Safe and Explainable
Critical Embedded Systems based on AI

Safety lifecycle for DL-based systems

Final Event

Javier Fernández – IKERLAN

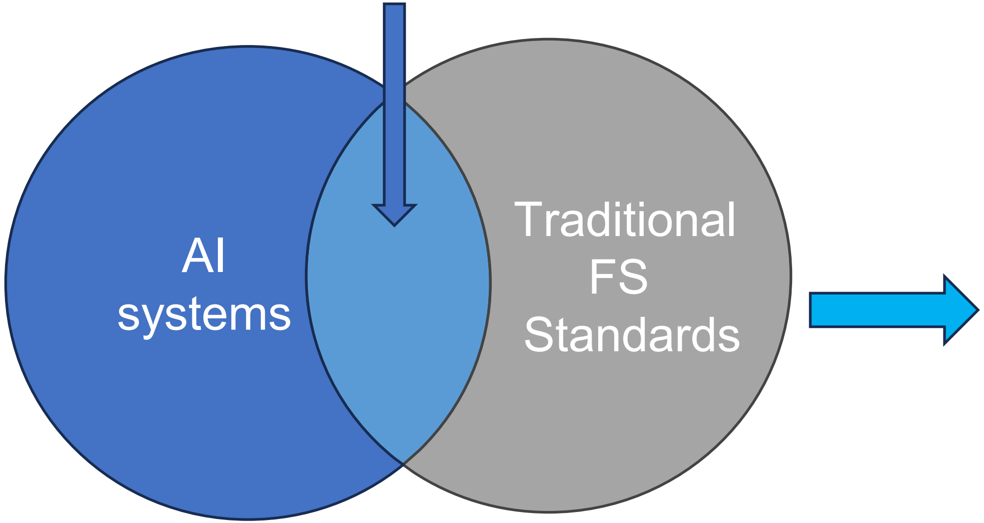
Giuseppe Nicosia, Carlo Donzella, Stefano Lodico – EXIDA ENGINEERING

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe programme under grant agreement number 101069595.

AI-FSM, Safety patterns and a V&V Strategy

Recommendations from emerging initiatives and standards



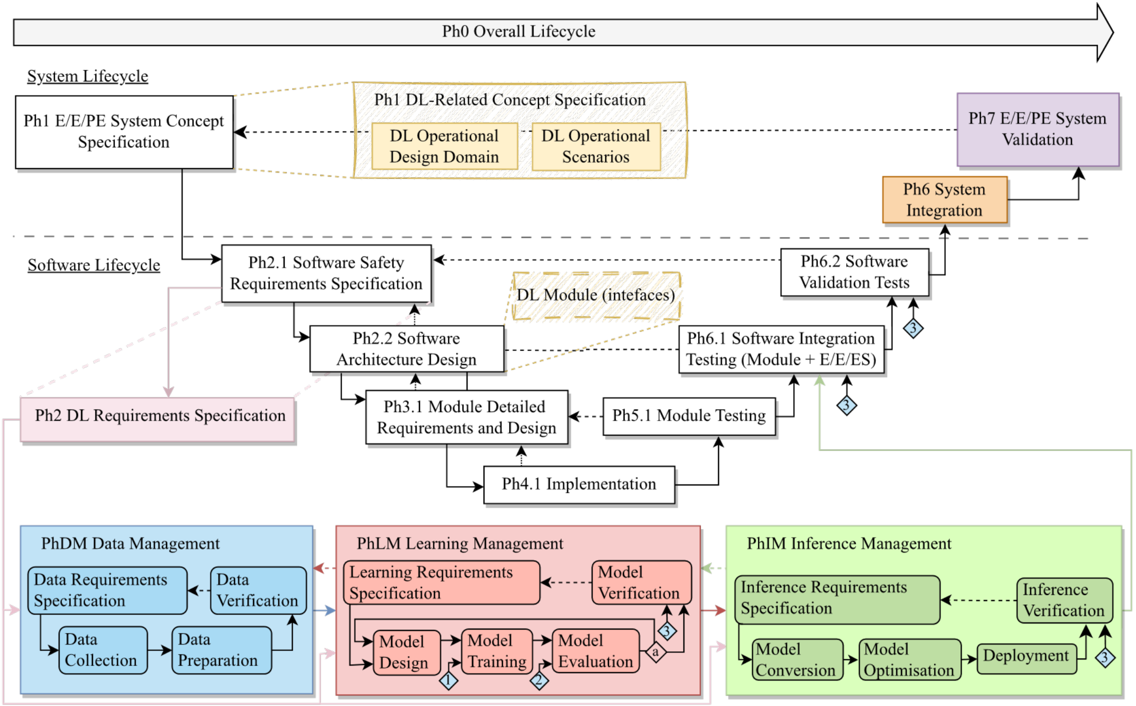
Contributions:

- *AI-FSM annex to be accommodated together with the traditional FSM*
- *Architectural safety patterns according to the capability level.*
- *V&V Strategy*

AI-FSM

AI-FSM lifecycle: DL spec., design and implementation

- IEC 61508 traditional functional safety lifecycle (Software V-model) + AI lifecycle



Define of AI-Functional safety management process based on **new initiatives and standards** (AMLAS, EASA, ISO/IEC TR 5469, SOTIF, Automotive SPICE)

Procedure

Guidelines

Templates

Internal review checklists

SAFEXPLAIN

AI-FSM lifecycle: DL spec., design and implementation

- AI-FSM reviewed by certification experts
  - Exida (as part of the consortium)
  - TÜV Rheinland (external certification body)

TÜVRheinland®
Precisely Right.

5. Conclusions

After having reviewed the updated versions 2.0 of the items under review as presented in chapter 3.2, no remaining deficiencies have been revealed. The items under review according to chapter 3.2 are considered as suitable for the intended purpose (incorporating artificial intelligence into IKERLAN's safety management system).

TÜVRheinland®
Precisely Right.

TÜV Rheinland InterTraffic GmbH
Assessment & Certification Rail Service

Independent Review Report
on the European SAFEEXPLAIN project w.r.t.
IEC 61508 / EN 5012x / ISO 5469

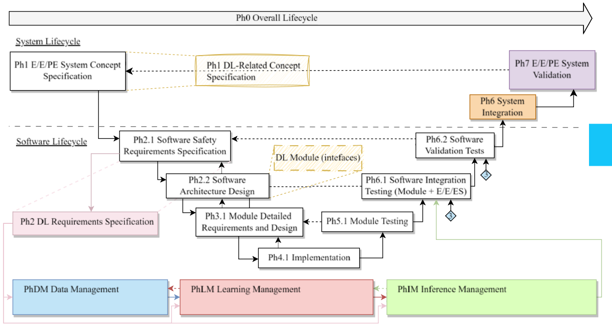
Report ID and Version	ACR/B 24/105-V1.0	
Report Date	2024-03-07	
Role	**Name**	**Signature**
Author	Dr. Hendrik Schäbe	Digital unterschrieben von Hendrik Schäbe Datum: 2024.03.07 15:32:50 +01'00'
Quality Reviewer	Dr. Ralf Röhrig	2024.03.08 14:39:13 +01'00'
Approved by	Dipl.-Ing. Peter Wigger	2024.03.08 19:46:36 +01'00'

SAFEXPLAIN

Architectural patterns for FuSa compliance

T2.3 DL Safety Architectural design Patterns

Reference software development process



Need for <u>runtime</u> safety mechanisms to deal with:

