Challenges to AI adoption in Safety Critical Systems

Coordinator perspectives on real-world barriers and breakthroughs

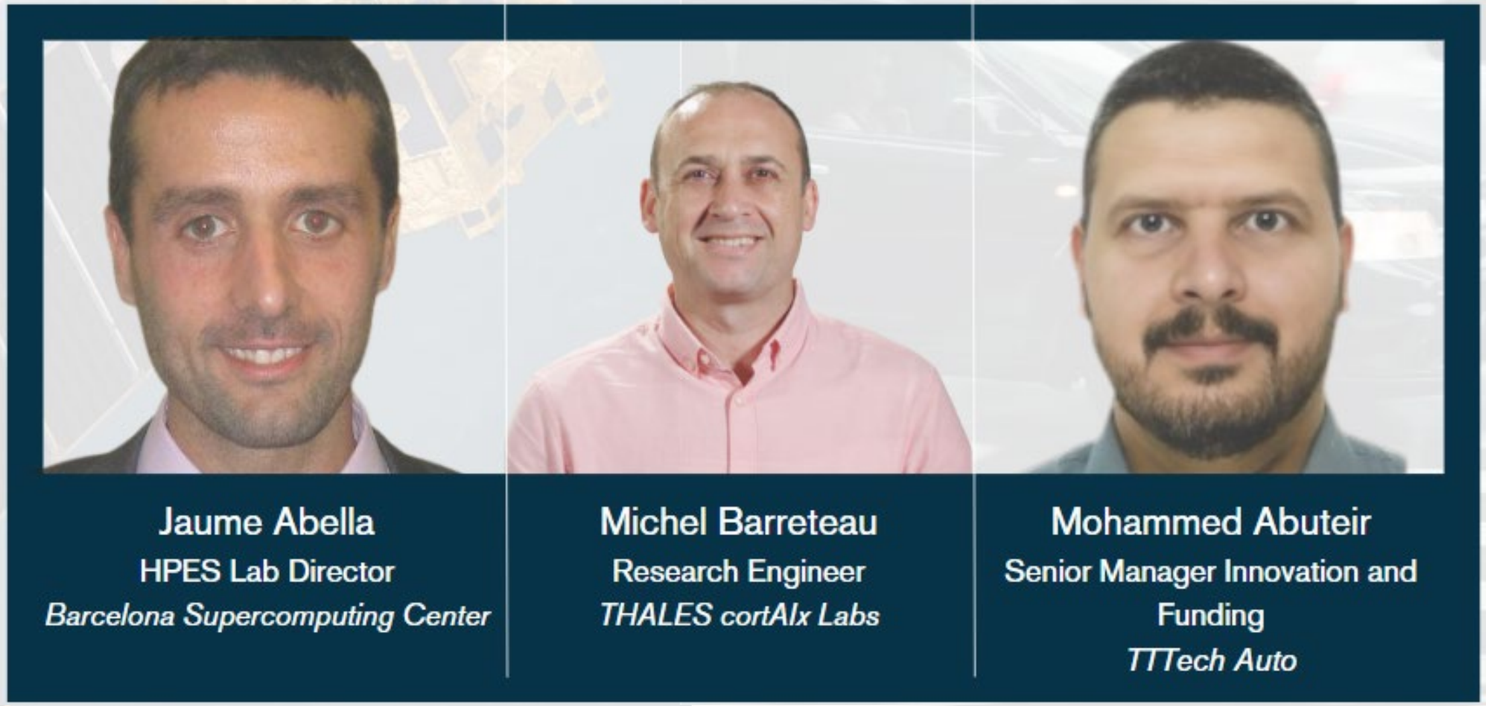
Jaume Abella (SAFEXPLAIN), Michel Barreteau (ULTIMATE), Mohammed Abuteir (EdgeAI-Trust)

Session moderated by Irune Agirre

Overview of the day

Time	Session	
9:45	Panel session: Challenges to AI adoption in Safety Critical Systems	
11:00	Coffee Break and Poster Session	
11:30	SAFEXPLAIN/ EdgeAI-Trust - Sala Master	ULTIMATE - Sala Agora
13:00	Lunch	
14:00	SAFEXPLAIN - Sala Master	Ultimate - Sala Agora
16:00	Coffee Break and Poster Session	
16:30	What's next for AI in safety-critical systems: Insights and the road ahead	

Project coordinators



Jaume Abella
HPES Lab Director
Barcelona Supercomputing Center

Michel Barreteau
Research Engineer
THALES cortAIx Labs

Mohammed Abuteir
Senior Manager Innovation and Funding
TTTech Auto

MODERATED BY
IRUNE AGIRRE (IKERLAN)

Setting the stage: Challenges of using AI in critical systems

- Failure or malfunction may result **severe harm** (e.g., casualties)
- Exhaustive **Verification and Validation** (V&V) process, and **safety measures** deployed to guarantee the safety goals are met
  - I.e. avoid errors if possible, and detect+correct them otherwise
- Each domain has its own guidelines and regulations for software and hardware

- ISO 26262, ISO 21448 (SOTIF) and ISO/PAS 8800 for automotive, DO178C and DO254 for avionics, EN5012x for railway, ECSS-Q-ST-80C for space

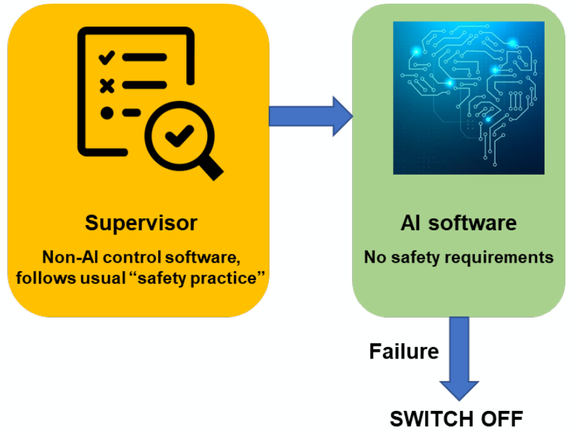
Setting the stage: Challenges of using AI in critical systems

- **AI techniques** are at the very heart of the realization of **advanced software functions** such as **computer vision for object detection** and tracking, path planning, driver-monitoring systems,…
  - E.g., You Only Look Once (YOLO) camera-based object detection system builds upon a Neural Network

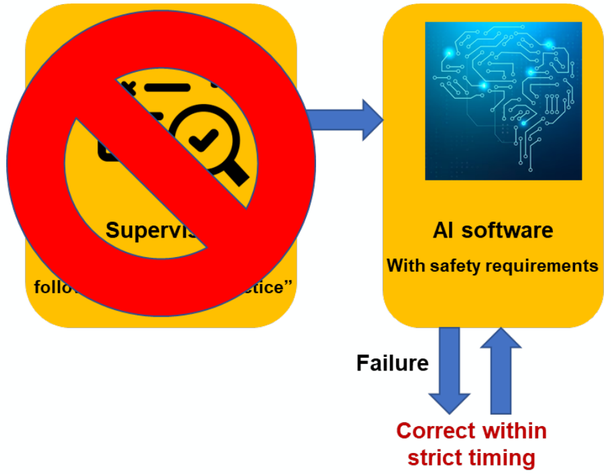
- Autonomous operation
  - epitome of safety-related applications of AI in Critical Embedded Systems
  - exemplifies the need for increasingly **high computing performance** whilst **making AI solutions to comply with Functional Safety requirements**

Setting the stage: Challenges of using AI in critical systems

- **AI used in fail-safe systems** (i.e. systems with a safe state)
  - E.g., Advanced Driving Assistance Systems (ADAS) can notify misbehavior and **transfer control to the driver**
- We can supervise the AI-based functionality and, **if it fails, switch it off**
  - or mitigate safety risks somehow (e.g., switching to a degraded mode)



- With **autonomous systems** (e.g., autonomous cars) this is **not yet solved**
  - We cannot simply switch AI parts off!!
  - E.g., car without steering wheel at 120km/h in the highway… **who's driving?**



Setting the stage: Challenges of using AI in critical systems

- Determine goals and specify **clear and unambiguous safety requirements** → **No "goals" or "requirements**. Just make it work **"as good as possible"**
- Propose a **design that meets all requirements** → Design based on **intuition/experience**
- **Decompose** into simpler pieces until getting software and hardware units → **No decomposition**, just an atomic piece of software
  - **Small enough** to be realized/implemented atomically → **Too large to be understood**
  - With specific interfaces and **correctness criteria** → Rather than correctness criteria, with **quality criteria**
  - Defining control and how data are used, but **not using data for definition** → **Data used for definition**. It sets all "parameters" holistically
- System becomes **"correct-by-construction"** → System **works well with some probability**
- Then we use **data to test**, as "extra" evidence → Test **data part of the empirical design loop** (optimize until it is "good enough")

