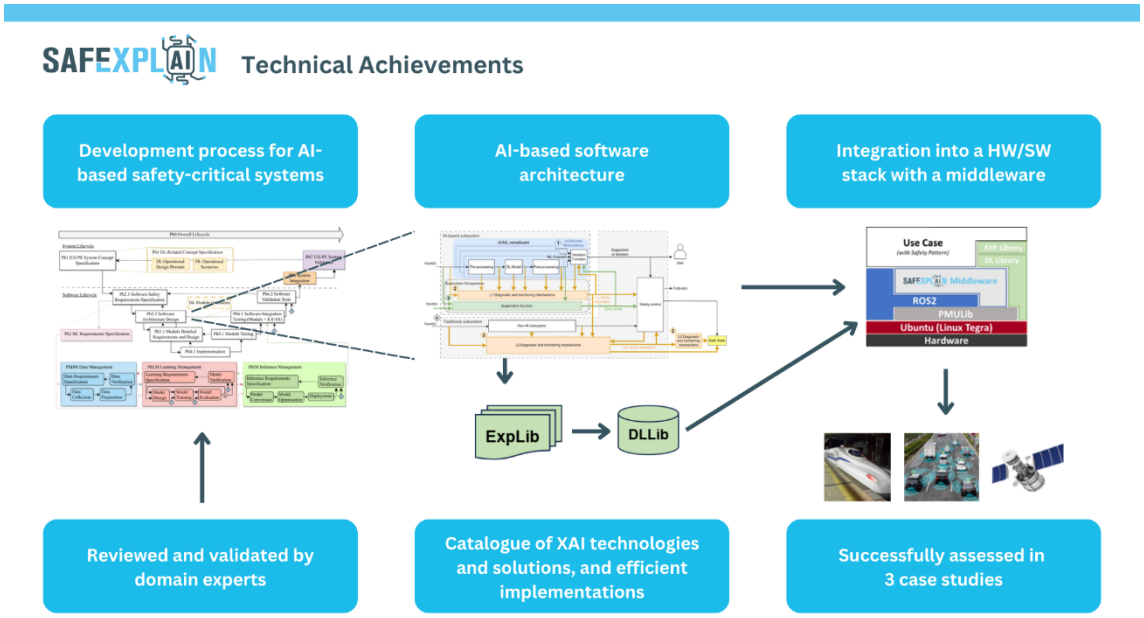
SAFEXPLAIN delivers final results: Trustworthy AI Framework, tools and validated case studies



Barcelona, 25 November 2025—After three years of research and collaboration, the EU-funded SAFEXPLAIN project, coordinated by the Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS) has released its final results: **a complete, validated and independently assessed framework for deploying safe, explainable and certifiable AI in safety-critical systems**. These results pave the way for industrial uptake across the **automotive, space, railway** and other safety-critical sectors.

The SAFEXPLAIN consortium, composed of leading research institutions and industrial partners across Europe, has demonstrated that **AI can meet the stringent safety, reliability and certification requirements of critical systems**, overcoming a critical obstacle to the widespread adoption of AI in these domains.

An End-to-End Trustworthy AI Methodology

SAFEXPLAIN's end-to-end approach integrates AI models with diagnostic mechanisms, supervision functions, and control logic as part of the proposed software architecture. This software architecture is the main product of the project's AI-FSM 2.0 life cycle, positively assessed by TÜV Rheinland, which rethinks the safety lifecycle for AI systems by merging traditional validation and certification principles with AI-specific development needs.

