Applying reference safety architecture patterns to decentralized edge AI systems

SAFEXPLAIN Final Event, Barcelona – 23.09.2025

Francisco J. Cazorla (Barcelona Supercomputing Center)

Agenda

1. Introduction to EdgeAI-Trust
2. Decentralized edge AI systems
3. Applying reference safety architecture patterns

EdgeAI-Trust

Introduction to EdgeAI-Trust

EdgeAI-Trust “Decentralized Edge Intelligence: Advancing Trust, Safety, and Sustainability in Europe” project has received funding from Chips Joint Undertaking (Chips JU) under grant agreement No 101139892. The Chips JU receives support from the European Union’s Horizon Europe research and innovation program and national authorities.

Introduction

- Need
  - **Rapid** evolution of **AI**,**ML**, and **edge** technologies
  - Need for improved infrastructure, tools, and solutions, particularly for **large-scale AI applications that require enhanced edge learning**
- EdgeAI-Trust
  - aims to develop a **domain-independent AI architecture**, along with **hardware and software solutions**, that will **support** these **AI advancements**.
  - Seeks to **increase acceptance** and **trust** in AI solutions, benefiting multiple sectors.

IF YOU ARE INTERESTED, PASS BY OUR POSTER

Enhancing Trust And Functionality In Edge AI Systems

The EdgeAI-Trust project aims to leverage cutting-edge technologies to tackle challenges associated with ensuring the trustworthy and real-time orchestration of critical applications using decentralized edge AI within the Functional Safety Continuum. The project also seeks to validate the effectiveness of these technologies through real-world use cases.

Develop a trustworthy domain-independent AI architecture

The architecture will support the development of collaborative AI-based systems with orchestration, upgradeability, manageability, reliability, real time, safety, security, and energy efficiency. The architecture will support a continuum of heterogeneous AI-based algorithms and devices along standard APIs for interoperability and trusted exchange. This will range from sensor-actuation, large scale device-connected systems, edge processing units to the cloud. While the architecture is domain-independent, it will enable domain-specific instantiations for trustworthy collaborative AI systems.

Create reliable and collaborative large-scale edge AI solutions

Develop the next generation decentralized HW/SW edge AI technologies, with a particular focus on safety-critical and security related systems. The technologies will support fully collaborative AI by allowing heterogeneous devices to learn, adapt at the edge to cognitive reasoning tasks. Decentralisation of resources and processes among different entities/devices enable dynamic reconfiguration of processes in a resource constrained environment.

Increase trustworthiness of edge AI solutions

Considering the upcoming EU AI Act, here the focus is on explainability, reliability, safety, security and robustness of edge AI solutions. We will achieve that by focusing on (i) rigorous plausibility checks, (ii) monitoring of AI decisions and anomalies, (iii) explanation of situational awareness, and (iv) also include reliability and fault tolerance in a dynamic zero trust environment.

Optimization and validation of decentralized edge AI solutions

The goal is to provide methodology and tools that enable the optimization and validation of AI systems based on the EdgeAI-Trust architecture and solutions from the development platforms to the finished product. Automation supports developers by taking care of the constraints (like accuracy, latency, resource constraints and reliability) due to the limitations of the edge AI systems. By that, rapid improvement and deployment of Edge AI-based systems is provided.

Large-Scale Impact and Economic Leadership through EdgeAI EDEM (EdgeAI Ecosystem Monitoring) Platform

Achieve large-scale, sustainable impact through the EdgeAI EDEM (EdgeAI Ecosystem Monitoring) platform while delivering a complete market analysis and an exploitation mechanism for financial sustainability and economic leadership in the global marketplace.