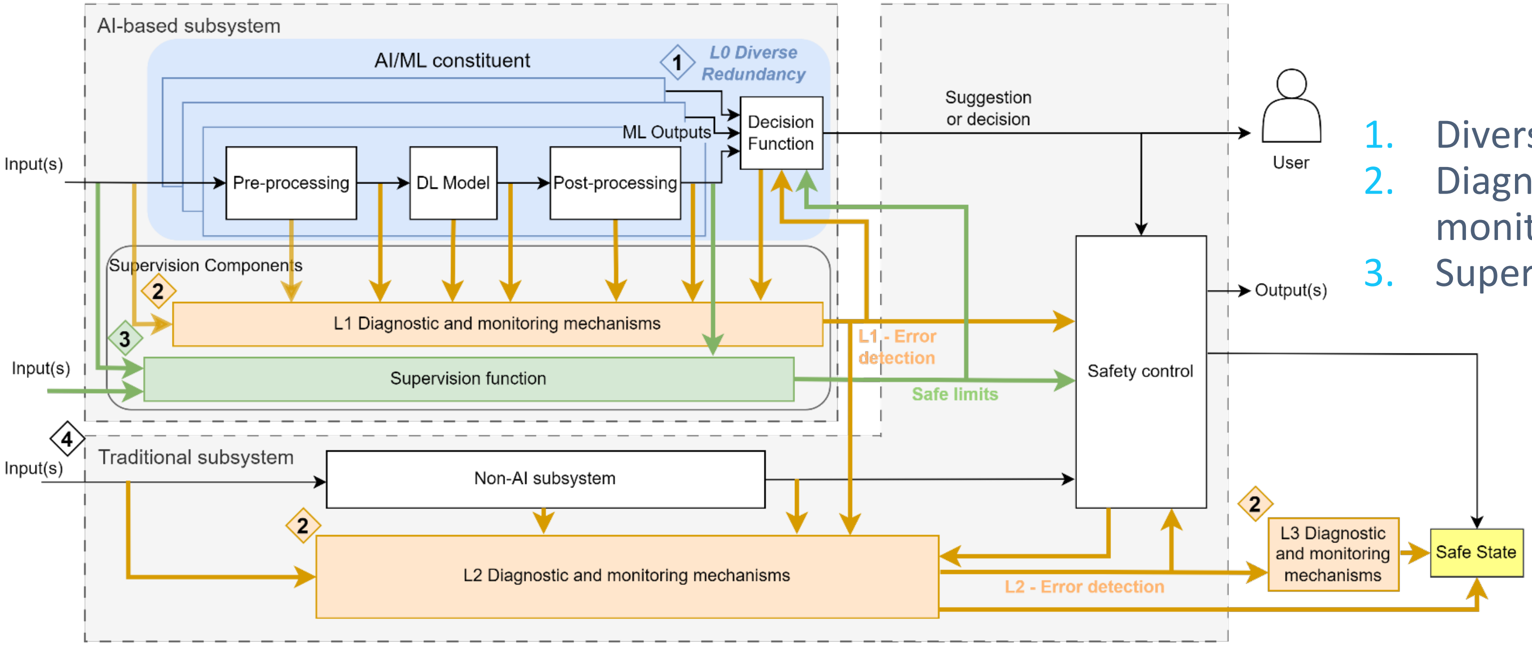
- *Random and systematic faults*
- *HW / SW platform complexity: integration problems (e.g., determinism, interferences on mixed-criticality approaches, use of resources…)*
- *DL model insufficiencies*
- *Support DL explainability*

…

GOAL: To provide reference safety architecture patterns for the adoption of DL in safety-critical systems with varying safety requirements

T2.3 DL Safety Architectural design Patterns

- Safety pattern: Generic solutions for commonly recurring design problems with the aim of simplifying and standardizing the design process



1. Diverse Redundancy
2. Diagnostic and monitoring mechanisms
3. Supervision function

T2.3 DL Safety Architectural design Patterns

- L0 Diverse Redundancy: Inference Platform diversity (inputs)

Image Transformation LockStep (ITLS)

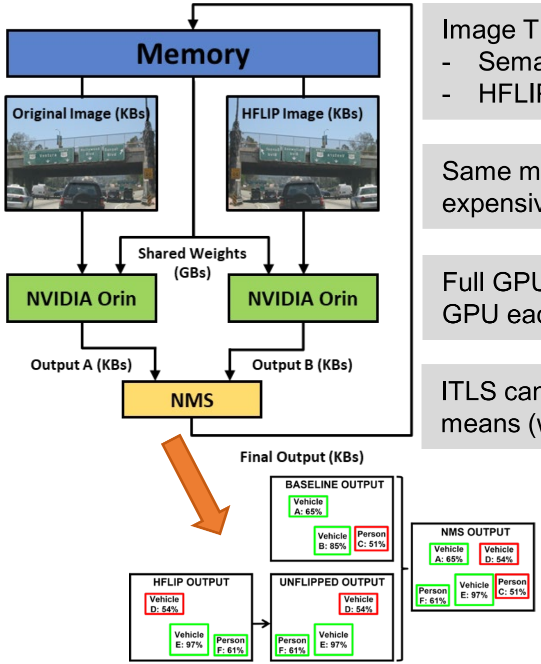


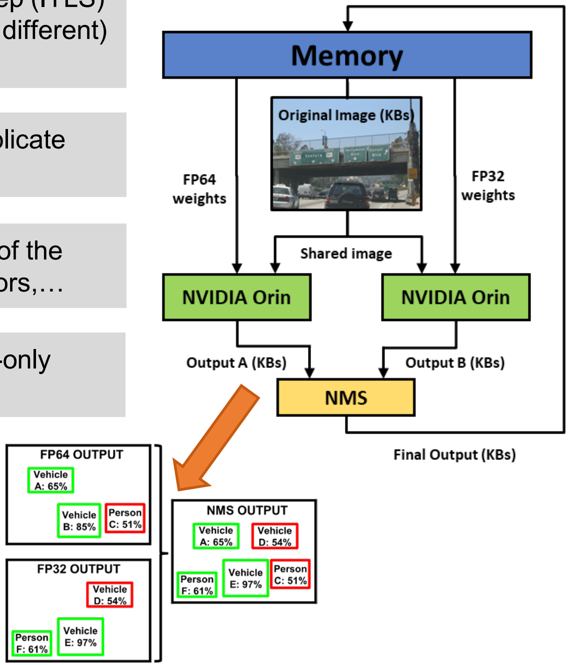
Image Transformation LockStep (ITLS)
- Semantic lockstep (bitwise different)
- HFLIP best choice

Same model twice (do not duplicate expensive weight fetching)

Full GPU both, different parts of the GPU each, different accelerators,…

ITLS can be realized with SW-only means (without HW support)

Diverse (native) data types (e.g., FP32 and FP64, or INT16 and INT32)



Multi data type lockstep (MDTLS)
- Semantic lockstep (bitwise different)
- FP easier (no calibration needed)

Same model twice if using "casts" inside (do not duplicate expensive weight fetching)

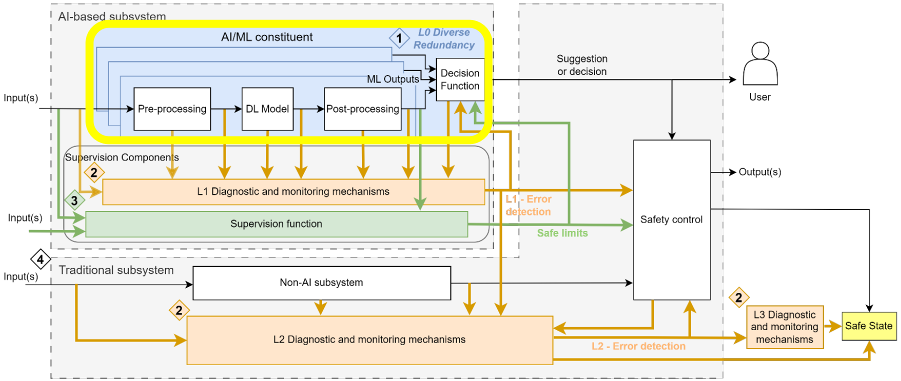
Full GPU both, different parts of the GPU each, different accelerators,…

