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Average angular error to reference trajectory	The average angular error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Average lateral error to center line	The average lateral error to the center of the line (Lateral and Heading Error)
Average angular error to center line	The average angular error to the center of the line (Lateral and Heading Error)
Max lateral error to reference trajectory	The maximum lateral error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Max angular error to reference trajectory	The maximum angular error to the reference trajectory generated by the motion planning module (Lateral and Heading Error)
Max lateral error to center line	The maximum lateral error to the center of the line (Lateral and Heading Error)
Max angular error to center line	The maximum angular error to the center of the line (Lateral and Heading Error)

Average difference of reference trajectory to center line [m]	The average difference between the reference trajectory generated by the motion planning module and the center of the line (Geometry)
Max difference of reference trajectory to center line [m]	The maximum difference between the reference trajectory generated by the motion planning module and the center of the line (Geometry)
Normalized max curvature [m]	The normalized maximum curvature of the reference trajectory generated by the motion planning module (Geometry)
Normalized average curvature [m]	The normalized average curvature of the reference trajectory generated by the motion planning module (Geometry)
Normalized max curvature difference to center line [m]	The normalized maximum curvature difference between the reference trajectory generated by the motion planning module and the center of the line (Geometry)
Normalized average curvature difference to center line [m]	The normalized average curvature difference between the trajectory generated by the motion planning module and the center of the line (Geometry)
Average computation time [ms]	The average computation time
Max computation time[ms]	The maximum computation time
Success rate [%]	The last vehicle exits the roundabout and there have not been any collisions (Decision Making)

Comparisons in decision-making and behavioural solutions are complex since its performance is heavily linked to the architecture in place. Nonetheless, EXP4 will focus on comparing the motion planning module, with the KPIs outlined above (Lateral Behaviour, Geometry, and Lateral and Heading Error) with the current SoA. Part of the challenge of the task lies in the comprehensive approach taken into the efforts to integrate several modules in WP4, and closely tied with different modules from WP3.

In simulations the performance of the vehicle and geometry will also be closely linked with the selection of the controller in place, so this is a variable to control.

6.2.2 Fail-safe Control

The objective is to evaluate a controller solution tightly integrated with motion planning task, with an additional layer for possible failure scenarios, and specifically test for an improved coverage for MPC failure to converge and the controller performance. The evaluations PIs are shown in Table 8.

Table 8: Evaluation PIs for the fail-safe module of EXP4

PI Category	**Evaluation PIs**

Error PI	Maximum Control Error in normal mode
	Maximum Control Error in mode transition (normal to degraded)
	Average Control Error in normal mode
	Average Control Error in mode transition (normal to degraded)
Time PI	Time to stable conditions after non-converging MPC [ms]

Tecnalia has previously worked on failure-tolerant architectures, with a focus on the decision stage [17]. The scope of the work was the execution of a fallback manoeuvre after the event of a GNSS failure. In this previous work, the fallback mechanism was the compute of a degraded localization and the generation of a safe trajectory. The motion control module was based on an MPC that executed the planned trajectory. In this project, the aim is to address the additional issue of an MPC failing to converge.

7. EXP5 (Perception only)

EXP5 is like EXP4 with two main differences. The first is that there is not self-assessment (T3.5) of the ego-vehicle. The second difference is that the motion planning involves path and speed planning as well as control of the different highway entering experiments. EXP5 is led by Hitachi (HIT-FR & HIT-UK) and Tecnalia (TECN), and CRF also participate.

In EXP5, the target evaluation objective is to evaluate the performance of the motion prediction solution against the SoA and against the ground truth.

The scenarios to evaluate the module would be performed in a hybrid setup in which HIT provides the detection and TECN assesses the solution with real-world data. The datasets Argoverse 1 and V2V4Real have been used to test and train the current implementation of the motion prediction module. The evaluations PIs that will be used are shown in Table 9.

Table 9: Evaluation PIs for the perception layer of EXP5

Evaluation PIs	Description
minADE (Minimum Average Displacement Error) [m]	The average distance between the best forecasted trajectory and the ground truth, compared against map-less SoA solutions.
minFDE (Minimum Final Displacement Error) [m]	The distance between the endpoint of the best forecasted trajectory and the ground truth, compared against map-less SoA solutions.
MR (Miss Rate) [%]	The number of scenarios where none of the forecasted trajectories are within 2 meters of ground truth according to endpoint error.
brier-minADE (Brier minimum Average Displacement Error) [m]	Similar to minADE with an added probability of the best forecasted trajectory.
brier-minFDE (Brier minimum Final Displacement Error) [m]	Similar to minFDE with an added probability of the best forecasted trajectory.