PROJECT PARTNERS

PROJECT COORDINATOR:

tttech auto

CONSORTIUM GERMANY

tttech auto, ZF, Infineon, MULYTIC LABS, asvin, DLR, Universität Siegen, HOCHSCHULE HAMM-LIPPSTADT, HOCH SCHULE OFFEN BURG

CONSORTIUM AUSTRIA

tttech auto, TTTech, AVL, Infineon, virtual vehicle

CONSORTIUM FRANCE

life.augmented, zetta scale, VRANKEN POMMERY, TECHNEXT®, Université de Reims Champagne-Ardenne, HAWAI.tech HARDWARE AI

CONSORTIUM ITALY

life.augmented, I-FEVS, Avimecc, exida DEVELOPMENT, NEOTERA

CONSORTIUM SPAIN

Barcelona Supercomputing Center, tttech auto, technica Member of KPIT Group, SOLVER Intelligent Analytics, RAPITA SYSTEMS A RapitaLaw COMPANY, FENTISS, Clue, Universitat Politècnica de València

CONSORTIUM DENMARK

NVIDIA, TEGNOLOGY

CONSORTIUM BELGIUM

IVEX, S

CONSORTIUM TURKEY

TOFAŞ, AVL, SMART UNIVERSAL, büyütech

CONSORTIUM GREECE