SAFEXPLAIN
Safe and Explainable
Critical Embedded Systems based on AI

ULTIMATE

EdgeAI-Trust

FINAL EVENT

TRUSTWORTHY AI IN SAFETY CRITICAL SYSTEMS

OVERCOMING ADOPTION BARRIERS

Funded by
the European Union

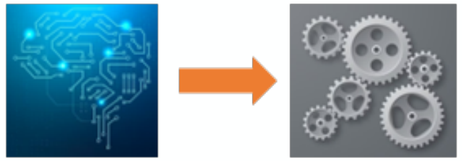
Safe and Explainable Critical Embedded Systems Based on AI

SAFEXPLAIN

www.safexplain.eu

SAFEXPLAIN ambition

- Architecting DL solutions **enabling certification/qualification**
  - Making them **adhere to "safety culture"**



  - Preserving **high performance**

  - Tailoring solutions to **varying safety requirements** (e.g., different safety needs for a coffee machine and a plane)





SAFEXPLAIN ambition achieved in breadth and depth

SAFEXPLAIN
Safe and Explainable
Critical Embedded Systems based on AI

BARCELONA SUPERCOMPUTING CENTER (BSC)
https://www.bsc.es/

IKERLAN, S. Coop (IKR)
https://www.ikerlan.es/

AIKO SRL (AIKO)
https://www.aikospace.com/

RISE RESEARCH INSTITUTES OF SWEDEN AB (RISE)
https://www.ri.se/

NAVINFO EUROPE BV (NAV)
https://www.navinfo.eu/

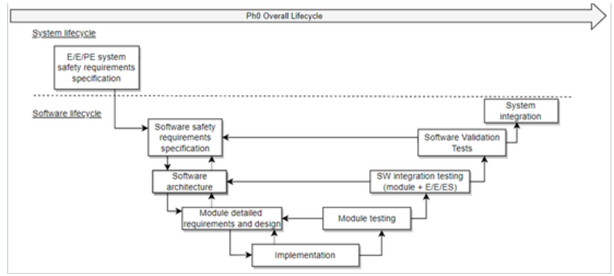
EXIDA DEVELOPMENT SRL (EXI)
https://www.exida-eu.com/

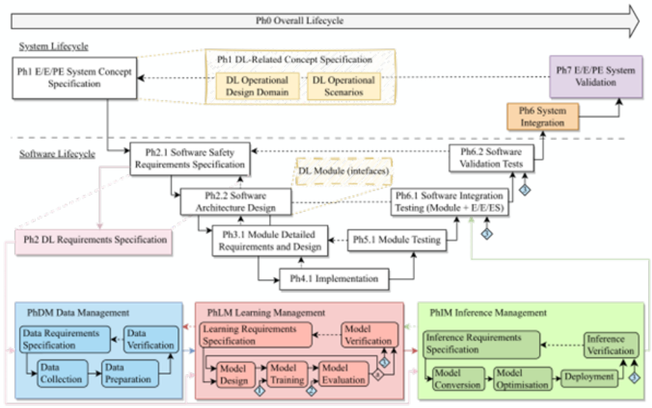
Jaume Abella
Project Coordinator

SAFEXPLAIN ambition and objectives

- Re-think safety lifecycle
  - **Keep principles** but with AI implementation in mind

- A **new AI-friendly lifecycle** has been devised and gone through a **positive assessment** by TÜV Rheinland

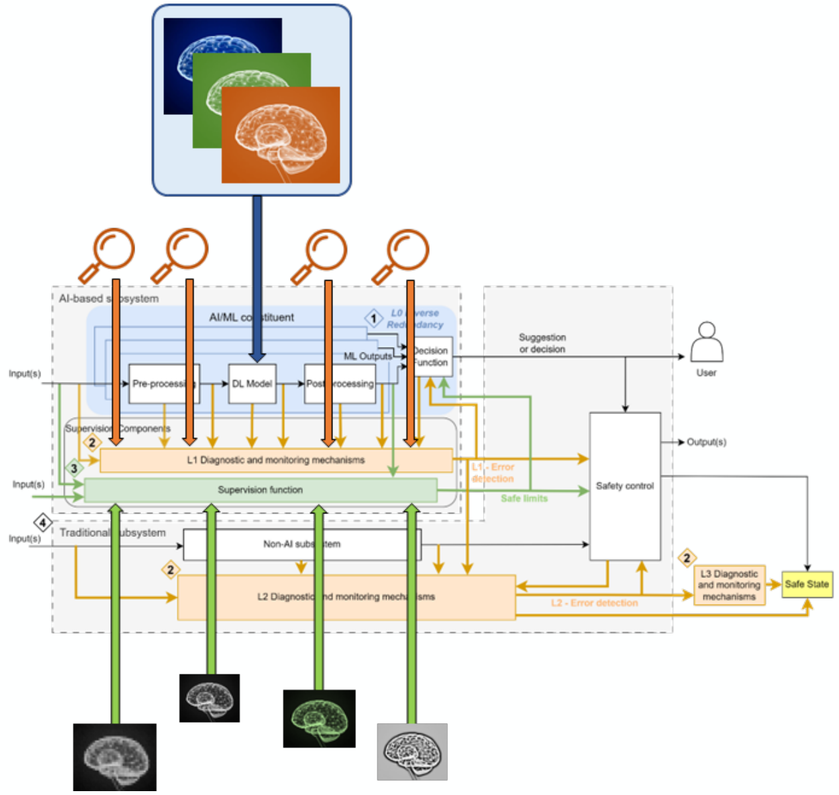




SAFEXPLAIN ambition and objectives

- Re-think AI software
  - Realize AI solutions **following safety principles** (redundancy, supervision, etc.)
  - Make **AI decisions explainable** (be able to understand why a given decision has been taken)

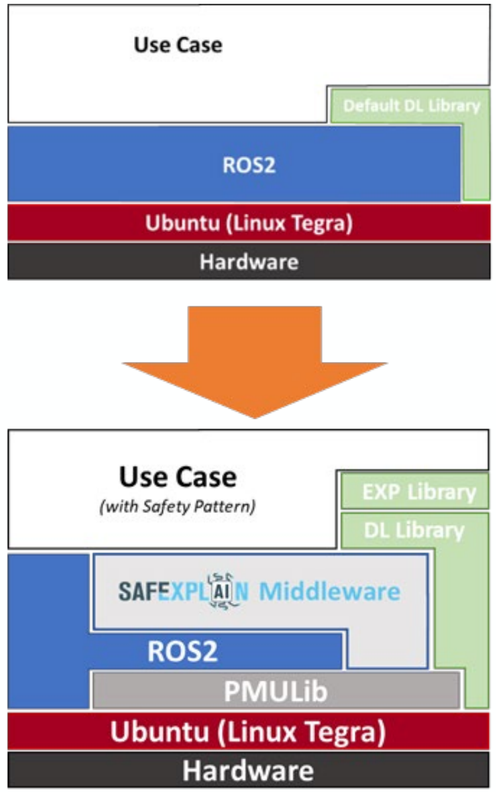
- A number of solutions have been devised matching the needs of the **new safety lifecycle**, including **safety partners** and appropriate **software architectures**

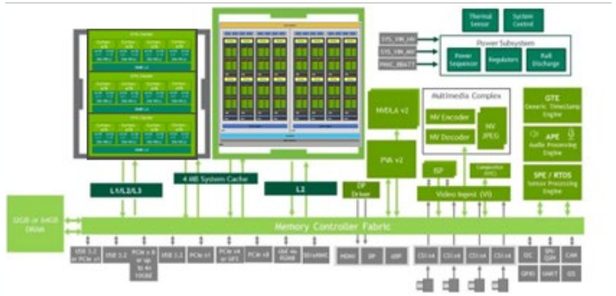


SAFEXPLAIN ambition and objectives

- Preserve performance and accuracy
  - Keep **high accuracy**
  - Keep **high performance**

- **Middleware** up and ready, **libraries** complete, and **case studies complete**
  - An end-to-end demo almost finalized





SAFEXPLAIN ambition and objectives

- Assess findings in three key domains (demos ready)

Railway

a relatively **controlled scenario** with more limited driving options

Automotive

a **complex scenario** with ability to change lanes and move freely, and many objects around