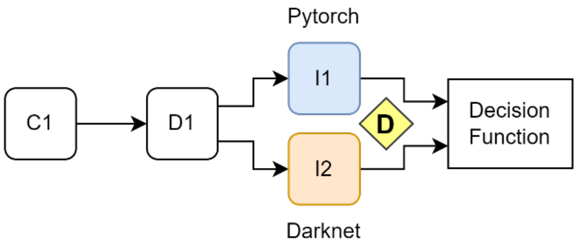
MDTLS can be realized with SW-only means (without HW support)

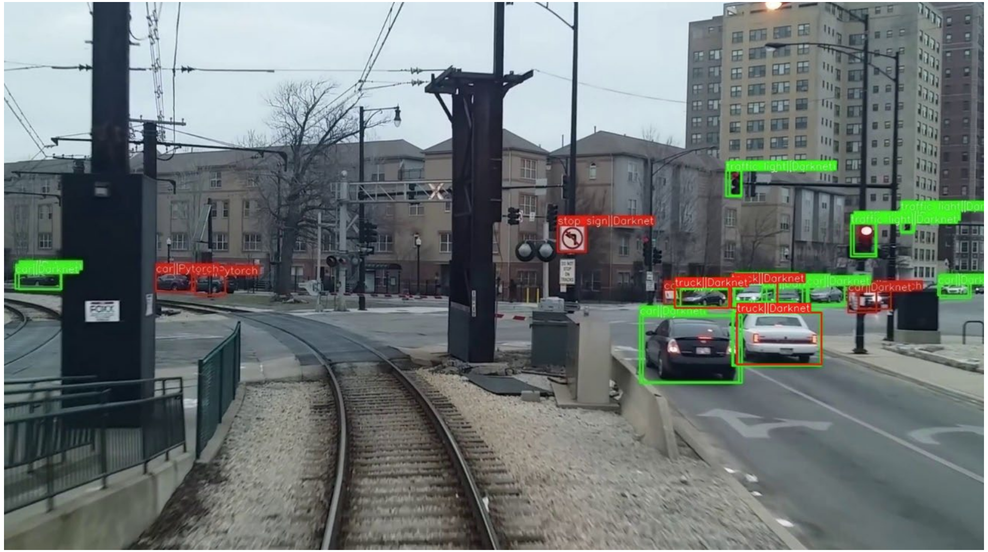
SAFEXPLAIN

T2.3 DL Safety Architectural design Patterns

- L0 Diverse Redundancy – Inference platform diversity using diverse redundant frameworks (i.e., Pytorch and Darknet).



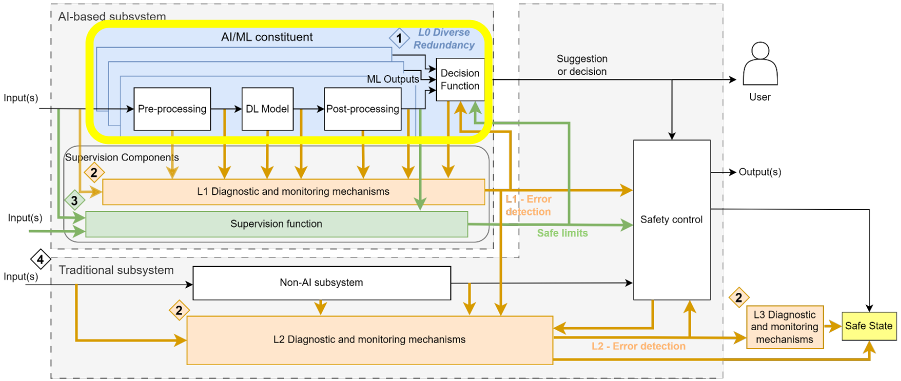


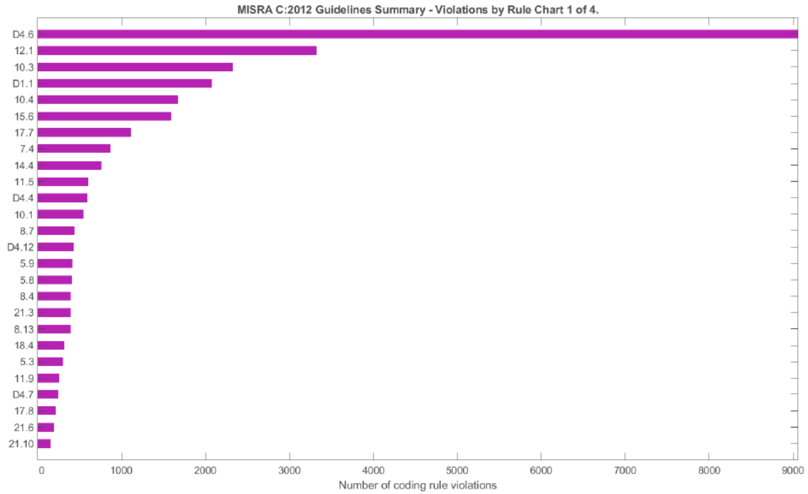


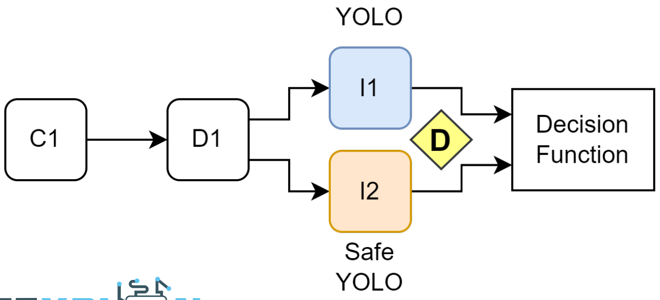
T2.3 DL Safety Architectural design Patterns

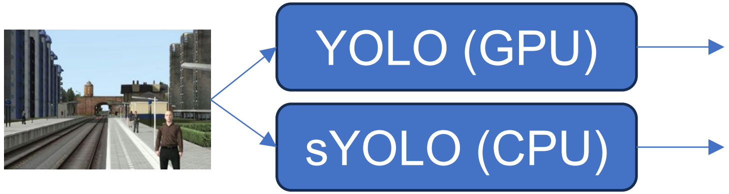
- L0 Diverse Redundancy – Inference platform diversity using diverse redundant frameworks (i.e., YOLO and SafeYOLO).