MediSys, Ελληνικό Μεσογειακό Πανεπιστήμιο

CONSORTIUM LATVIA

SMARTSOL

CONSORTIUM CYPRUS

velti AI, AI-powered Mobile Marketing & Gaming, itml innovation applied

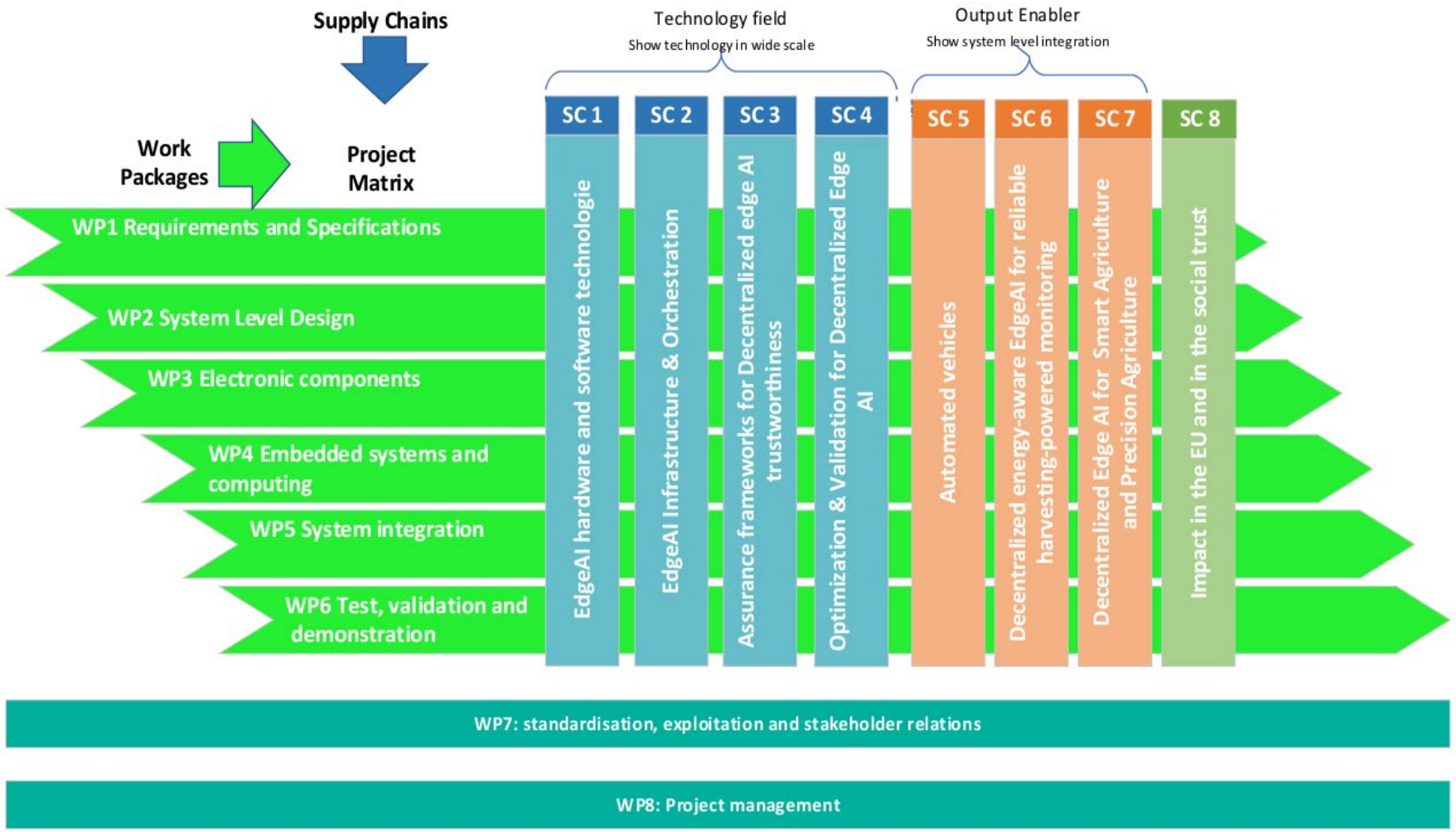
CONSORTIUM SLOVENIA

Jožef Stefan Institute

CONSORTIUM SWITZERLAND

zhaw School of Engineering, InES Institute of Embedded Systems, UE Uyan Electronics, miromico

WPs and SCs



EdgeAI-Trust matrix structure of WPs and SCs.

EdgeAI-Trust

Decentralized edge AI systems

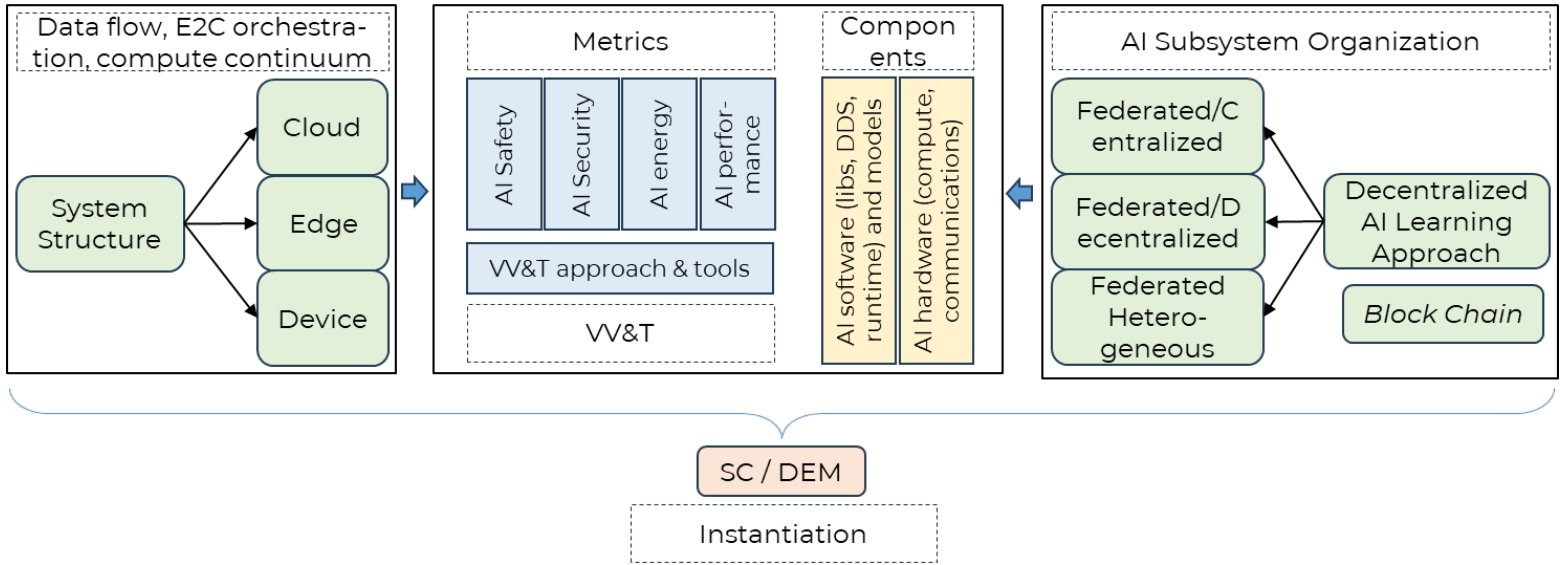
EdgeAI-Trust “Decentralized Edge Intelligence: Advancing Trust, Safety, and Sustainability in Europe” project has received funding from Chips Joint Undertaking (Chips JU) under grant agreement No 101139892. The Chips JU receives support from the European Union’s Horizon Europe research and innovation program and national authorities.