8. EXP6 (Perception only)

EXP6 concerns the sensing of small objects and semantic representation of these objects (relative position, height, object velocity, over-drivability and estimation of time to collision) within diverse weather conditions where the object might not be clearly visible to the human eye and a critical decision on the vehicle behaviors shall be taken to either avoid a potential frontal collision if the object is not over-drivable by braking or avoid a potential rear collision with other vehicles driving behind if the object is over-drivable due to unnecessary braking. EXP6 is led by APTIV (APTIV-DE & APTIV-FR).

The evaluation of EXP6's perception algorithm will be done using real test data collected on a test track and using simulations.

On the test track, data for different objects recorded at different vehicle approach speeds will be collected.

Additionally, simulations will be used to evaluate our perception system for a wider range of variations in vehicle states and object dimensions and pose.

KPIs with goal values that ensure safety and comfort will be used to evaluate our perception system. The goal values will be derived as part of task T5.3.

The confusion matrix, shown in Table 10, will be used to differentiate between true positive predictions (TP), false position predictions (FP), false negative predictions (FN) and true negative predictions (TN) made by the perception system.

Table 10: Confusion matrix

		Predicted	
		Overdriveable	Non-driveable
Actual	Overdriveable	TN	FP
	Non-driveable	FN	TP

The perception level KPIs, shown in Table 11, will be evaluated on the real test track data and on simulation data.

Table 11: Evaluation PIs of the perception layer of EXP6

Evaluation PIs	**Description**
Distance to debris [m]	Distance to the object at the timestamp when we have the same prediction class for at least 0.25 s
TTC to debris [m]	TTC to the object at the timestamp when we have the same prediction class for at least 0.25 s

Mean accuracy of longitudinal position of detected object [m]	The mean longitudinal position error value must be within the desired range. The closer to 0 the better. The error is computed by comparing the object cluster position provided by the perception system compared to the ground truth position.
Upper bound accuracy of longitudinal position of detected object [m]	The mean longitudinal position error value plus 3 standard deviations of the error must be less than the goal value. The closer to zero the better.
Lower bound accuracy of longitudinal position of detected object [m]	The mean longitudinal position error value minus 3 standard deviations of the error must be greater than the goal value. The closer to zero the better.
Mean accuracy of lateral position of detected object [m]	The mean lateral position error value must be within the desired range. The closer to 0 the better.
Upper bound accuracy of lateral position of detected object [m]	The mean lateral position error value plus 3 standard deviations of the error must be less than the goal value. The closer to zero the better.
Lower bound accuracy of lateral position of detected object [m]	The mean lateral position error value minus 3 standard deviations of the error must be greater than the goal value. The closer to zero the better.
Recall/True Positive Rate	This is the ratio between TP predictions and the actual ground truth positive class. $$TPR = \frac{TP}{TP + FN}$$
Precision/Positive Predictive Value (PPV)	This is the ratio between the TP predictions and all the predicted positive class. $$PPV = \frac{TP}{TP + FP}$$
F1-score	This is the harmonic mean between the precision and recall. $$\text{F1score} = 2\frac{TPR * PPV}{TPR + PPV}$$
True Negative rate	This is the ratio between TN predictions and the actual ground truth negative class. $$TNR = \frac{TN}{TN + FP}$$
Short range consecutive FP predictions	The median number of consecutive FP predictions when the object is less than 50 m away from the vehicle.

Long Range consecutive FP predictions	The median number of consecutive FP predictions when the object is less than 250 m away from the vehicle.
Short range consecutive FN predictions	The median number of consecutive FN predictions when the object is less than 50 m away from the vehicle.
Long Range consecutive FN predictions	The median number of consecutive FN predictions when the object is less than 250 m away from the vehicle.
Perception latency	The time difference between the moment the object is sensed by the perception system (object radar detection available) and the moment in which a stable classification is outputted. Stable classification implies that the output class is constant for 0.25 s.

The above defined KPIs have a direct influence on the end to end system KPIs. The end to end system KPIs, shown in Table 12, will not be tested since the DM module is not being addressed in this experiment, hence are only being provided for completeness sake.