Space

a **complex scenario** involving 3D navigation and extreme lighting conditions

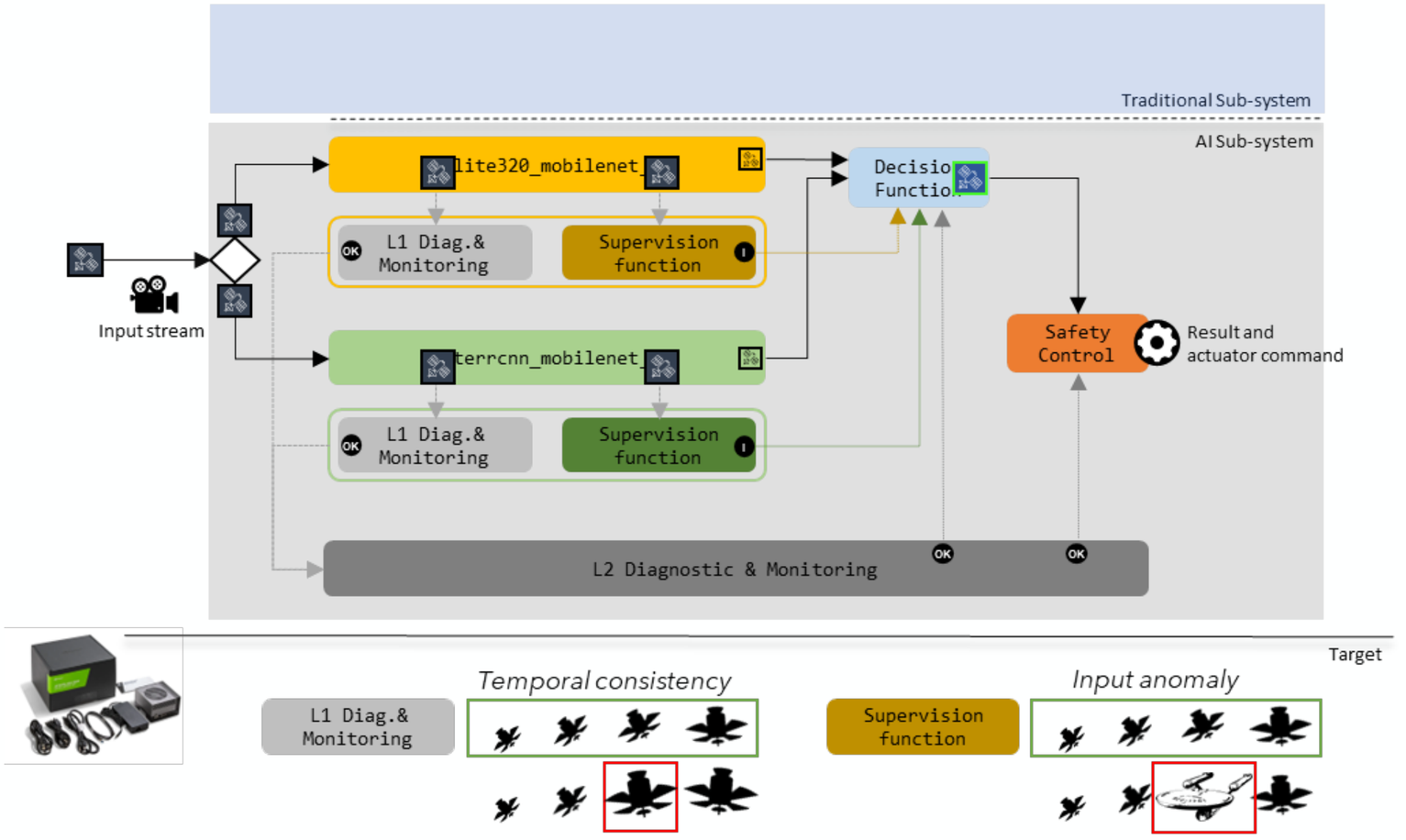
Open source case study for public distribution

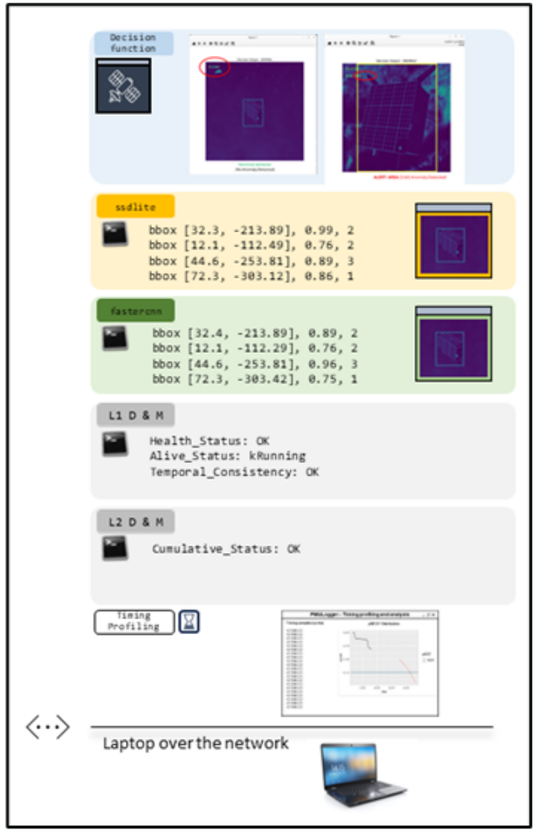
Demo example (based on Space)

a **limited scenario** with a subset of the conditions of the space case

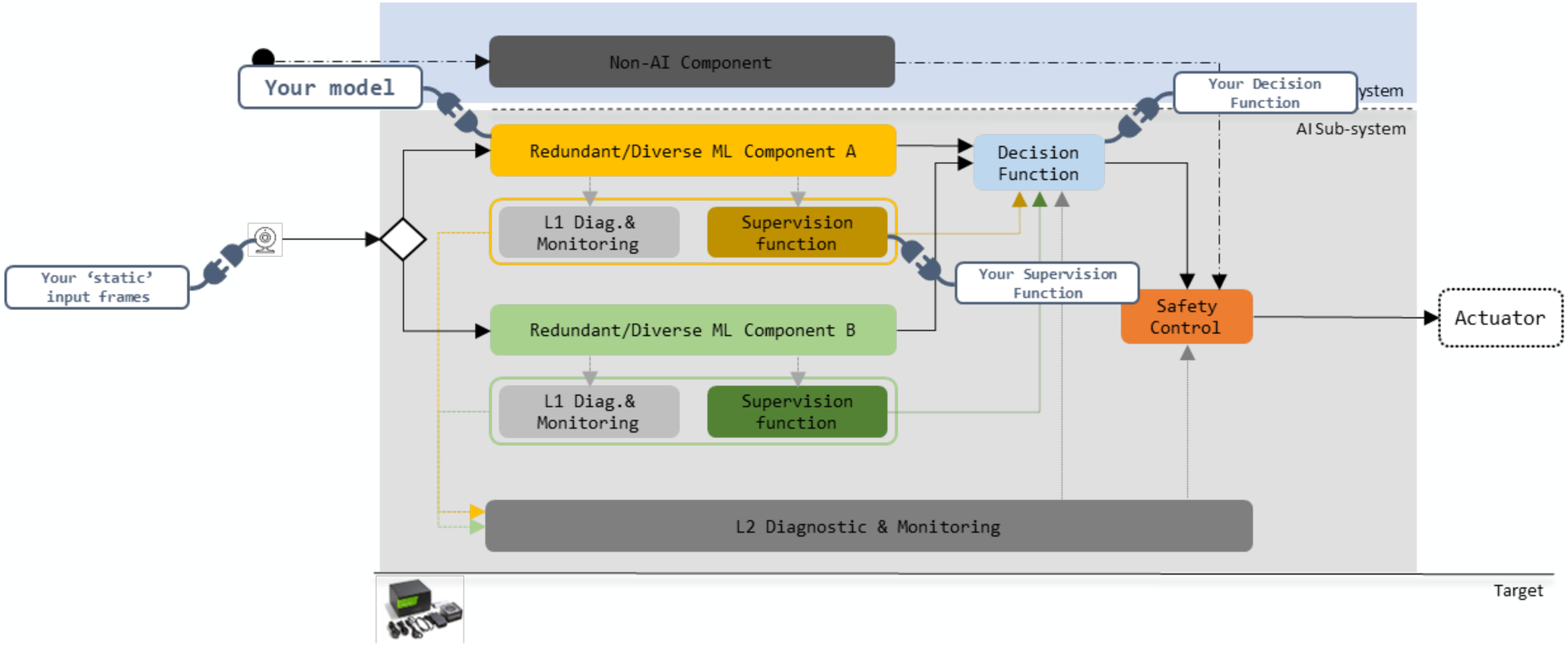
- **Influence the different relevant standards** on AI for safety-critical systems
  - Already in synch with ASPICE ML workgroup, among others

SAFEXPLAIN CORE Demo (Space)





SAFEXPLAIN CORE Demo Generalization



Summary

- An AI-amenable **safety lifecycle** has been developed and positively assessed by TÜV Rheinland
- **Software architectures** for AI-based systems already defined for multiple safety patterns
- **AI solutions** already devised for the different AI components in the software architectures
- **Platform support** (middleware, libraries, and software mechanisms) to enable AI-based safety-critical real-time software execution complete
- **Case studies ported** and appropriate demos generated
- **Open source case study** finalized
- External industrial and standard committees **feedback highly positive** so far across domains

