SAFEXPLAIN

Making certifiable AI a reality for critical systems: Core Demo

safexplainproject@bsc.es

Learn more

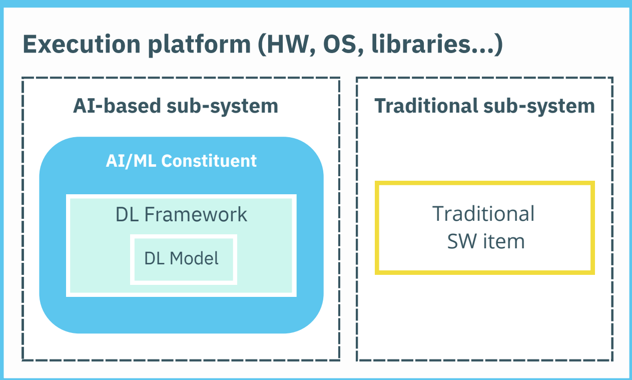
As fully autonomous systems become more and more crucial for advanced functions, AI takes on an increasingly more dominant role in critical systems. However, current AI software is developed as a black-box that is hard to verify, lacks clear requirements, and clashes with the rigorous, certifiable and explainable processes required by safety-critical systems.

The SAFEXPLAIN project is working to close this gap through its next-generation open software platform designed to show how AI can meet the stringent requirements of functional safety in autonomous technologies. The SAFEXPLAIN Core Demo is a direct, small-scale concretization of our end-to-end approach to making certifiable AI a reality for critical systems by ensuring that systems are safe by construction and follow the latest relevant standards.

This small scale, fully functional and configurable teaser shows how SAFEXPLAIN technology can accommodate scenarios with critical functionalities in three selected 'toy' examples from the automotive, rail and space domains. SAFEXPLAIN offers a pivotal concept in platform-level support. Its middleware works to ensure compliance with safety patterns by design while preserving modularity.

The SAFEXPLAIN technology deployed in this demo is beneficial for developers and decision-makers focused on transport and mobility in Critical Autonomous AI-Based Systems (CAIS) who need evidence of what type of evaluation and support will be available in the near future regarding the certifiability of their CAIS products.

FuSa Compliant Safety Pattern 2



Human-machine interaction only potentially affects a safety function

- Traditional FuSa risk factors (systematic/random)
- HP MPSoCs platform integration risk factors
- AI performance insufficiency
- AI FuSa risk factors (LOW/MED integrity levels)

The core demo provides a generic skeleton that accomodates simple, functionally relevant examples. It focuses on ‘Safety Pattern 2’ where an AI/ML constitutent partially affects the decision-process. Figure 1 visualizes this relationship.

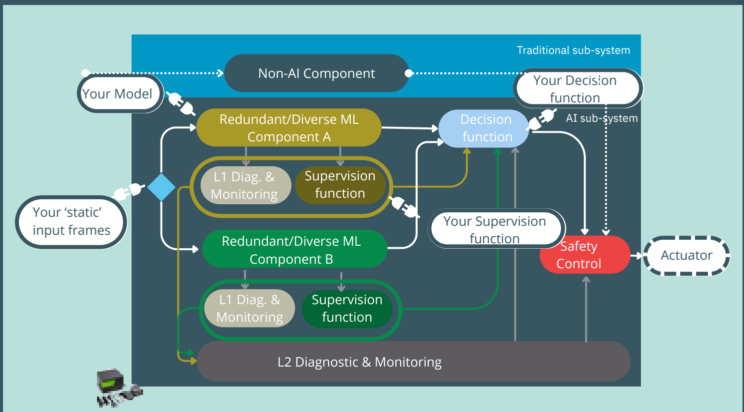


Figure 1: Core demo skeleton

Plug-in ‘functional’ code

- *ML nodes*
- *Supervisor*
- *Decision function*
- *Control logic*

This demo is available for early users. It provides a preview of what is to come with the car, rail, and space full demos, equipped with completely developed user cases with ODD-based scenarios and test-suites. These full demos will be available in September 2025 as final results of the SAFEXPLAIN project.