Decentralized edge AI systems

- Design a **comprehensive, modular AI architecture** capable of **operating seamlessly on edge devices across multiple domains** and applications
  - The resulting **decentralised edge AI architecture** will be domain-independent, providing **scalability**, **efficiency** and **flexibility** for different applications and environments
- Develop **AI architectures** that **incorporate** features such as **explainability**, **interpretability**, **traceability** and the use of **AI ensembles**.
  - **Minimise the risk of failure** during construction, **improve observability and controllability**, and **provide support for safety measures** to enable the construction of safety cases
- Subtasks.
  - Research & Analysis: **Investigate current edgeAI architectures** to understand **strengths**, **weaknesses** and areas for improvement. Studying **domain-specific edgeAI solutions** to identify common **patterns** and **requirements**.
  - Modularity Design: **Create an AI architecture design** that emphasises **modularity**, allowing for **easy integration** and **modification** of components based on specific domain needs.
  - Efficiency & Scalability: **Optimise the architecture** for efficient resource utilisation, ensuring **scalability** as more edge devices or more complex AI models are added.
  - Interoperability: **Ensure** that the designed architecture can **communicate** and **operate** with different software, hardware and network protocols, enabling a wide range of applications.

Proposed Architecture

- The proposed modular AI architecture covers 3 main areas:
  - **System Structure** (left) components cover the **compute continuum** determining the computations performed in the cloud, edge, and device. It defines **data flow** and the **edge2cloud orchestration**.
  - **AI subsystem** organization (right) components cover the **AI subsystem organization** with focus on the **decentralized AI approach** and the usage of blockchain technologies.
  - **Technological system** (center top) covers
    - the metrics to assess the proposed solution like AI safety, security, energy and performance;
    - the V&V and testing (VV&T) solutions to support the fulfillment of requirements sent on those metrics.
    - the software and hardware components to implement the solution.
- The **instantiation** (center bottom) of the proposed architecture requires realizing these areas as described later in this document for the different DEM.