SAFEXPLAIN
Safe and Explainable
Critical Embedded Systems based on AI

ULTIMATE

EdgeAI-Trust

FINAL EVENT

TRUSTWORTHY AI IN SAFETY CRITICAL SYSTEMS

OVERCOMING ADOPTION BARRIERS

Funded by
the European Union

mUlti-Level Trustworthiness to IMprove the Adoption of hybrid arTificial intelligencE (ULTIMATE)

ULTIMATE

CONTEXT (2021)

- AI grows in sophistication, complexity, autonomy …
- Need of large-scale data (often difficult to access)
- Lack of trust by end-users (e.g. not explainable, far from reality … and not suited for AI regulation)

- Investigation of Trustworthy Hybrid AI as a promising solution

ULTIMATE

MAIN CHALLENGES

- Investigation of:
  - Trustworthy (technically & ethically speaking)
  - Hybrid AI (good trade-off between data-driven AI and knowledge-driven AI)
  - For end-users (industrial adoption)
- What suggests an end-to- end AI trustworthiness life cycle and related activities ideally

ULTIMATE

THREE USE CASES

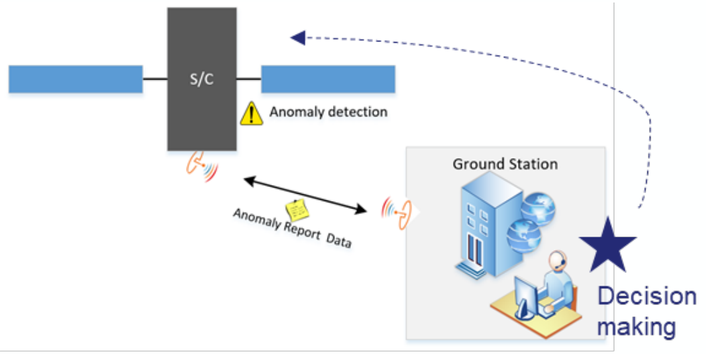
TO ADDRESS THESE CHALLENGES

SMART FDIR for SATELLITE (TAS UC)

- Equipment failure prognosis on a satellite board (high fidelity simulator)

- Objectives
  - Autonomous detection of Reaction Wheels degradation or anomalies

- Challenges
  - Early detection of anomalies (weak signals)
  - Classification of the anomaly type and severity
  - Transmission of a trigger to ground teams for check and decision



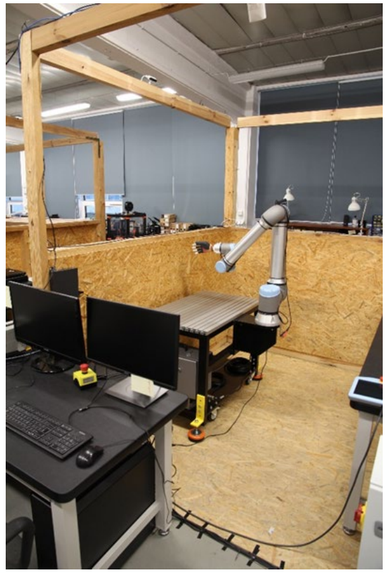
Funded by the European Union

ROBOTIC WORKSHOP (PIAP UC)

- 1 AMR, 2 stationary manipulators and human workers

- Objectives: human detection and tracking for robotic operation efficiency improvement (overall performance improvement of more than 15%)

- Challenges
  - For workers:
    - Secure interaction with dangerous tools and robots
    - Reduced direct contact between dangerous products and human worker
  - For the product:
    - Increased involvement of system features to automate and support the production process
  - For the efficiency of the process:
    - During the daily operation (8h), robots are stopped more than 2h

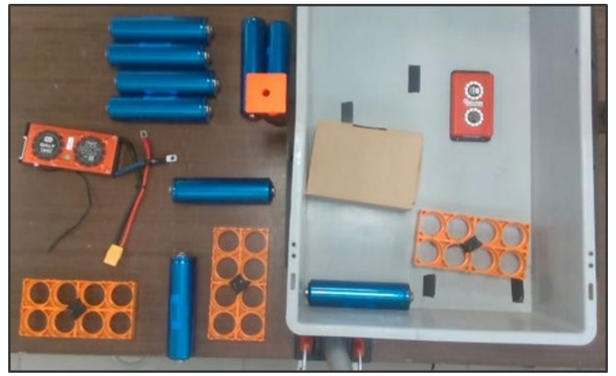


INDUSTRIAL ROBOTS (ROB UC)

- 2 Mobile manipulators and human workers

- Objectives
  - Adaptive behaviour to new situations regarding objects and navigation

- Challenges
  - Object detection & manipulation (e.g. kitting)
    - Ability to detect a specific object among others
  - Safe navigation
    - Avoid collision with humans to autonomously navigate to the target destination
  - Efficient cooperation between humans and robots