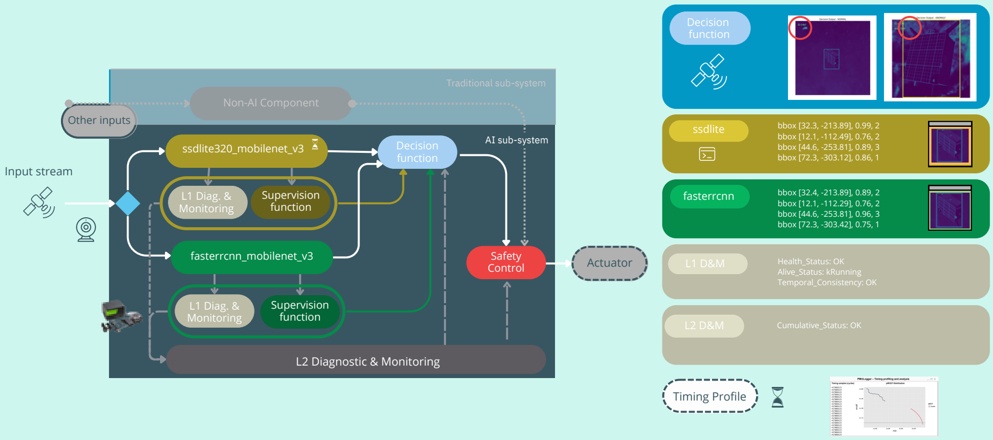
Join the final event

How it works: Core demo instantiated in 3 cases

SPACE DOMAIN

Goal: *Identify the target, estimate its pose, and monitor the agent position, to signal potential drifts, sensor faults, etc*

- AIKO open models + Input images
- L1 D&M Temporal consistency
- Supervision VAE
- Decision Ensemble

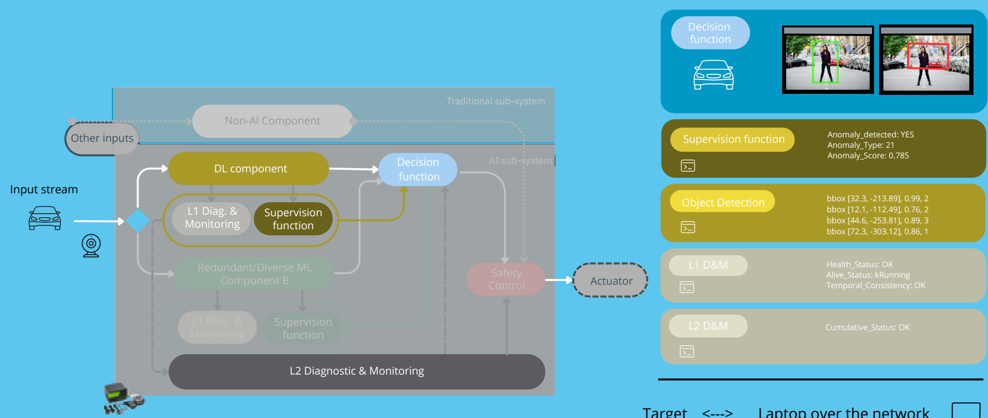


Target <---> Laptop over the network

AUTOMOTIVE DOMAIN

Goal: Validate the system's capacity to detect pedestrians, issue warnings, and perform emergency braking

- YOLO v11 pretrained model + frames from CARLA scenario
- L1 D&M Temporal consistency
- Supervision Function from EXPLib (anomaly detectors)
- Decision Function (mainly visualization oriented)

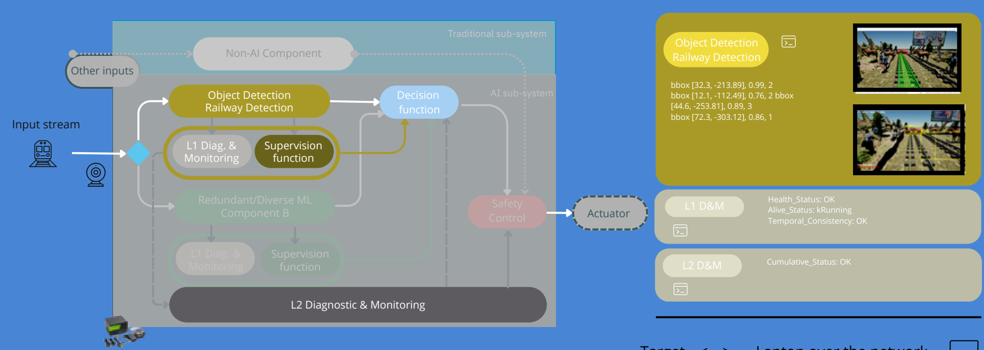


Target <---> Laptop over the network

RAIL DOMAIN

Goal: Validate the system's capacity to detect cars, issue warnings, and perform service braking

- YoloV8 pretrained model + frames from Unreal Engine
- L1 D&M Temporal consistency Bounding boxes for detected obstacles
- Data-quality and temporal-consistency scores



Target <---> Laptop over the network

Funded by the European Union

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