Table 12: End-to-end KPIs

Evaluation PIs	**Description**
Maximum deceleration [m/s^2]	Maximum deceleration required to come to stop for non-driveable object or to slow down for an overdriveable object
Smallest longitudinal distance to non-driveable object [m]	Smallest longitudinal distance to non-driveable object after stopping in lane
Smallest longitudinal distance to an overdriveable object [m]	The smallest distance along the direction of the travel of the ego vehicle with respect to the closest overdriveable object in the ego lane shall be 0 m, as the ego vehicle is expected to drive over the object.
Ego vehicle velocity when driving over an overdriveable object [m/s]	The ego vehicle velocity when driving over an overdriveable object shall be between 0.9* goal value and the goal value.
Ego vehicle velocity in case of collision with non-driveable object [m/s]	The relative velocity with respect to the non-driveable object in case of collision should be less than the safety goal value.

9. EXP7

This experiment focuses on the development of an integrity monitoring mechanism for estimating the distance to the leading vehicle in urban and highway environments under adverse operational domain conditions. The mechanism should reliably indicate the point in time when the relative localization of the ego-vehicle with respect to the leading vehicle must not be trusted and/or the object detection and tracking becomes unreliable. Another objective (not related with the self-assessment objective) is to study the effects of adverse weather conditions on a perception module performing other vehicles' behaviour prediction. EXP7 is led by WMG and ICCS also participates.

9.1 Perception Layer

EXP7 consists of only a perception layer, which in turn incorporates a self-assessment layer and a prediction module.

9.1.1 Perception/Self-assessment layer

The perception system in EXP7 comprises camera-based and lidar-based object detection. The output of the developed self-assessment system for perception is binary classification (Error vs No-error) determining whether to trust the current perception frame or not. For camera-based detection, a perception error is declared when the mAP is estimated to be less than 0.5, while for lidar-based 3D object detection, a perception error is declared when at least one object (vehicle or pedestrian) in the input frame is not detected. For performance evaluation lidar data will be collected in public roads. In addition, public datasets will be used to assess the performance of the self-assessment mechanism both for camera-based and lidar-based object detection, as follows:

- Target evaluation objective: Evaluating the performance of a DNN-based self-assessment system for camera-based 2D object detection and lidar-based 3D object detection. For the object detection tasks YOLOv8 (2D) and Centrepoint (3D) are used.
- Evaluation scenarios: Data collection (lidar point clouds) in public roads during: (i) Urban driving under normal weather as well as (ii) Motorway driving under normal and adverse weather conditions.
- Datasets: The introspection system has been trained using: (i) YOLOv8 in KITTI and BDD for 2D object detection and (ii) Centrepoint in NuScenes for 3D object detection. Note that the NuScenes dataset contains inputs under adverse weather too. For performance evaluation using lidar data from public roads, a small labelling task is required, i.e., generating binary labels indicating whether

a pedestrian or a vehicle is missed in the input frame. To facilitate labelling, camera data is also collected at the same time.

- Baseline: The developed introspection system has been trained using: (i) activation maps associated with mAP (2D introspection), and (ii) activation maps associated with at least one missing object (3D introspection). Only the vehicle and pedestrian classes are considered by the introspection system. The baseline introspection system for comparison uses some statistical features of the activation maps instead of the full activation map.

Regarding the self-assessment of the distance-estimation to the leading vehicle, the distance to the leading vehicle, in EXP7, is estimated using a lidar. Centrepoint is used for object detection and the closest object in terms of longitudinal distance is declared to be the leading vehicle. Objects in the same lane and within a specified maximum longitudinal distance to the ego vehicle are only considered. Under adverse weather, the performance of the lidar detections can severely degrade and a radar sensor is used to obtain the ground truth distance. The self-assessment mechanism is trained based on the activation maps and ground truth data collected using a radar to predict whether the distance estimation error to the leading vehicle is significant or not. Significance depends on the application, i.e., urban or motorway driving. The distance estimation error is calculated based on the longitudinal distance of the detected and ground truth bounding box centers of the lead vehicle. If the lead vehicle is missed the distance error equals the distance between the ground truth and the maximum longitudinal distance for lidar detection. If a ghost lead vehicle is detected the distance error equals the distance between the prediction and the maximum longitudinal distance for lidar detection. Self-assessment of the distance estimation to the leading vehicle is part of the ego vehicle's ACC.