MISRA C:2012 Guidelines Summary - Violations by Rule





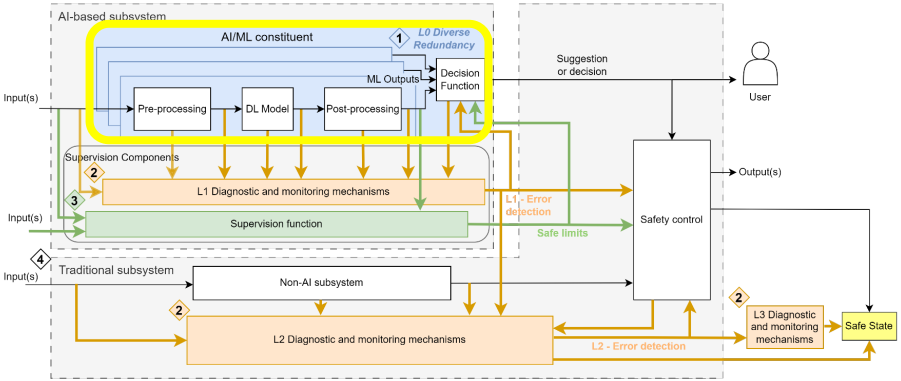


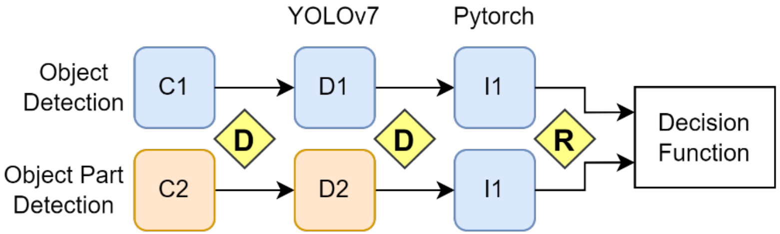


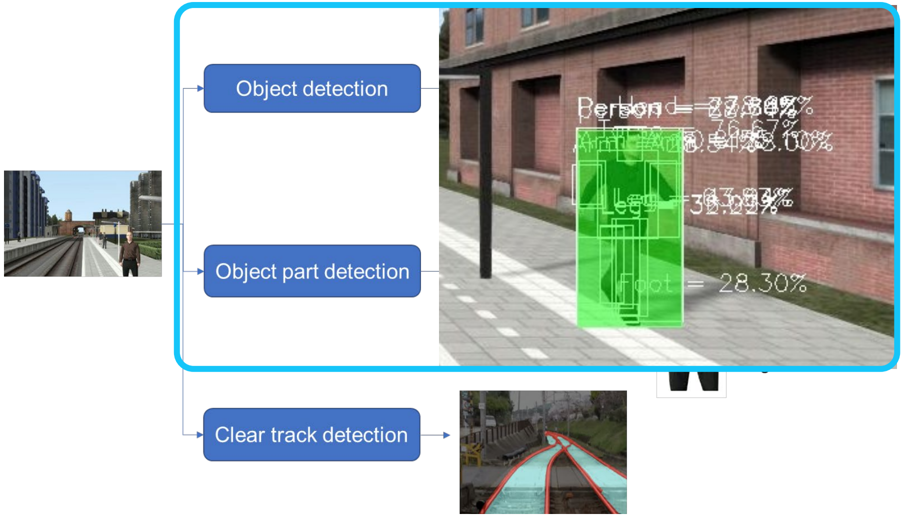
SAFEXPLAIN

T2.3 DL Safety Architectural design Patterns

- L0 Diverse Redundancy – Concept diversity using diverse concepts (i.e., Object Detection and Object Part Detection).

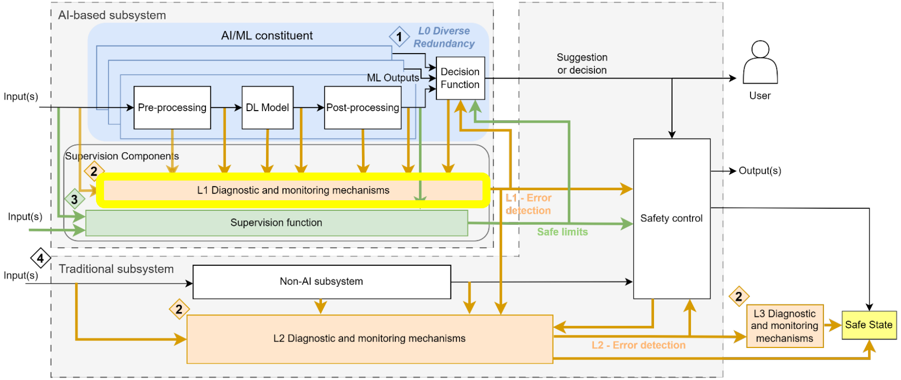






T2.3 DL Safety Architectural design Patterns

- Diagnostic and monitoring mechanisms – L1DM mechanisms – AI subsystem



1. Inputs: diagnostic mechanisms for input correctness, data quality, data redundancy, temporal consistency…
2. Model: diagnostic and monitoring mechanisms for execution errors, timing, program sequence, neuron activation patterns…
3. Outputs: diagnostic mechanisms for outputs, plausibility checks, input-output correlation, temporal consistency …
4. Resource usage: monitoring mechanisms for resource usage (e.g., CPU/GPU usage, memory usage)

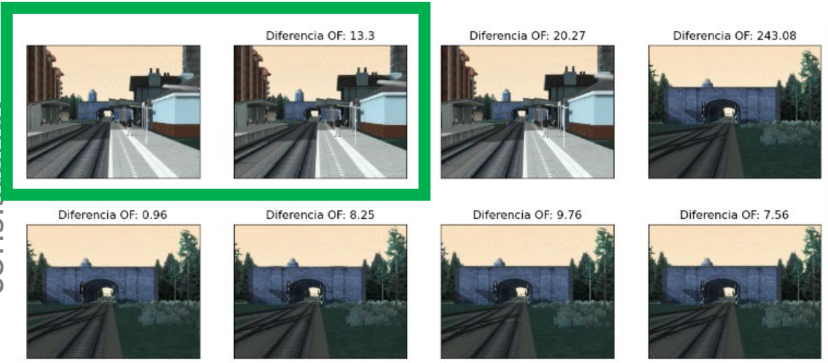
SAFEXPLAIN

T2.3 DL Safety Architectural design Patterns

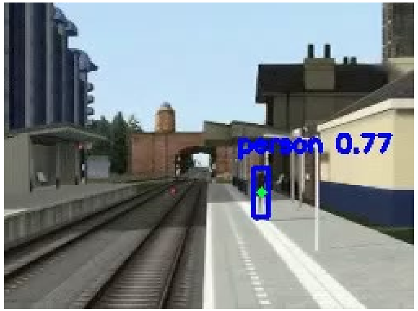
- Diagnostic and monitoring mechanisms – L1DM mechanisms – AI subsystem