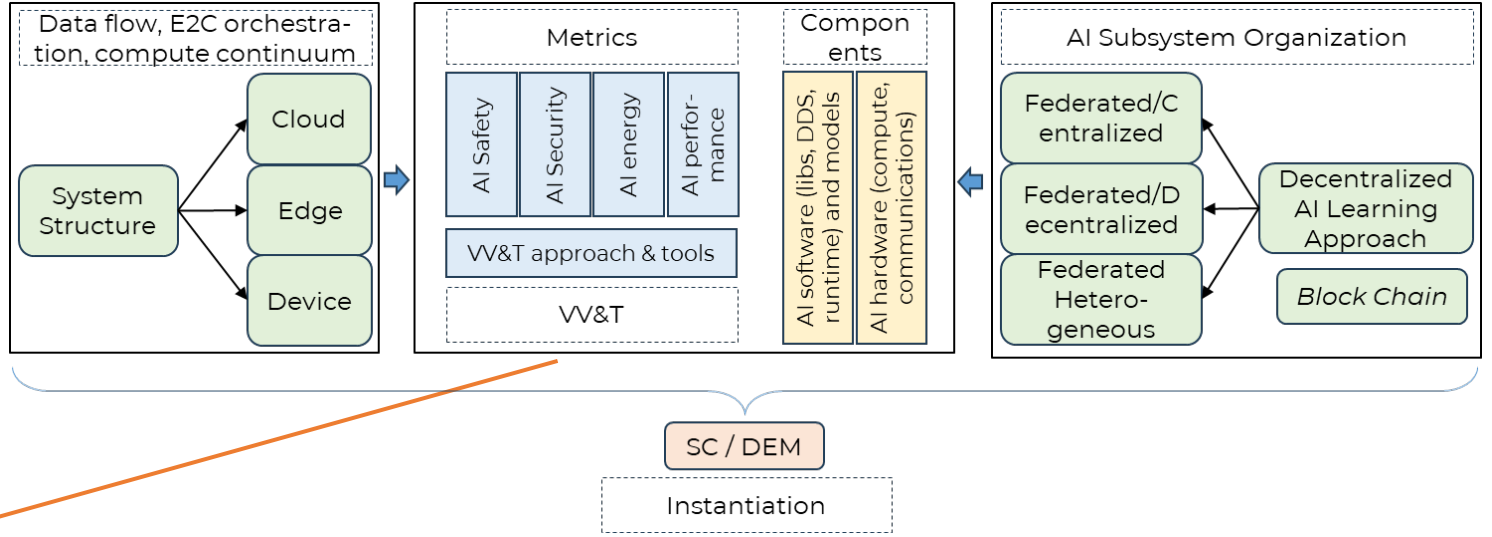


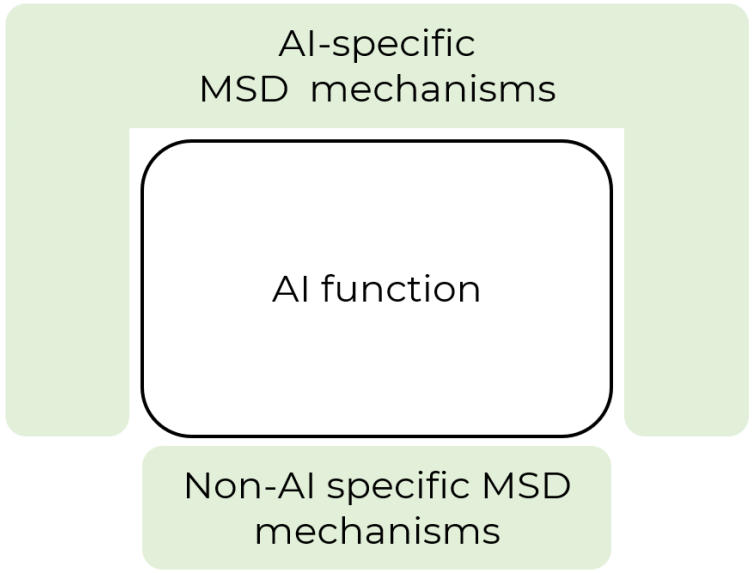
EdgeAI-Trust

Applying reference safety architecture patterns

EdgeAI-Trust “Decentralized Edge Intelligence: Advancing Trust, Safety, and Sustainability in Europe” project has received funding from Chips Joint Undertaking (Chips JU) under grant agreement No 101139892. The Chips JU receives support from the European Union’s Horizon Europe research and innovation program and national authorities.