- Target evaluation objective: Evaluating the performance of a DNN-based self-assessment system of lidar-based lead vehicle detection and distance estimation to the Ego vehicle. For lead vehicle detection Centrepoint (3D) is used.

- Evaluation scenarios: Data collection for: (i) urban driving under normal weather as well as (ii) motorway driving under normal and adverse weather conditions.

- Datasets (train, test): The introspection system has been trained using Centrepoint in NuScenes including adverse weather. For testing data collection is carried out. For the ground truth of the distance to the leading vehicle, radar data is used.

- Baseline (to which system are you comparing against): Not yet decided.

Regarding the integrity monitoring for GNSS-based localisation:

- Target evaluation objective: Evaluating the performance of an integrity monitoring mechanism for GNSS-based localization. The mechanism leverages neural networks to calculate the horizontal protection level, which is an upper bound to the horizontal positioning error.
- Evaluation scenarios: Data collection at fixed locations while being static. The average of the best position signals across a time interval shall be considered as the ground truth and the position errors are calculated accordingly.
- Datasets (train, test): Not yet decided.
- Baseline: Receiver-autonomous integrity monitoring (RAIM) using weighted least-squares estimation.

Evaluation PIs	**Description**
Frame-level error performance	Assuming that frame-level errors are associated with the positive class of the self-assessment mechanism, its performance is evaluated using class specific recall values and the AUROC (the higher the better). The inference time is also measured.
Binary classification [True or False]	Binary classification whether the estimated distance to the leading vehicle can be trusted or not. The inference time is also measured
Horizontal protection level [m]	Evaluating the integrity monitoring for GNSS-based localisation

9.1.2 Prediction module

ICCS will contribute to the maneuver and trajectory prediction task (assuming given object bounding boxes). The scenarios evaluated are per EXP7 UCs (highways) and Open Road (runtime) evaluation is considered.

In accordance with KPIs defined previously in D2.1, D3.1, D4.1 and with the current State-of-the-art literature, the evaluation KPIs for T6.1 are reported and described in Table 13 and Table 14.

Table 13: Maneuver prediction KPIs and expected performances after module completion

Evaluation PIs	**Description**	**Formula**

Classification Accuracy (Acc.)	Percentage Value of correctly predicted maneuvers (online) of each non-ego vehicles during the prediction horizon, averaged per time horizon (weights), scene agents	$\frac{1}{NxT}\sum_{i=0}^{i=N-1}\sum_{t=0}^{t=T-1} w_t \frac{TP+TN}{TP+FP+TN+FP}$ w_t : weighting scheme, $w = [0.44,0.33,0.22]$ P: positive class (Lane Change Right) N: negative class (Lane Change Left)
ROAUC	Area under receiver operating curve, averaged over time and agents	$\frac{1}{NxT}\sum_{i=0}^{i=N-1}\sum_{t=0}^{t=T-1}\int_0^1 TPR\, dFPR$ TPR: True Positive Rate FPR: False Positive Rate
Precision	Precision averaged over time, agents and classes	$\frac{1}{NxT}\sum_{i=0}^{i=N-1}\sum_{t=0}^{t=T-1} w_t \frac{TP}{TP+FP}$
True Negative Rate	TNR averaged over time, agents and classes	$\frac{1}{NxT}\sum_{i=0}^{i=N-1}\sum_{t=0}^{t=T-1} w_t \frac{TN}{TN+FN}$

Table 14: Trajectory prediction KPIs and expected performances after module completion

Evaluation PIs	Description	Formula
Minimum Average Displacement Error (minADE)	Minimum displacement error of each non-ego vehicle during the prediction horizon, averaged per time (weights) scene agents and weights	$\frac{1}{NxTxK}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1}\sum_{t=0}^{t=T-1} w_t * \left(x_{i,t} - \widehat{x_{i,t}}\right)^2$ $x_{i,t}$: Logged trajectory (2D) ground truth $\hat{x}_{i,t}$: Predicted 2D waypoint
Minimum Final Displacement Error (minFDE)	Final displacement error of each non-ego vehicle during the prediction horizon, averaged per time (weights) and scene agents	$\frac{1}{NxTxK}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1} w_t * \left(x_{i,T} - \widehat{x_{i,T}}\right)^2$
Minimum Average Displacement Error over top k scored trajectories (minADEk)	Minimum displacement error of each non-ego vehicle during the prediction horizon, averaged per time (weights), scene	$\min_{k*\in 1,\dots K}\frac{1}{NxT}\sum_{i=0}^{i=N-1}\sum_{t=0}^{t=T-1} w_t * \left(x_{i,t,k*} - \widehat{x_{i,t,k*}}\right)^2$ $k^* = \text{argmax}_{I\subset K,\lvert I\rvert=N} P\left(S(k)\right)$