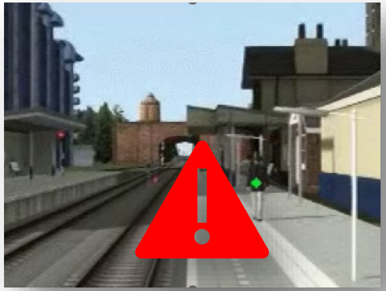
Diff 13.3 Diff 20.27 Diff 243.08

Input temporal consistency

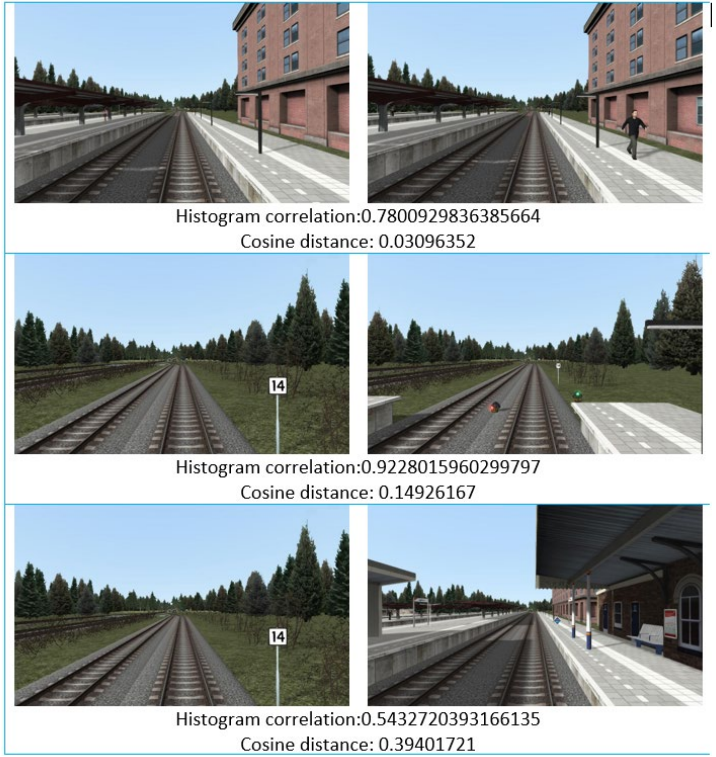


Output trajectory prediction



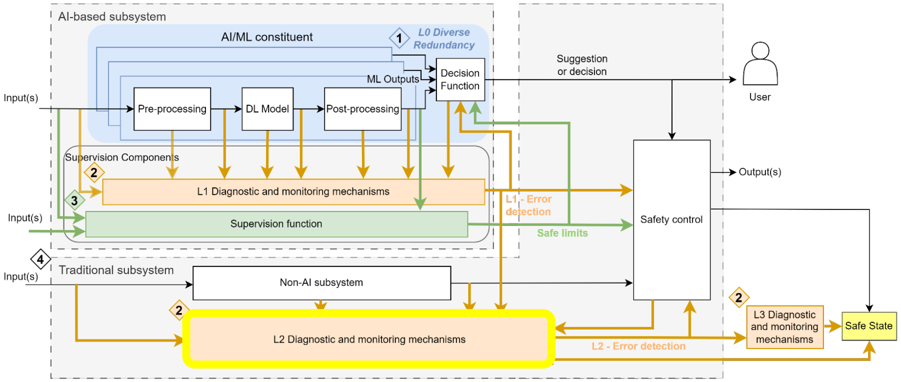


Redundant input consistency



T2.3 DL Safety Architectural design Patterns

- Diagnostic and monitoring mechanisms – L2DM mechanisms – Traditional subsystem



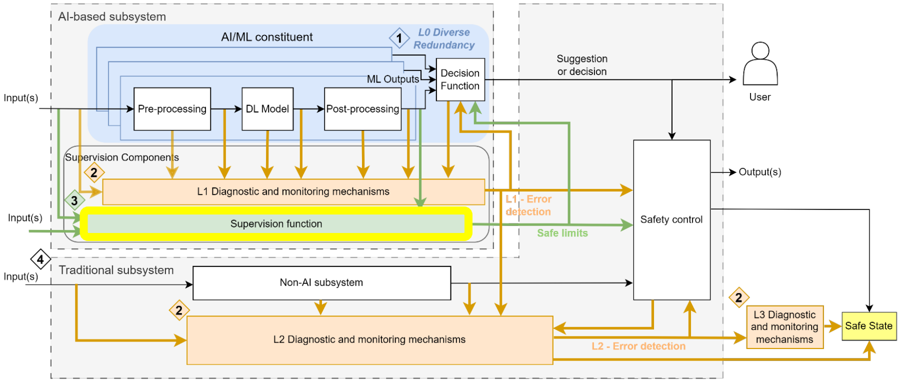
WP4 → HPC Platform monitoring approaches

1. Traditional functional safety diagnostics
2. Advanced diagnostic approaches for high-performance platforms
   - L2DM Configuration check – check that the configuration defined and verified during system development is kept at runtime.
   - L2DM Interference mitigation and control – check that all critical platform SW components meet with their deadline and if this is the case, it refreshes an external watchdog (L3).
   - L2DM Health Management – is the responsible of triggering the required reaction whenever L0, L1 or L2 diagnostics and monitoring detects an error

SAFEXPLAIN

T2.3 DL Safety Architectural design Patterns

- Supervision function

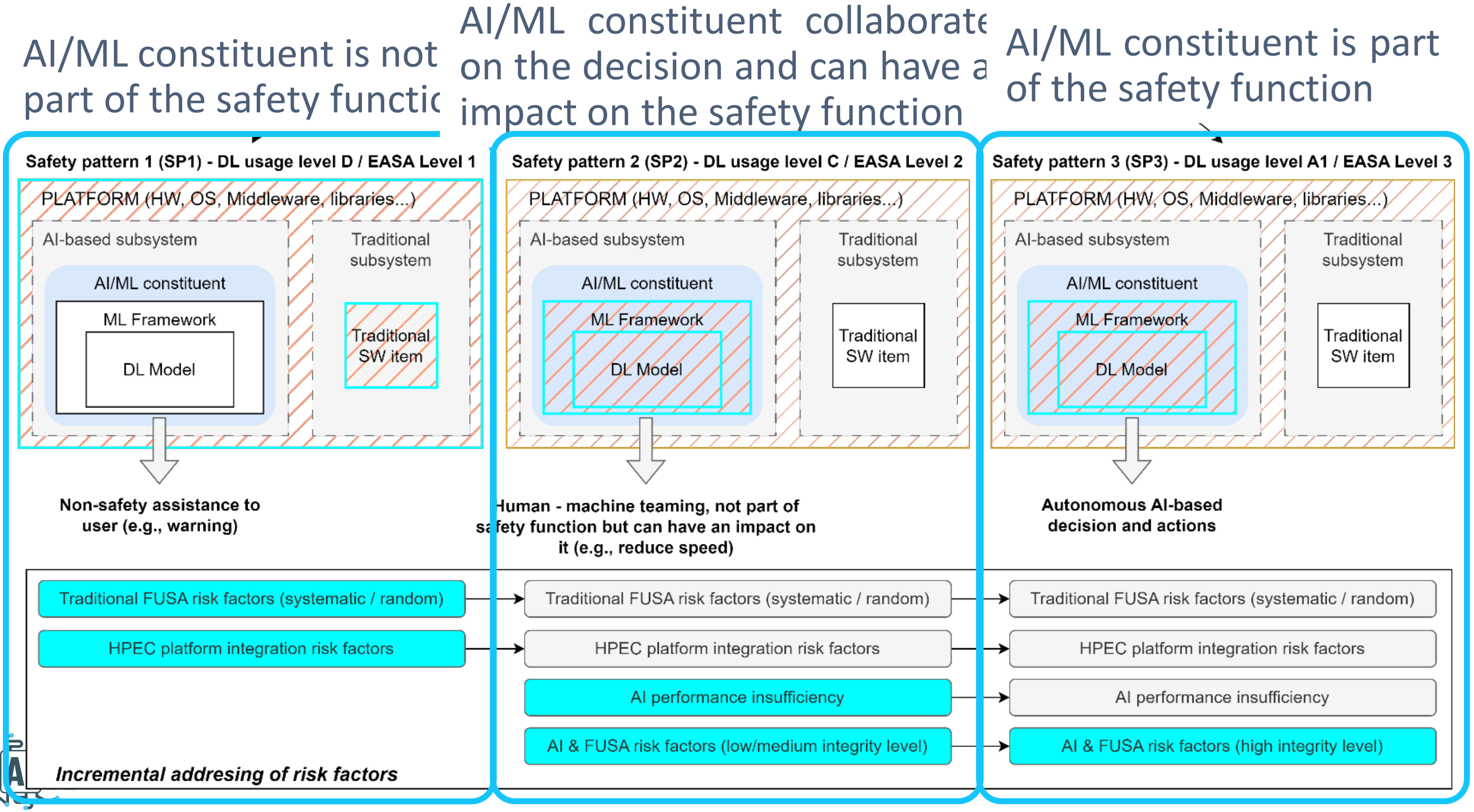


1. Check the appropriateness of the environment
2. Supervise the output of the ML constituent to identify unsafe situations
3. Stablish the limits for safe operation, providing a safe envelope
4. Provide explanations on the DL model