ULTIMATE

SAFEXPLAIN
Safe and Explainable
Critical Embedded Systems based on AI

ULTIMATE

EdgeAI-Trust

FINAL EVENT

TRUSTWORTHY AI IN SAFETY CRITICAL SYSTEMS

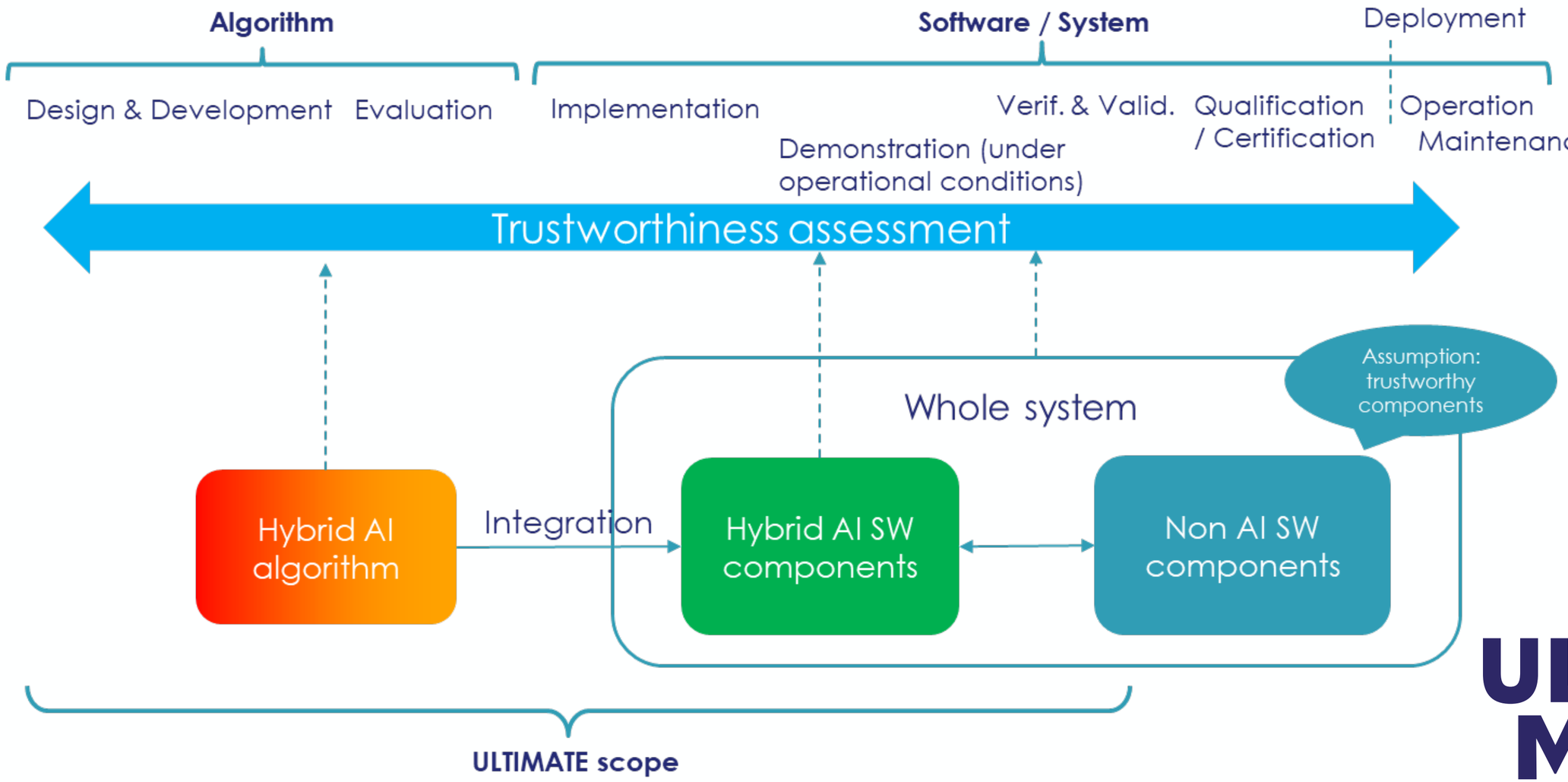
OVERCOMING ADOPTION BARRIERS

Funded by
the European Union

ULTIMATE APPROACH

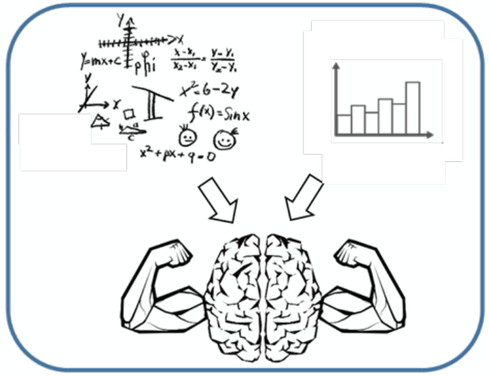
ULTIMATE

ULTIMATE HYBRID AI LIFE CYCLE



AMBITION: PIONEER THE TRUSTWORTHINESS OF INDUSTRIAL-GRADE HYBRID AI

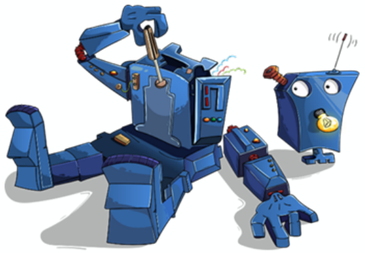
Hybrid AI



Develop **innovative architectures** to construct and train hybrid AI algorithms

Design & Development

Design **rigorous evaluation methodologies** with appropriate properties



Evaluation

Ensure the **ethical compliance**

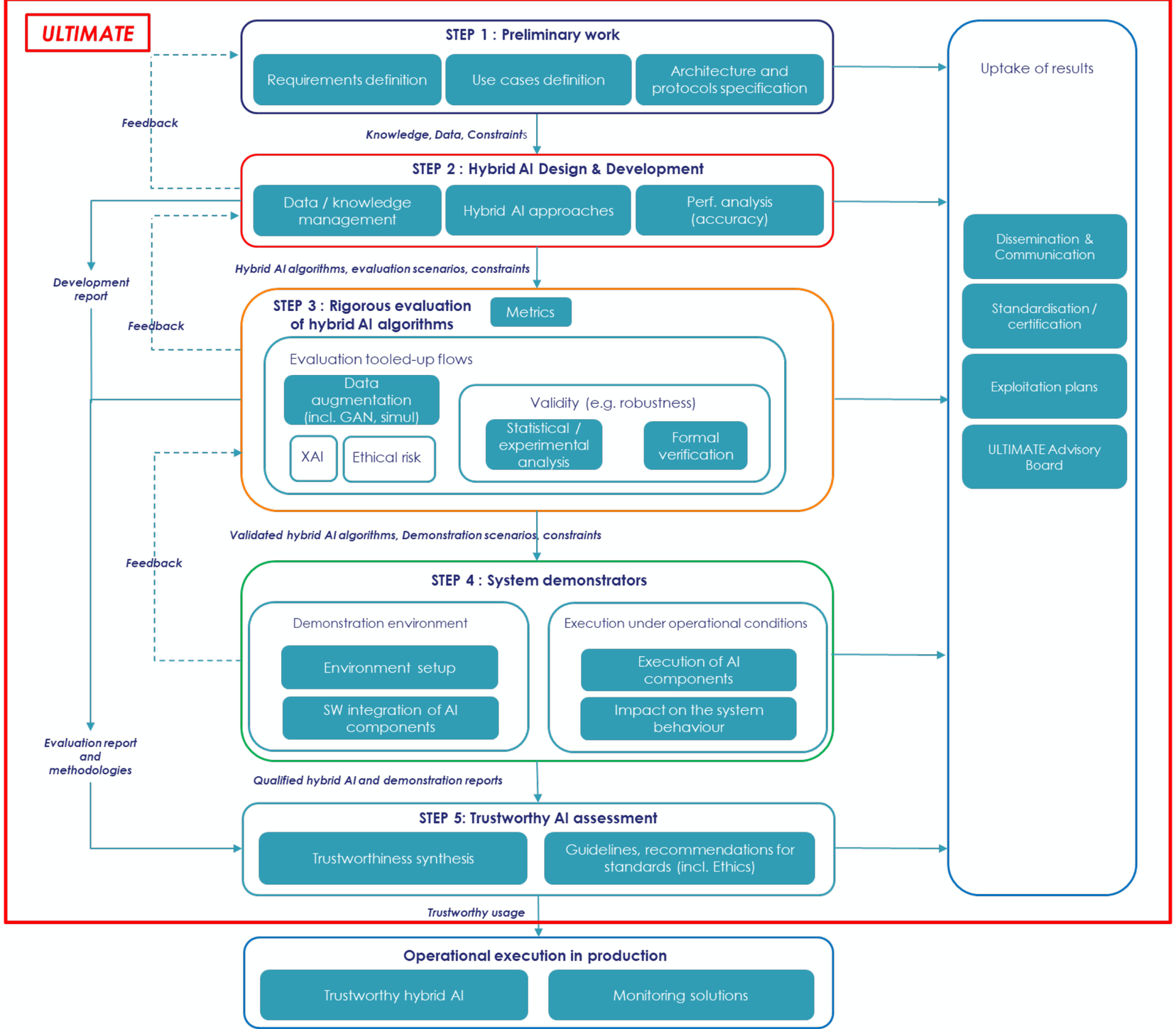
Trustworthiness

Implement the hybrid AI algorithms under **operational conditions**

Pre-industrialisation

ULTIMATE

ULTIMATE GLOBAL FLOW



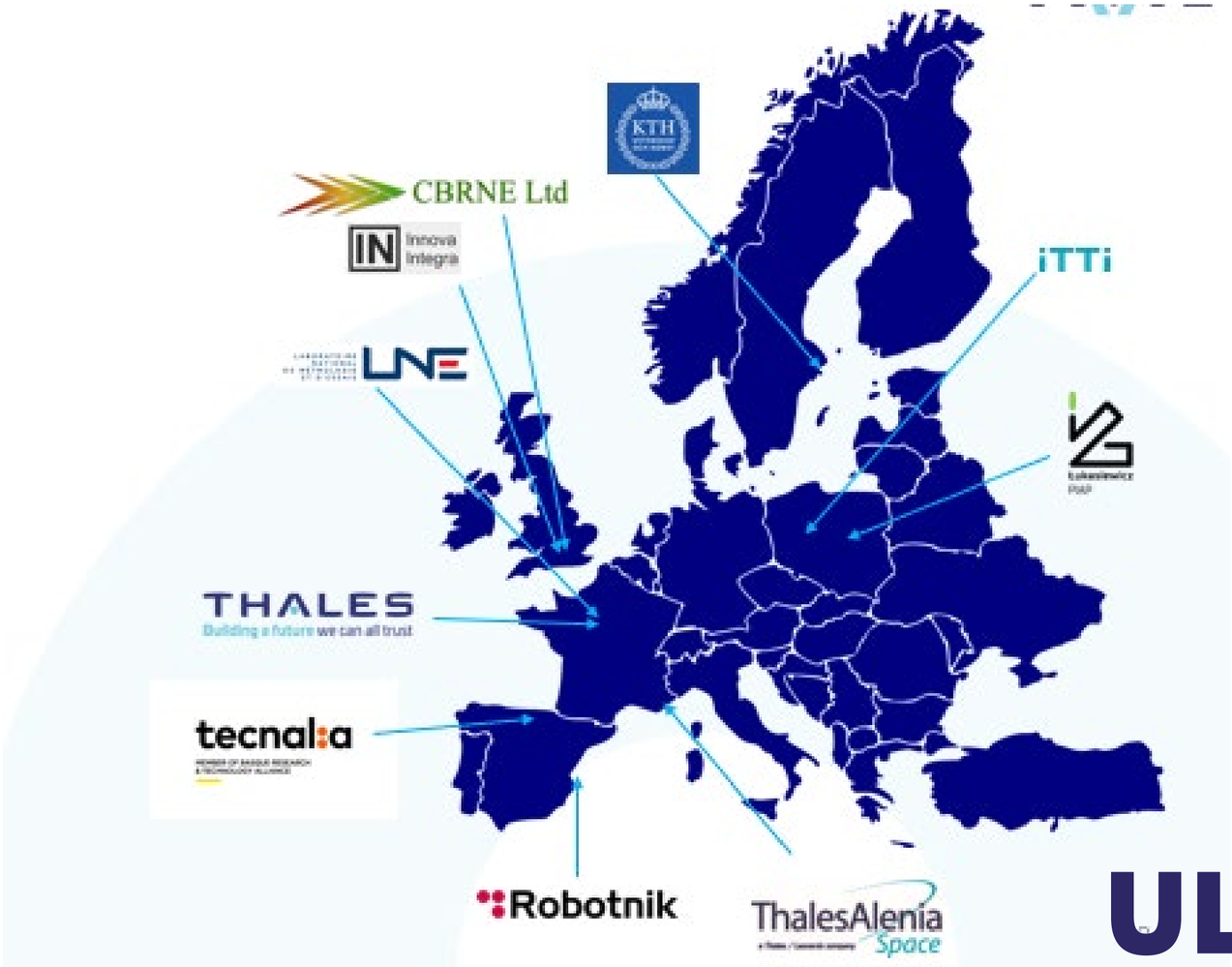
Contact

michel.barreteau@thalesgroup.com

https://ultimate-project.eu/

Find us on

https://www.linkedin.com/groups/12735469/



ULTIMATE

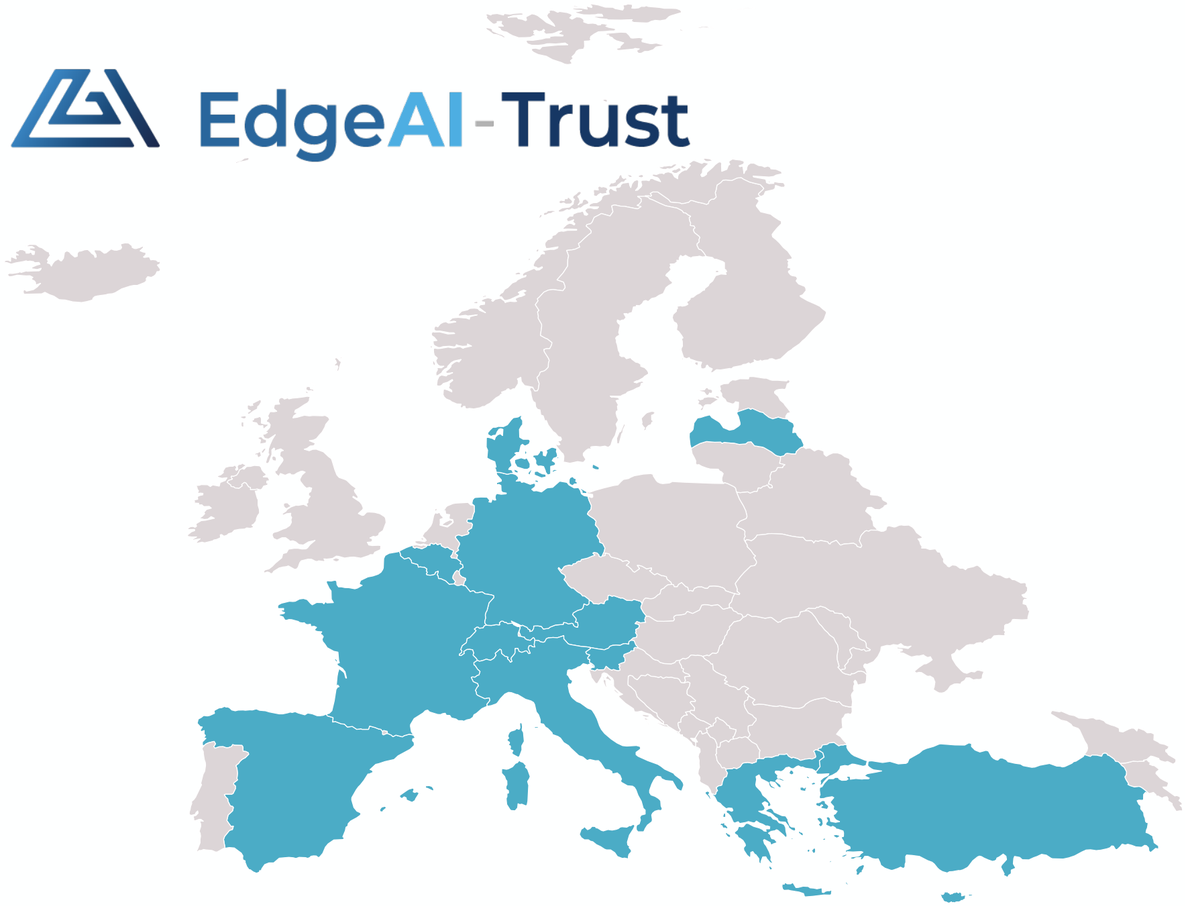
Challenges to AI Adoption in Safety Critical Systems: Insights from EdgeAI-Trust

Mohammed Abuteir , TTTech Auto

Content

Project profile

- Project start: 01.05.2024
- Duration: 3 years
- Partner: 53
- Involved countries: 13
- Budget: ~38M€
- Supply Chains: 8
- Demonstrators: 21



Main topic: Trustworthy EdgeAI Ecosystem

VISION