	agents and over top k best generated trajectories (multimodality)	P(S): Power set of S S(k): Planning-aware scoring function of trajectory k
Minimum Average Displacement Error in Dynamic evaluation (Dynamic minADE)	Minimum average Displacement Error of each non-ego vehicle during the prediction horizon, averaged over time and scene agents, in dynamic evaluation (simulated motion model of non-ego agents)	$\min_{k*\in 1,\ldots,K} \frac{1}{\mathrm{NxT}} \sum_{i=0}^{i=N-1} \sum_{t=0}^{t=T-1} \mathrm{w_t} * \left(\mathrm{x_{i,t,k*}} - \mathrm{x^{sim}_{i,t,k*}}\right)^2$ $k^* = \mathrm{argmax}_{I\subset K,\lvert I\rvert=N} P\big(\mathrm{S}(k)\big)$

The model baselines for EXP7 were considered in a top-down approach, following the recent advancements in the manoeuvre prediction field, as shown in Table 15.

Table 15: Model's baselines

Method	Description
CNN	CNN on preprocessed RGB Ego-view images augmented with context and history [7]
CNN Encoder + LSTM	CNN encoder (ResNet) and LSTM for prediction [8]
Bayesian Network	Bayesian network for lane change prediction
Interaction Network	GNN on vectorized scene and agent features, pretrained on VectorNet [9]

The developed module is to be evaluated both in real and simulation datasets, as shown in Table 16.

Table 16: Simulation and real-world datasets for EXP7

Dataset	Description	Type
PREVENTION	RGB Video of Ego View highway scenes	Real
BDD100k	RGB Video of Ego View in highway, urban scenes in US	Real
KITTI	Grayscale stereo video of Ego View in highway, urban scenes	Real

Synthetic simulation dataset (ICCS)	Highway multi-agent driving scenes developed in CARLA by ICCS	Simulation

10. EXP8

The low atmospheric visibility in adverse weather conditions like fog, snow, and rain reduces the maximum viewing distance of LiDAR sensors. This in turn decreases the object detection and localization performance and cause safety hazards. Weather conditions have effect on sensing and therefore on perception and localization of automated driving system. Use case provides possibility to evaluate the on-board visibility-based localization performance estimate. Safe vehicle control is necessary in case the weather conditions worsen and fail-safe behavior in case of exiting the ODD completely due to extreme weather. EXP8 is led by TUD and Perciv.AI (PERCIV) also participates.

In EXP8, the evaluation plan includes work on the perception layer and on the decision-making layer, performed by PERCIV and TUD respectively.

10.1 Perception Layer

With regards to the perception layer, a detailed evaluation of two key modules will be conducted, a radar point cloud segmentation and a radar-based ego-motion estimation, developed in WP3 (T3.2). The evaluation will focus on the reliability and accuracy of these modules to ensure they are dependable for subsequent development phases.

The primary evaluation objective is to validate the reliability of the perception modules to serve as a trustworthy source for subsequent development stages. The evaluation will test the modules' performance in real-world conditions using PERCIV's proprietary dataset, which encompasses real-world driving conditions. This dataset includes a variety of environmental and traffic scenarios, providing a comprehensive base for evaluation.

Each module will be compared to specific baselines to underscore the advancements made by PERCIV, as follows:

- Radar Point Cloud Segmentation:
  - Ghost Target vs. Real Radar Targets: Performance will be compared to the method proposed by Chamseddine et al., "Ghost target detection in 3D radar data using point cloud based deep neural network."
  - Moving vs. Static Targets: We will benchmark against statistical methods for detecting moving targets, such as those discussed by Palffy et al. in "Detecting Darting Out Pedestrians with Occlusion Aware Sensor Fusion of Radar and Stereo Camera."
  - Road User Segmentation: This will be evaluated against our state-of-the-art deep learning-based top-down radar object detection

approach, which estimates bounding boxes for each object. We will use these boxes to "paint" the radar points with class information.

- Radar-Based Ego-Motion Estimation: This module's performance will be benchmarked against camera-based odometry systems, demonstrating radar's effectiveness even in poor visibility conditions.