WP3 → Detailed techniques

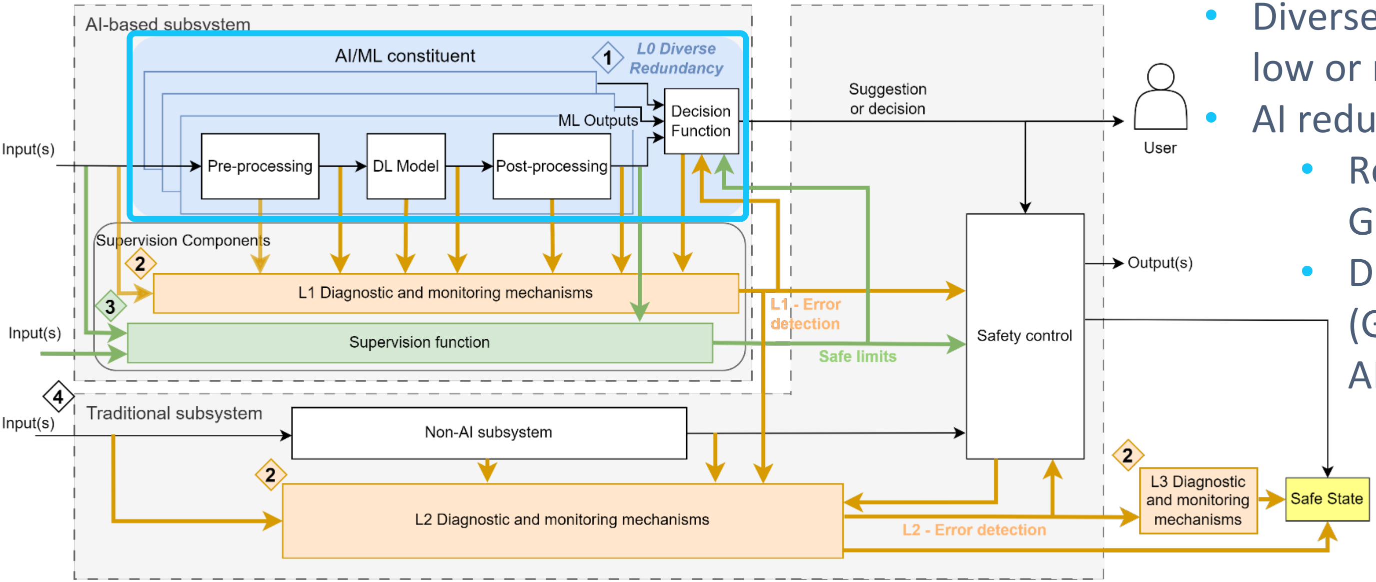
T2.4 Safe DL execution platform

- Incremental strategy for AI adoption in safety critical systems



T2.4 Safe DL execution platform

- Safety Pattern 2 (SP2) – L0 Diverse redundancy

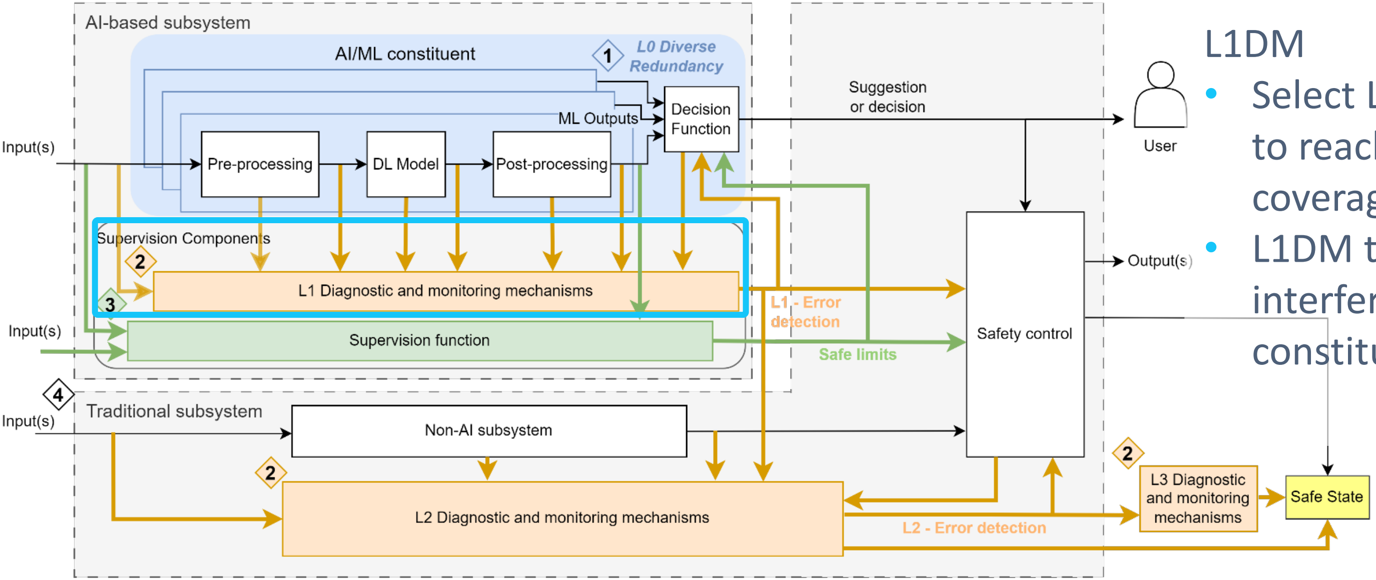


- Diverse redundancy schemes with low or medium DC
- AI redundancy
  - Redundant instances on the GPU
  - Diverse processing resources (GPU and CPU, GPU and other AI accelerators)

SAFEXPLAIN

T2.4 Safe DL execution platform

- Safety Pattern 2 (SP2) – Diagnostic and monitoring mechanisms



L1DM

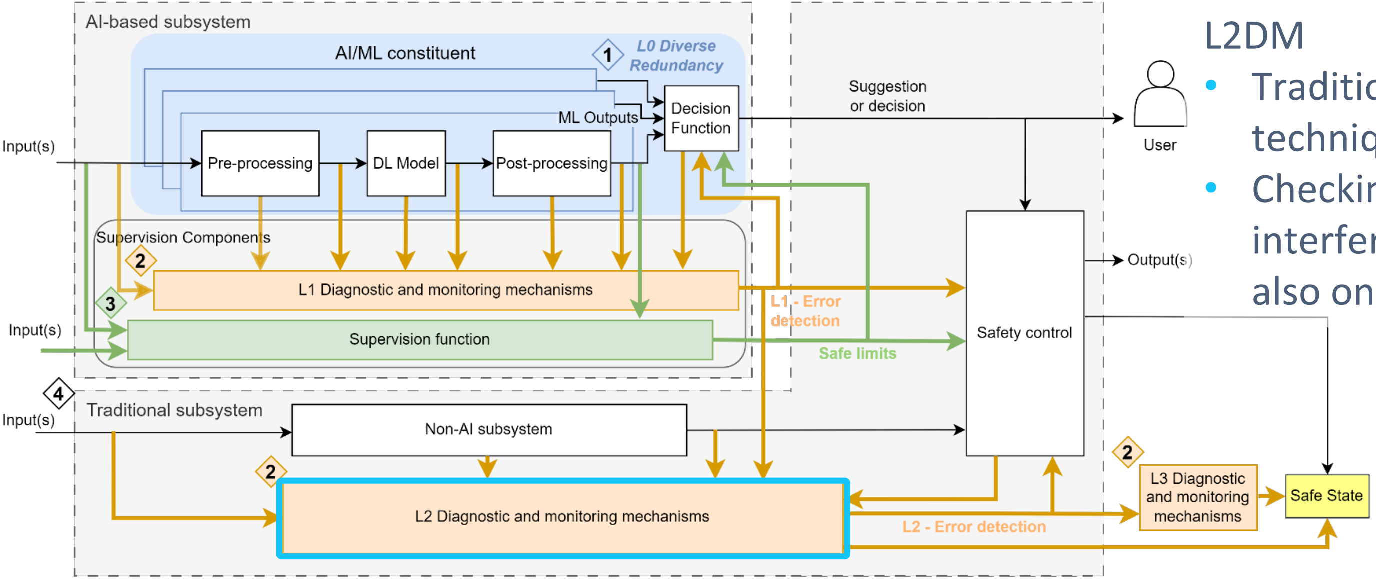
- Select L1DM diagnostics mechanisms to reach the required diagnostic coverage
- L1DM to guarantee freedom from interference, also in the AI/ML constituent