Create a **trustworthy edge AI ecosystem including an AI architecture, components, orchestration techniques, development tools and a community** enabling the real-time collaboration of heterogeneous edge devices, while maintaining the highest levels of security, scalability, sustainability and robustness.

Targeted key innovations:

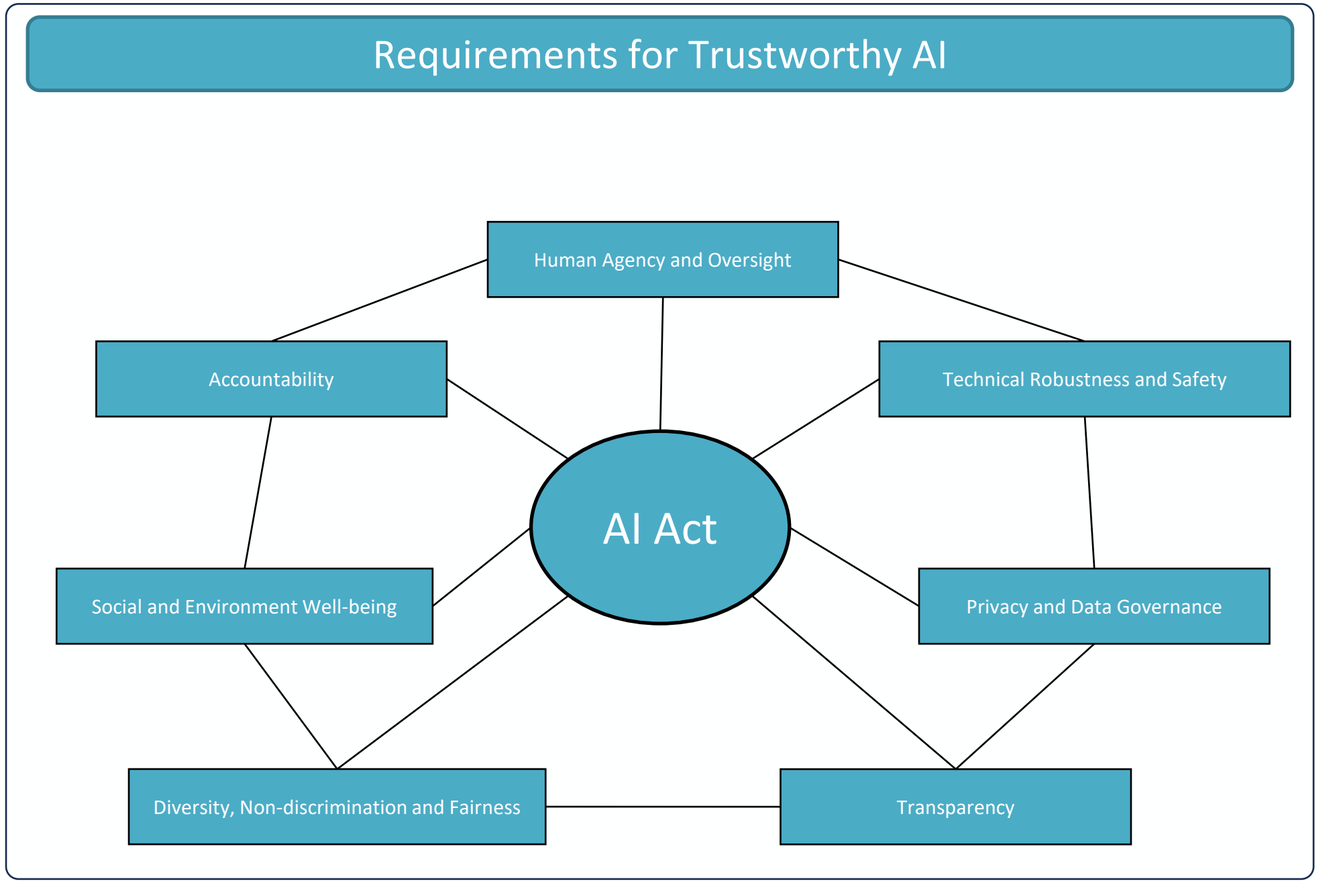
- Federated learning, decentralized AI, distributed machine learning and Probabilistic AI
- Data management, data model, data analysis for data-driven approaches and models
- Heterogeneity of computing/communication/AI resources, applications, and requirements
- Fault tolerance, robustness, resilience, and trust
- Energy efficiency and resource optimization
- Real-time capabilities and adaptability
- Virtualization control and sensor-based AI
- Interoperability, Scalability, Open source, Regulatory compliance

MISSION

EdgeAI-Trust will develop a **domain-independent end-to-end collaborative AI architecture and large-scale edge AI technologies**. The core of our approach increases trustworthiness, reliability, safety, security, energy efficiency, sustainability, and societal acceptance of AI through orchestration of AI components, virtualization of heterogeneous distributed AI resources for federated learning, rigorous plausibilization, explanation and monitoring of AI decisions and standardized interfaces and toolchains for optimizing and validating edge AI solutions.