The evaluation will utilize scenarios from PERCIV's real-world dataset to ensure the testing environment reflects typical operational conditions and the evaluation PIs are shown in Table 17.

Table 17: Evaluation PIs of the perception layer of EXP8

Area	Evaluation PIs
Radar Point Cloud Segmentation	The Intersection over Union (IoU) will be used to quantify the accuracy of segmentations [%].
Radar-Based Ego-Motion Estimation	Metrics will include drift over time and average mean square error, assessing the precision of the module's motion estimations.

10.2 Decision-Making Layer

With regards to the decision-making layer, the target evaluation objective is to assess the motion planner with decision-making to avoid collision and keep a vehicle stable. TUD will validate the motion planner using Monte-Carlo simulations with high-fidelity multibody software IPG/CarMaker. The evaluation scenario includes TUD's demo vehicle performing an emergency evasive manoeuvre on a slippery road under rain conditions to avoid a low-speed dynamic obstacle.

The evaluations PIs that are going to be used to assess the performance of motion planner with decision-making are shown in Table 18.

Table 18: Evaluation PIs of the decision-making layer of EXP8

Evaluation PIs	Description
Vehicle-to-Obstacle distance [m]	It is a Euclidian distance between the ego-vehicle and the obstacle.
Time to Collision (TTC) [ms]	Essential to evaluate the decision-making logic
Time to Brake (TTB) [ms]	Essential to evaluate the decision-making logic
Time to Steer (TTS) [ms]	Essential to evaluate the decision-making logic
Vehicle sideslip angle [degrees]	The angle between the vehicle centre line and the vehicle velocity vector. The absolute sideslip angle is small and below 5 degrees in stable driving situations

Computation time [ms]	It should remain below the allocated planning time (below 50 ms)

The developed planner will be compared against a SoA approach based on nonlinear Model Predictive Control. It combines motion planning, path tracking and vehicle stability into a single controller. The prediction model is based on a nonlinear single-track and a Fiala tyre model [16]. The obstacles and the vehicle are represented as circles, so the distance between the vehicle-to-obstacle is constantly monitored, and the controller uses it to prioritize obstacle avoidance rather than path tracking in an emergency. The prediction model kinematics is formulated using a Frenet reference system [22], so the vehicle-to-obstacle distance is also measured using this coordinate framework.

11. Summary & Conclusion

This Deliverable (D6.1), which is associated with task T6.1, describes the methodology and scope for the evaluation of the EVENTS experiments. The aim of the corresponding work package (WP6) is to evaluate the various experiments, EXP1 to EXP8, specified by the EVENTS project partners. A summary of the eight EVENTS experiments with their respective AD stack layers and the partners that work on each of these layers is shown in Figure 3.

Experiment	Perception & Prediction	Decision-making	E2E System
EXP1	TUD	TUD	✓
EXP2	ICCS, TECN	TECN	✓
EXP3	UULM	N/A	×
EXP4	HIT, WMG	TECN	✓
EXP5	TECN, HIT	N/A	×
EXP6	APTIV	N/A	×
EXP7	ICCS, WMG	N/A	×
EXP8	PERCIV	TUD	✓

Figure 3: Summary of EVENTS experiments

In addition, a summary of all the partners working on different datasets for the evaluation of their modules is shown in Figure 4.

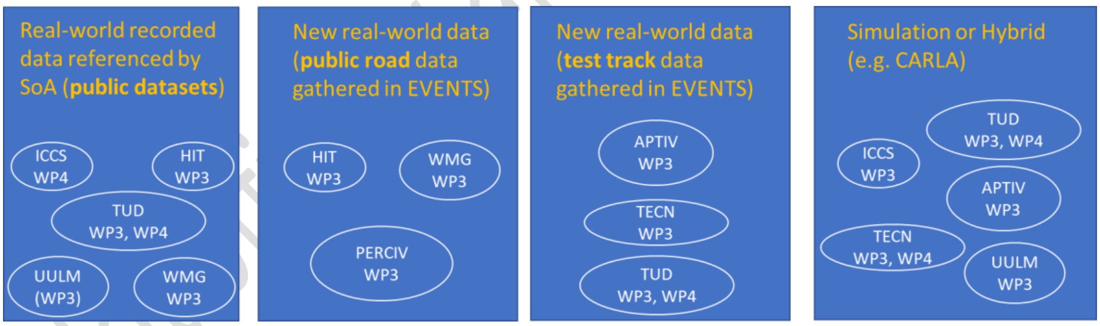

Figure 4: Data for module evaluation

The eight experiments, described in this document, particularly with regards to their perception and prediction layer, have presented and will use a variety of methods and subsequently different PIs for their evaluation. It should be noted that while the reported methods will be thoroughly evaluated, they will not be benchmarked between them, since each experiment tackles different challenges.