WP4 → specific events that can be monitored

SAFEXPLAIN

T2.4 Safe DL execution platform

- Safety Pattern 2 (SP2) – Diagnostic and monitoring mechanisms



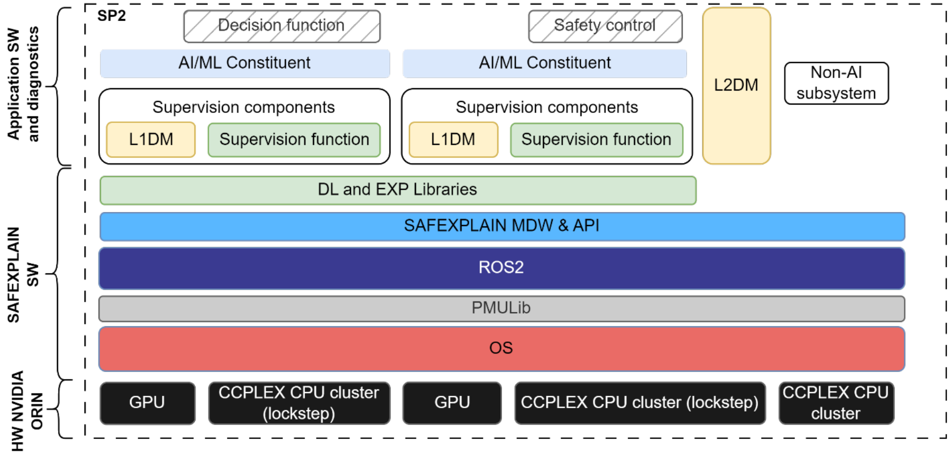
L2DM

- Traditional functional safety techniques and measures
- Checking system configuration for interference mitigation and control, also on the AI/ML constituent

SAFEXPLAIN

T2.4 Safe DL execution platform

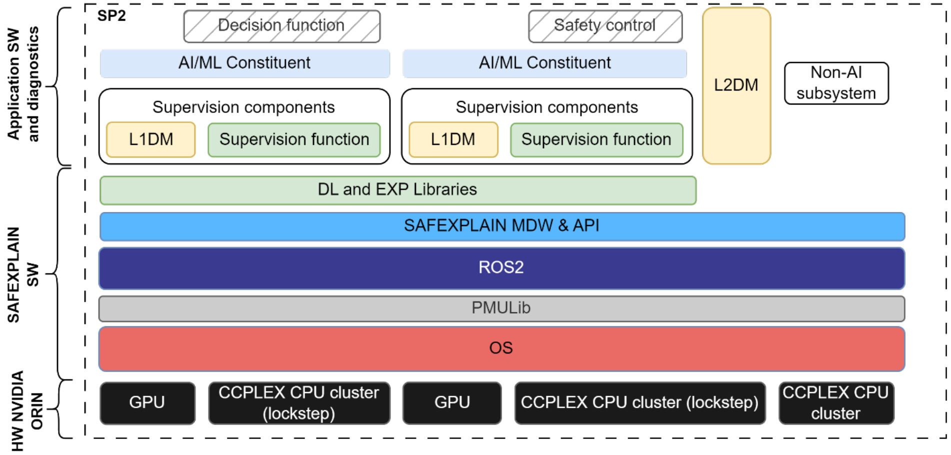
- SP2 to NVIDIA Orin resource allocation and configuration option



SP2 Element	Safety / non-safety	SP2 - A NVIDIA Orin resources and configuration
AI/ML constituent	AI based safety SW	Two instances, each in one separate CCPLEX CPU Cluster (Cortex A78) in lockstep configuration GPU for AI inference (depending on the DRS CPU or other computing resources could also be used to improve diversity) Memory controller fabric and traffic from CPU cluster to GPU MMUs for spatial independence SAFEXPLAIN SW Stack
Supervision components	Traditional or AI based safety SW	Each AI/ML constituent has each own L1DM and optionally each own supervisor function (depends on user application). Depending on the implementation of the supervision component, it may need GPUs for improved performance (e.g., AI based supervision function). The supervision components can share same CCPLEX CPU Cluster (Cortex A78) in lockstep configuration as the AI/ML constituent. MMUs for spatial independence SAFEXPLAIN SW Stack

T2.4 Safe DL execution platform

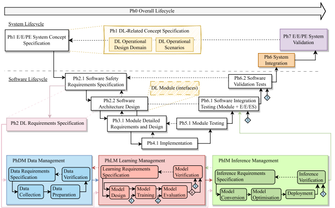
- SP2 to NVIDIA Orin resource allocation and configuration option



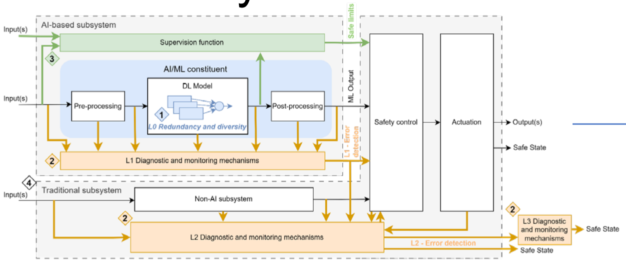
SP2 Element	Safety / non-safety	SP2 - A NVIDIA Orin resources and configuration
Decision function	Safety traditional SW	These SW components can run on any of the CCPLEX CPU Cluster (Cortex A78) in lockstep configuration used for the AI/ML constituent with the same configuration assuming they have the same integrity level.
Safety control	Safety traditional SW	
L2DM	Safety traditional SW	
Non-AI subsystem	Non-safety traditional SW	CCPLEX CPU Cluster (Cortex A78) or SPE (no need for lockstep configuration). MMUs for spatial independence L4 cache partitioning or disabled SAFEXPLAIN SW Stack or different OS on top of SPEs or hypervisor

Safety concept Assessment

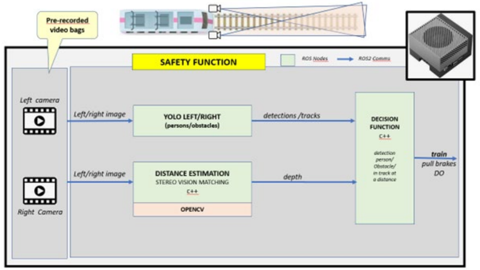
AI-FSM



Safety Patterns



Railway case study



Railway safety concept

- System description
- Safety Requirements
- Safety Architecture
- Safety techniques and measures
- Etc.

V&V methods for FuSa compliance

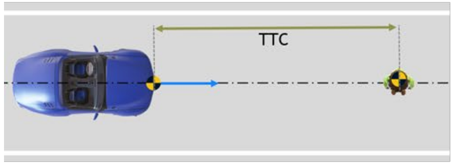
AI-based safety-related systems V&V Strategy

- Within the SAFEXPLAIN project, Exida contributed to the definition of a strategy for the right-hand side of the V-model for AI-based safety-related systems, with the goal to provide a structured approach to verification, validation and testing of Deep Learning (DL) software
- The strategy is structured on two levels:
  - Vehicle-level approach, based on a scenario catalogue and the corresponding test case specification.
  - Component-level approach, based on the identification of triggering condition and the corresponding test case specification.
- The V&V Strategies are applied in 3 key domains:
  - The automotive use case, led by NAVINFO, it is related to AEB (Autonomous Emergency Braking) implemented in a car vehicle.
  - The railway use case, led by IKERLAN, it is related to ATO (Autonomous Train Operation) implemented in a train (GoA 2).
  - The aerospace use case, led by AIKO, it is related to GNC (Guidance, Navigation and Control) implemented in a space vehicle.

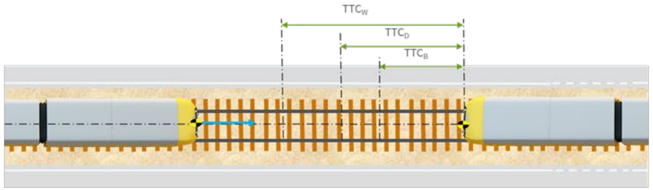
Vehicle-level V&V strategy

- The vehicle-level V&V activity begins with the creation of a scenario catalogue created, defined according to the applicable ODD.
- The purpose of the scenario catalogue is to collect all the hazardous operational scenarios that could impact vehicle safety and led to a hazardous situation.
- Once the scenarios are defined, they must be validated by means of test case specification.