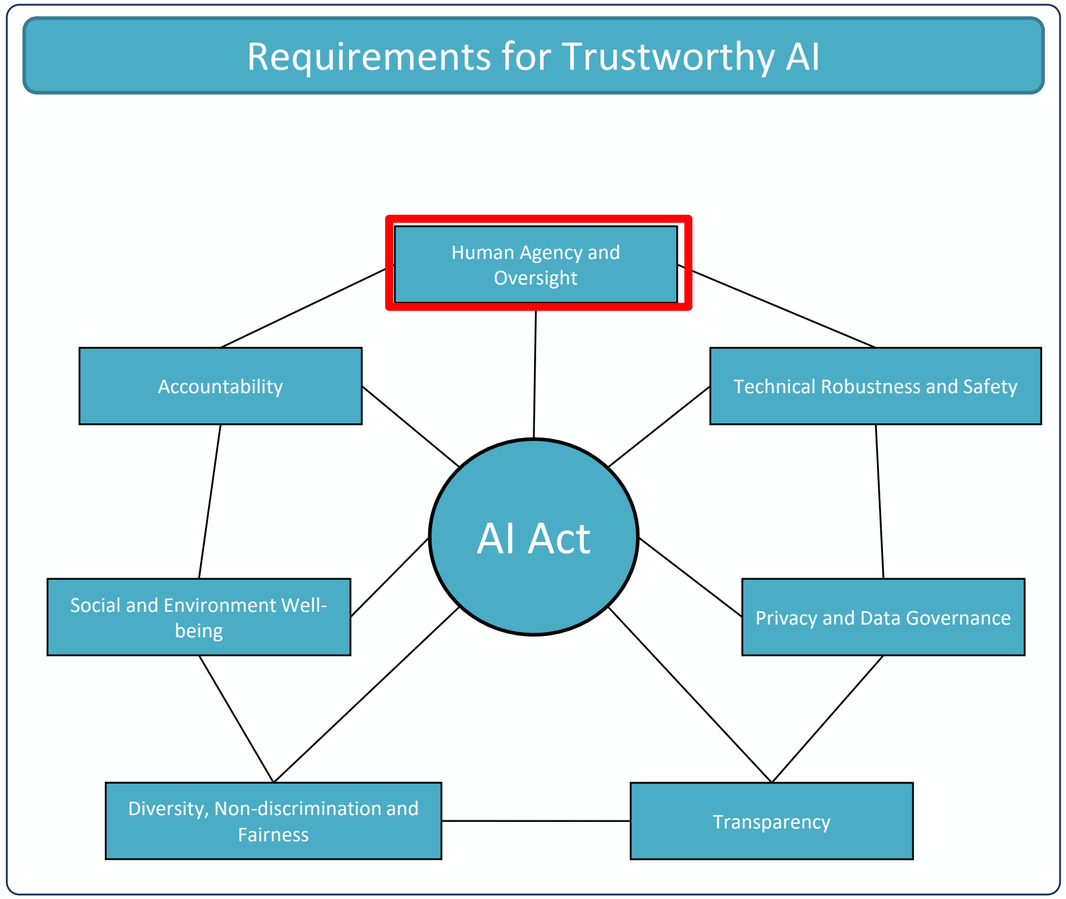
Link to the EU AI Act

The scope of EdgeAI-Trust is fully in line with the EU AI Act requirements!



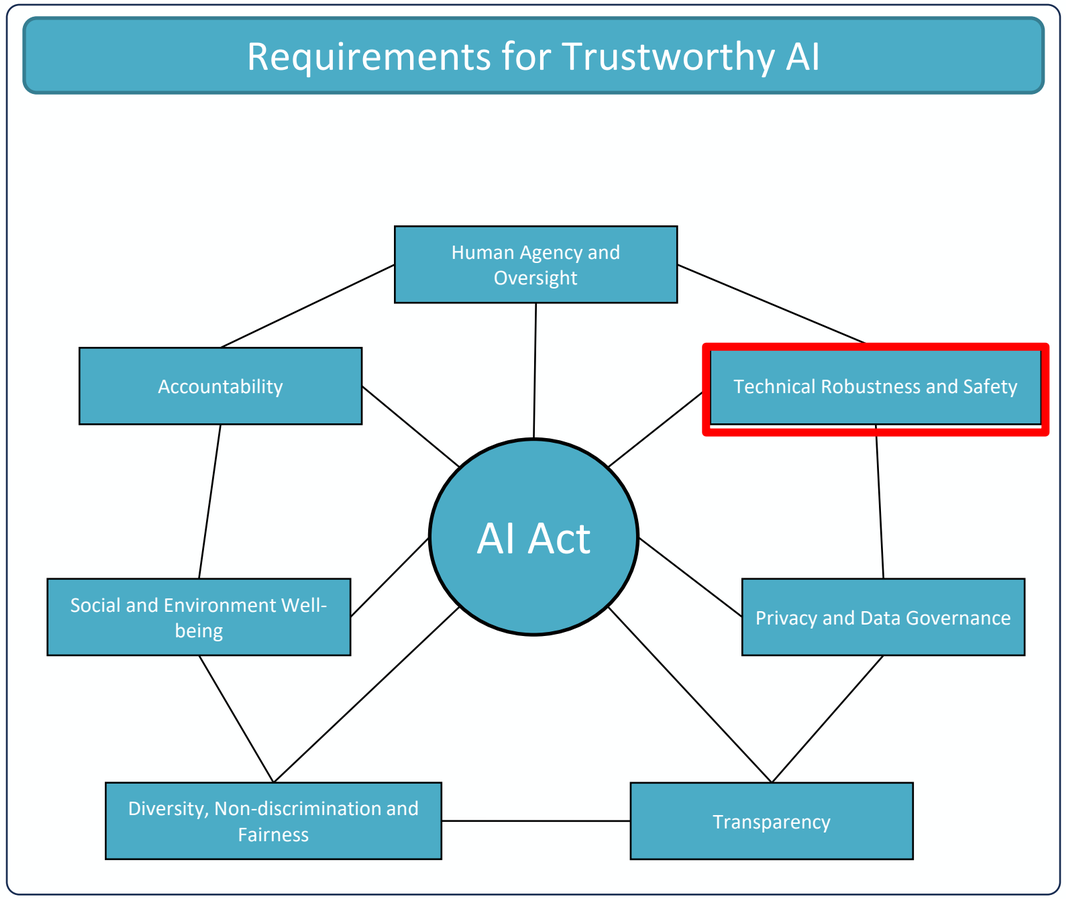
Human Agency and Oversight

- **Human Oversight Importance**
  - Ensuring humans maintain control over AI decisions is critical in safety-sensitive fields like healthcare and transportation.
- **Explainability and Transparency**
  - AI systems must offer clear explanations and transparent monitoring to avoid black-box operations and build trust.
- **Mitigating Automation Bias**
  - Robust human-in-the-loop designs reduce the risk of over-reliance on AI recommendations in critical decisions.
- **EdgeAI-Trust Solution**
  - Integrates orchestration, explainability, and real-time monitoring tools empowering users to audit AI decisions effectively.