The forthcoming deliverable D6.2 (Technical evaluation results), will report on the results of the technical evaluation and small-scale SOTIF verification (part of task T6.3), which will be based on data collected from the various abovementioned experiments in controlled real-world environment or in simulations (part of task T6.2).

Disclaimer of Warranties

The information and views set out in this deliverable are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the following information.

References

[1] EVENTS Deliverable D2.1: User and system requirements for selected use cases, 2023.

[2] EVENTS Deliverable D2.2: Full Stack Architecture & Interfaces, 2023.

[3] EVENTS Deliverable D2.3: Vehicle System Hazard Analysis & Risk Assessment, 2023.

[4] EVENTS Deliverable D3.1: Perception Components Methods, 2023.

[5] EVENTS Deliverable D4.1: Initial version of motion planning and behavioural decision-making components, 2024.

[6] Hi-Drive Deliverable D4.5: Effects evaluation methods. Co-funded by the European Union under Horizon 2020 programme. Grant Agreement No 101006664. https://www.hi-drive.eu/app/uploads/2023/09/Hi-Drive-SP4-D4.5-Effects-evaluation-methods-v1.0-.pdf.

[7] Izquierdo, R., Quintanar, A., Parra, I., Fernández-Llorca, D., & Sotelo, M. (2019). Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (pp. 3657-3662).

[8] Bahram, M., Hubmann, C., Lawitzky, A., Aeberhard, M., & Wollherr, D. (2016). A Combined Model- and Learning-Based Framework for Interaction-Aware Maneuver Prediction. IEEE Transactions on Intelligent Transportation Systems, 17(6), 1538-1550.

[9] Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., & Schmid, C. (2020). Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11525-11533).

[10]Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. International Journal of Robotics Researc, 32(11): 1231-1237.

[11]Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... & Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. Conference on Computer Vision and Pattern Recognition, pp. 11621-11631.

[12]Alibeigi, M., Ljungbergh, W., Tonderski, A., Hess, G., Lilja, A., Lindström, C., ... & Petersson, C. (2023). Zenseact Open Dataset: A large-scale and diverse multimodal dataset for autonomous driving. International Conference on Computer Vision, pp. 20178-20188.

[13]https://autowarefoundation.github.io/autoware.universe/main/planning/behavior_path_avoidance_module/

[14]https://autowarefoundation.github.io/autoware.universe/main/planning/obstacle_avoidance_planner/

[15]Yin, T., Zhou, X., & Krahenbuhl, P. (2021). Center-based 3d object detection and tracking. Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11784-11793).

[16]Brown, M., Gerdes, J.C. (2019). Coordinating tire forces to avoid obstacles using nonlinear model predictive control. IEEE Transactions on Intelligent Vehicles, 5(1): 21-31.

[17]J. A. Matute-Peaspan, J. Perez, and A. Zubizarreta, “A fail-operational control architecture approach and dead-reckoning strategy in case of positioning failures,” Sensors, vol. 20, no. 2, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/2/442.

[18]Berk, M. (2019). Safety assessment of environment perception in automated driving vehicles (Doctoral dissertation, Technische Universität München).

[19]Rahmathullah, A. S., García-Fernández, Á. F., & Svensson, L. (2017). Generalized optimal sub-pattern assignment metric. In 2017 20th International Conference on Information Fusion (Fusion) (pp. 1-8). IEEE.

[20]https://www.sae.org/news/2021/08/designing-and-assessing-vehicle-safety-functions-with-a-use-case-approach.

[21]Menzel, T., Bagschik, G., Isensee, L., Schomburg, A., & Maurer, M. (2019, June). From functional to logical scenarios: Detailing a keyword-based scenario description for execution in a simulation environment. In 2019 IEEE Intelligent Vehicles Symposium (IV) (pp. 2383-2390). IEEE.

[22]Avanzini, G. (2004). Frenet-based algorithm for trajectory prediction. Journal of guidance, control, and dynamics, 27(1), 127-135.

[23]International Organization for Standardization. (2018). *Safety of the intended functionality* (ISO Standard No. 21448:2022).