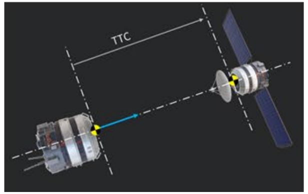
Automotive



Railway



Aerospace

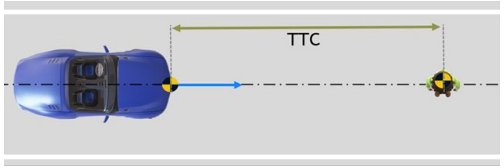


SAFEXPLAIN

Vehicle-level V&V strategy: Description of the Scenario/Constraints

- Driving situation description:
  - When the distance with vulnerable users (adult, child) decreases so that the driver or vulnerable users are in dangerous zone (possible collision) the intended functionality shall warn the driver and, if no driver reaction occurs and the collision is imminent, shall decelerate the vehicle.
  - The probability of Exposure (duration) of these scenario conditions is E4, considering the following combinations:
  - Driving in a city– E4 (>10 % of average operating time)
    - E.g., 10% of 8000h = 800 h
  - Persons within danger zone (ca. 1 vehicle lenght in front of vehicle) – E3 (1% to 10% of average operating time)
    - E.g., from 80 h to 800 h

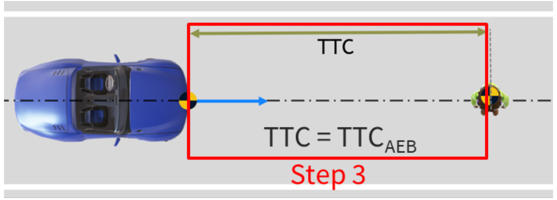
- Scenario Conditions/Constraints:
  - The Ego vehicle speed range is [5 km/h, 50 km/h]
  - The pedestrian crosses the road at 5 km/h (± 0,1 km/h)
  - The following environmental conditions shall be present:
    - Dry and daylight with minimum 1000 lux and Sun angle >15° to horizon
    - …
  - The following Pre-conditions shall be respected:
    - Ego vehicle shall keep steady speed and path
    - No override condition shall be present



Car towards a pedestrian

V&V strategy at vehicle level – Example: Test Specification – Steps Description

- Test specification steps description:
  - (Step 1) The ego vehicle is approaching the vulnerable users (adult, child).
  - (Step 2) When the distance, between the ego vehicle and the VRUs, is equal to the **Time To Warning (TTW),** the intended functionality shall evaluate the VRUs as collision relevant and provide at least 0,8 s before the start of the emergency braking the visual and audible warning to the driver (UN Regulation N° 152 clause 5.2.1.1, 5.5.1).
  - (Step 3) When the distance, between the ego vehicle and VRUs, is equal to the **Time To Collision AEB (TTC AEB),** the intended functionality shall ,if no driver reaction occurs, shall decelerate the vehicle providing at least 5.0 m/s2 (UN Regulation N° 152 clause 5.2.1.2).

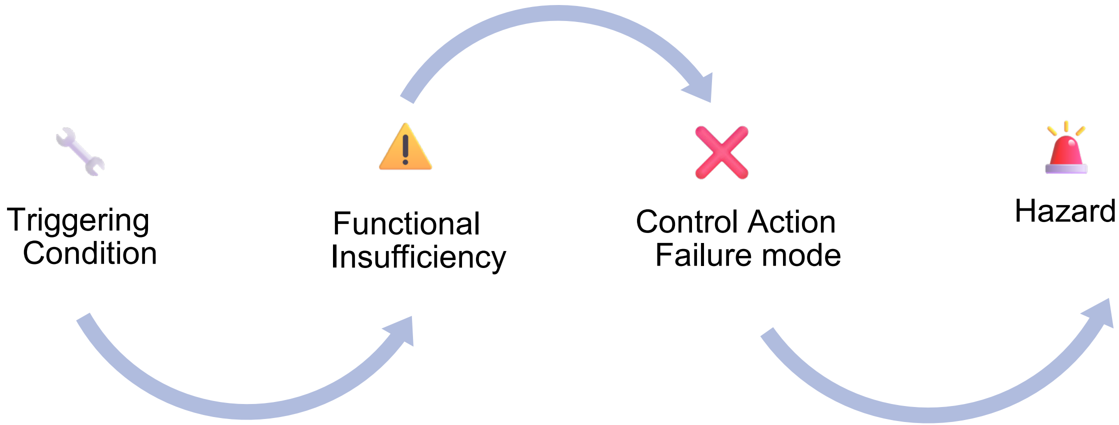


- Test specification (Step 3):
  - Ego vehicle status:
    - KL 15 = On;
    - …
    - Intended functionality state: Active Intervinging
  - …
  - Distance between Ego vehicle position and target vehicle
    - TTC == TTC AEB
  - Expected result:
    - Warning = Present and provided 0,8 s before emergency braking trigger
    - Braking = Present with deceleration equal to 5.0 m/s^2

SAFEXPLAIN

Component-level V&V Strategy

- The Component V&V activity begins with the identification of Triggering conditions based on system functionalities description and design by using STPA analysis methodology.
- The purpose is to collect all the component and functions malfunctions that could create a functional insufficiencies and led to a hazardous situation.
- Once the Triggering conditions are identified, they must be validated by means of test case specification.



Component-level V&V Strategy: Railway use case triggering condition

- Triggering conditions:
  - One side of the stereo camera feed is missing (e.g. left or right image not available for a number of frames), making depth estimation impossible.
  - …
  - Missing detection of sensor decrease in performance (e.g. damaged lens, aeging effects, glare conditions, blockage condition of the FOV, extreme weather conditions, wrong camera position/camera calibration...).
- Functional Insufficiency:
  - The ATO does not provide the requested braking intervention.
- Control Action Failure Mode:
  - Service braking is not performed when a collision is imminent.
- Hazard:
  - The ATO doesn't provide service braking when needed.

Component-level V&V Strategy: Railway use case test case specification

- Railway Triggering Condition Test Specification & Test Matrix:

Triggering Conditions ID	Triggering Conditions	Test ID	Test Description	Test Preconditions	Test Operating Elements	Test Steps (Operating)	Acceptance criteria
TC-0055	One side of the stereo camera feed is missing (e.g. left or right image not available for a number of frames), making depth estimation impossible	TestCase_17	The test verifies that the ATO detects missing stereo image input and transitions to safe state if the condition persists	• Kl.15 = off • No warning message available • Camera lens is damaged • Intended functionality state: active	• Camera Sensors • CAN Bus Simulator • ATO ECU	Step 1: • Start ATO system	• No warning message provided • Intended functionality state: boot state
						Step 2: • Simulate stereo camera feed where only one image is received (e.g. right image missing)	-
						Step 3: • Monitor image reception Count number of incomplete stereo frames (e.g. missing one side) • Threshold: e.g. ≥5 incomplete stereo frames in last 10	-
						Step 4: • Check whether the threshold exceeds	• ATO is suppressed within [x] ms • Intended functionality state: deactivated
TC-0054	Missing detection of sensor decrease in performance (e.g. damaged lens, aeging effects, glare conditions, blockage condition of the FOV, extreme weather conditions, wrong camera position/camera calibration...)	TestCase_05	The test verifies whether the ATO system correctly detects repeated black frames (indicative of damage, obstruction, or misalignment) and transitions to safe state. NOTE: we may need to define some terms, like: • Black Frame Definition: a frame where average luminance (Y channel) < threshold (e.g. <10). • Temporal Condition: configurable number of black frames in last N frames (e.g. 5 of 10)	• Kl.15 = off • No warning message available • Camera field of view is covered • Intended functionality state: active	• Camera Sensors • CAN Bus Simulator • ATO ECU	Step 1: • ATO system is started	• No warning message provided • Intended functionality state: boot state
						Step 2: • Monitor image reception Count number of black frames • Check the number of the acquired black frames	• More than 5 of 10 frame received, are black frame • Safe state is triggered

Thanks

SAFEXPLAIN

Safe and Explainable
Critical Embedded Systems based on AI

Follow us on social media:

www.safexplain.eu

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe programme under grant agreement number 101069595.