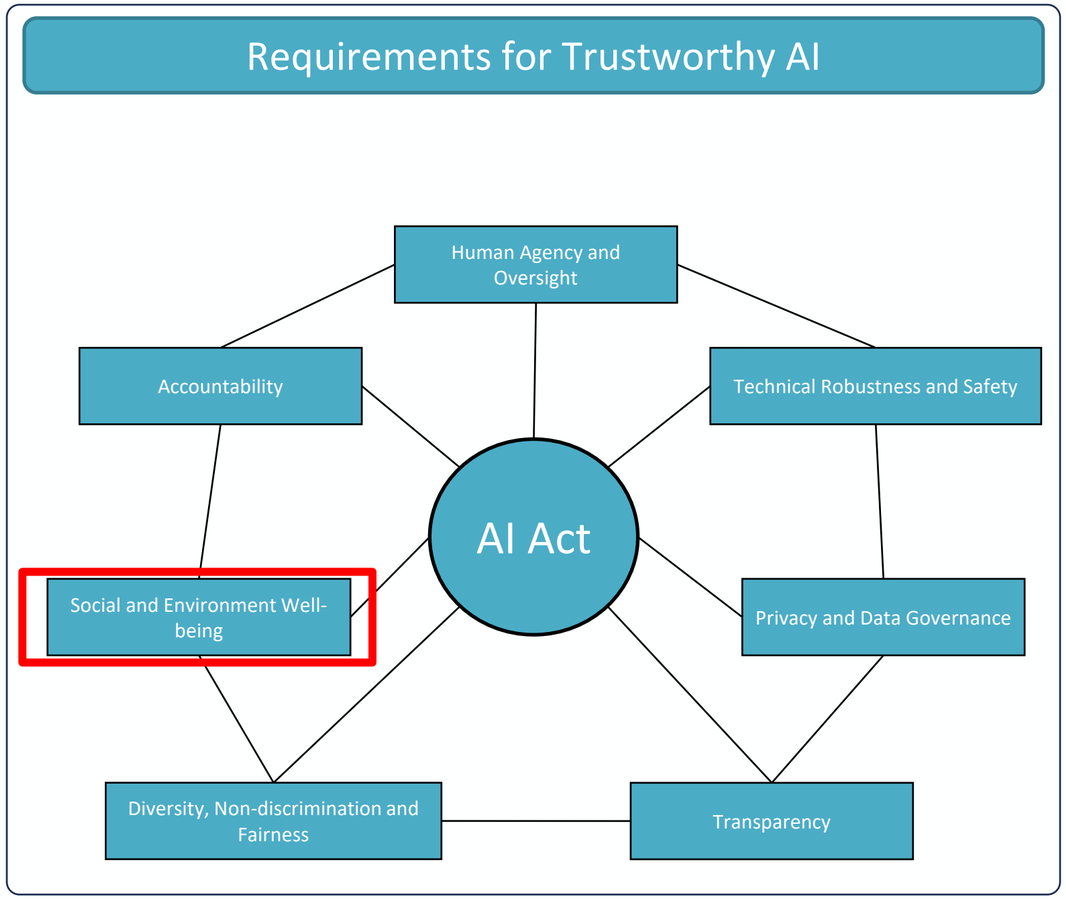
Technical Robustness and Safety

- **Resilience to Challenges**
  - AI systems must handle unexpected inputs, adversarial attacks, and operational anomalies to ensure safety.
- **Rigorous Validation and Monitoring**
  - Continuous validation and monitoring are essential to maintain reliability and detect failures early.
- **Fail-Safe Mechanisms**
  - Implementing fail-safe designs helps AI systems gracefully handle errors and system degradations.
- **EdgeAI-Trust Approach**
  - EdgeAI-Trust emphasizes advanced validation, plausibilization, and robust design to prioritize AI safety.



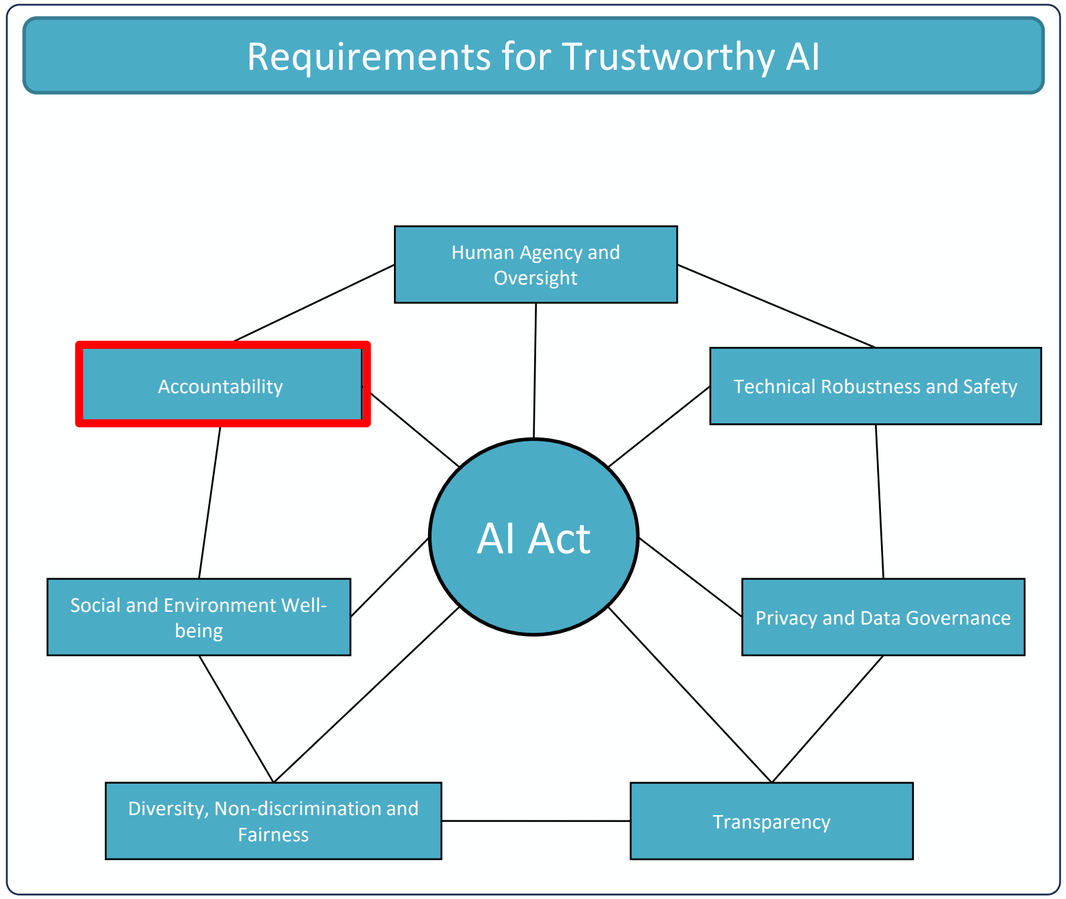
Social and Environmental Well-being

- **Energy Efficiency in AI**
  - Minimizing energy consumption is critical for reducing the carbon footprint of AI systems in safety-critical domains.
- **Social Impact of AI Automation**
  - Ensuring AI-driven automation respects employment and societal well-being is essential for responsible AI deployment.
- **Sustainable AI Design Principles**
  - EdgeAI-Trust promotes sustainability through energy-efficient algorithms, scalable architectures, and responsible innovation.



Accountability

- **Importance of Accountability**
  - Accountability ensures safe and ethical AI deployment in safety-critical systems, protecting stakeholders and users.
- **Challenges in Responsibility**
  - Defining and assigning responsibility for AI actions is complex in distributed and AI-driven environments.
- **Tools for Transparency**
  - Standardized toolchains and traceability enable transparent auditing and decision attribution in AI systems.
- **Fostering Trust and Compliance**
  - Accountability frameworks support trust, legal, and regulatory compliance for AI stakeholders and operators.



Technological Approaches to Overcoming Challenges

- **Collaborative AI Architecture**
  - Enables seamless integration and real-time collaboration of heterogeneous AI components across edge devices.
- **Federated Learning and Virtualization**
  - Supports secure, privacy-preserving model training across distributed resources using virtualization techniques.
- **Explainability and Monitoring Tools**
  - Ensures AI decisions are reliable, interpretable, and continuously monitored for safety-critical systems.
- **Energy Efficiency and Standardization**
  - Focuses on sustainable AI solutions with standardized interfaces to optimize and validate edge AI systems.

Remaining Barriers and Research Opportunities

- **Certification and Compliance Challenges**
  - Evolving standards create challenges in certification and regulatory compliance for AI in safety -critical systems.
- **Scalability and Integration Issues**
  - Scaling AI across domains and integrating with legacy systems demand continuous innovation and adaptation.
- **Societal Trust and Acceptance**
  - Building societal trust requires transparency, stakeholder engagement, and proven safety records.
- **Future Research Directions**
  - Research focuses on explainability, validation, stakeholder involvement, and edge AI performance improvements.

Digging deeper: Insights from the coordinators

Panel moderated by Irune Agirre

Questions?

THANK YOU!

Want to know more about SAFEXPLAIN/ EdgeAI-Trust?

Join us here, in the Sala Master, after the coffee break

Want to know more about ULTIMATE?

Join us next door, in the Sala Agora, after the coffee break