Technological System





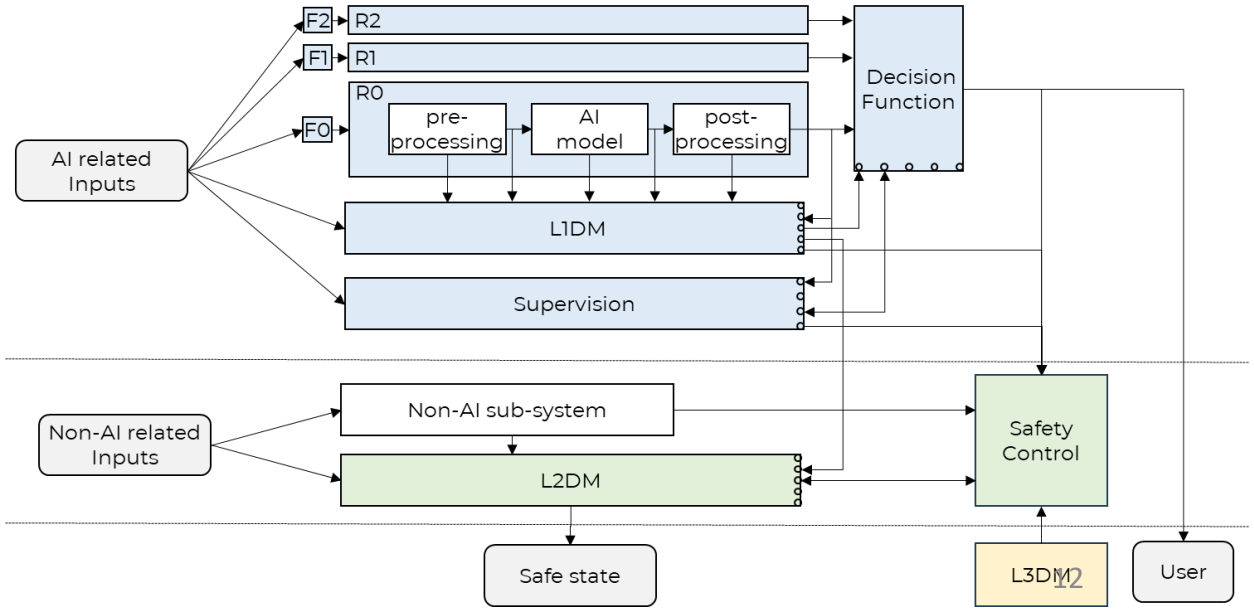
MSD = Monitoring, Supervision, and Decision

- For the **realization of the technological components of our decentralized AI architecture**, **we build on an approach created for centralized AI and safety** that was developed as part of the **SAFEXPLAIN** project [AAY24].
- We **extend** it to **cover decentralized AI** and **several nonfunctional metrics**.

[AAY24] Jaume Abella (BSC), Irune Agirre (IKR), Irune Yarza (IKR). SAFEXPLAIN Project D2.2 DL safety architectural patterns and platform. 2024 [link]