Combined with tailored software architectures and components, this AI-friendly lifecycle ensures traceability, explainability and accountability by design, enabling people to understand, validate and certify how AI reaches its decisions. This approach is essential for meeting existing functional safety standards such as

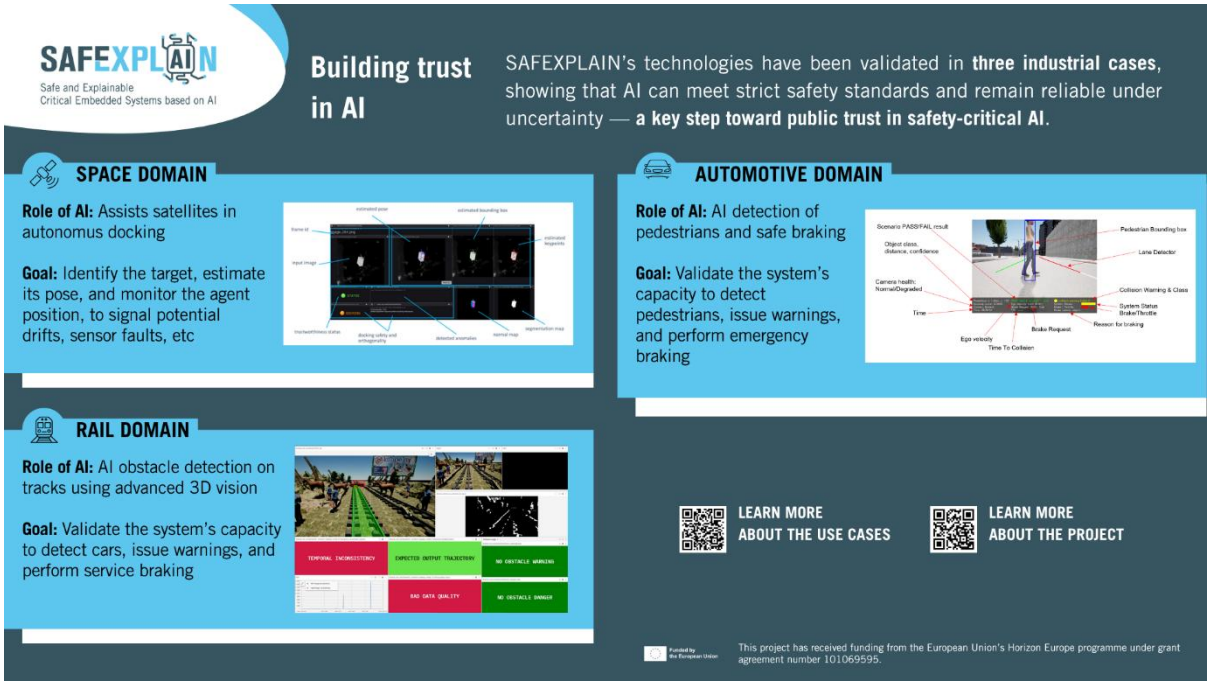
ECSS (space), ISO26262/ ISO21448/ ISO8800 (automotive), and EN 50126/8 (rail), which define the process and evidence necessary to certify that systems operate safely even under faults or uncertainty.

Ready-to-use technologies and resources

SAFEXPLAIN has delivered a series of tested tools, libraries, and multimedia material that enable implementation, validation and demonstration of trustworthy AI on industrial platforms. The project's webinar series, walks through the project's top technologies and their integration, culminating in the release of the Core Demo, an open, complete SW platform that demonstrates critical functionalities on selected examples from the space, automotive and rail use cases.

Validation across industrial case studies

The SAFEXPLAIN framework has been validated across three real-world case studies (space, automotive and rail), each tackling distinct safety-critical challenges.



Testing under operational and stress conditions showed that SAFEXPLAIN's diagnostic and supervisory components enhance anomaly and fault detection and management, ensuring AI can operate safely even under uncertainty, an essential step for building public trust in AI systems that impact human safety.

A foundation for the Future of Trustworthy AI in safety-critical systems

The project's technologies and open resources will remain available to the research and industrial community, fostering the continued evolution of safe and explainable AI.

"With SAFEXPLAIN, we have demonstrated that safety is not an afterthought in AI. It can be integrated so systems are correct by construction", said Jaume Abella, project coordinator and head of the BSC's High Performance Embedded Systems Lab. "The project provides both a safety-relevant development process and its concrete realization through software architecture, libraries and case studies, offering a complete path from concept to certifiable AI".

More information and open resource are available on the project website https://safexplain.eu/ .