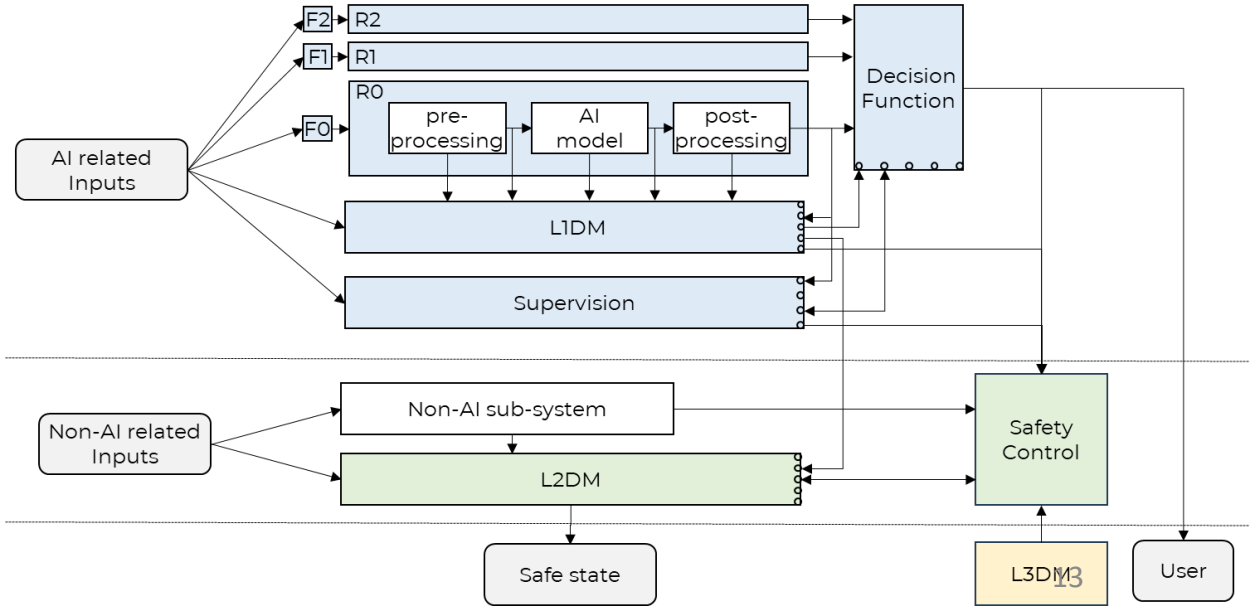
Approach \1

- **AI-related** (blue) **subsystem**:
  - It encompasses all AI-related components.
  - Monitoring, Supervision, and Decision (MSD) components in this subsystem handle more AI-specific features and hence are naturally “closer” to the AI function
  - **Redundancy element** (RED):
    - Several redundant copies (R0, R1, and R2) of the same AI functionality to cover different NFR.
    - Redundant copies can given inputs after they are filtered (F0, F1, and F2) for instance to increase model robustness or to realize redundancy with lower cost than full replication.
  - **L1 Diagnosis & Monitoring element** (L1DM)
    - Complements RED to detect (functioning) runtime errors in any NFR or model insufficiencies w.r.t. any metric.
  - **Supervision element**:
    - It focuses on input and outputs of the AI component.
  - **Decision function element**:
    - Combines the output of the several redundant copies of the AI component and information from the L1DM & supervision elements.



Approach \2

- **Non-AI-related** (green) **subsystem**:
  - It implements part of the functionality of the application that is non-AI related
  - It encompasses MSD components present in traditional (i.e. not necessarily AI) systems.
  - These MSD handle generic monitoring, supervision, and decision logic according to consolidated safety standards.
  - **L2DM**:
    - it collects information from the elements in the non-AI subsystem, i.e. the non-AI subsystem component; the safety control component; non-AI inputs; and the L1DM
    - It manages errors related to SW violating their deadlines, or providing invalid outputs (out of range)
  - **L3DM**:
    - Following traditional functional safety practices, it implements DM on an external device (watchdog)
  - **Safety Control**
    - Preserves safe operation of the overall system, e.g., orchestrating errors reported by the components and taking system level decisions



Approach \3

- **Data**
  - It sets no constraints on the means and formats of communication to exchange data.
- **Assurance level**
  - It is not bound to any particular safety or security assurance level
  - or any other metric that has its criticality stipulated in the applicable standards.
- **Mapping:**
  - The architecture itself sets no constraints on whether a component is mapped to the cloud/edge/device
  - In fact, the functionality of an MSD module can be split into several of these levels.
  - It is rather the timing requirements of the monitoring (control) loops the ones affecting this decision.
    - For instance, it is natural that the DM components in the AI-subsystem are placed in the same HW component where the AI-subsystem component is executed.
- **Metrics**
  - In every MSD module several metrics can be checked/assessed against specific values.
- **Complexity**
  - The complexity of each MSD module depends on the particular instantiation and its particular needs for the different functional metrics
    - On one extreme of the spectrum, the MSD module can be left either empty or include basic quality measures when the requirements on nonfunctional metrics are low.
    - On the other extreme, for those systems to be safety and security compliant according to specific standards, MSD modules will follow a compliant design.
      - In fact, it is worth noting that the proposed architecture emanates from the SAFEXPLAIN project that was designed to be compliant with ISO 26262 and ISO 21448 (aka SOTIF) standards.

EdgeAI-Trust

Thank you

EdgeAI-Trust “Decentralized Edge Intelligence: Advancing Trust, Safety, and Sustainability in Europe” project has received funding from Chips Joint Undertaking (Chips JU) under grant agreement No 101139892. The Chips JU receives support from the European Union’s Horizon Europe research and innovation program and national authorities.