



Safe and Explainable
Critical Embedded Systems based on AI

AI explainability solutions to build safety-relevant AI-based software architectures

Barcelona, 23 September 2025

Consortium partners: NavInfo Europe, BSC, Ikerlan, AIKO, RISE, Exida

Thanh Bui (RISE)



Funded by
the European Union

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

Agenda

- Challenges
- Strategy
- Explainable AI enabled compliance

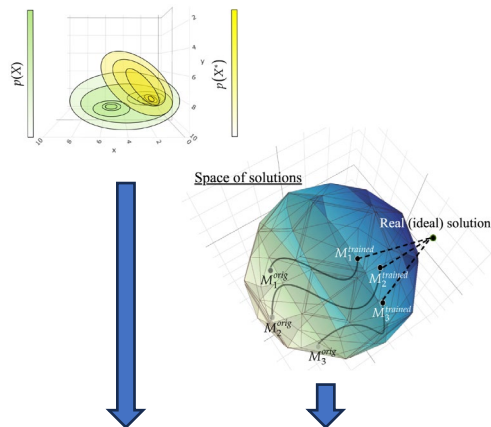


Challenges



WHY DL software is so different for safety culture?

- **Domain uncertainty:** Datasets correctly represents the problem
- **Model epistemic uncertainty:** model correctly models the problem
- **Aleatoric uncertainty:** Irreducible uncertainty concerned with the nature of the problem

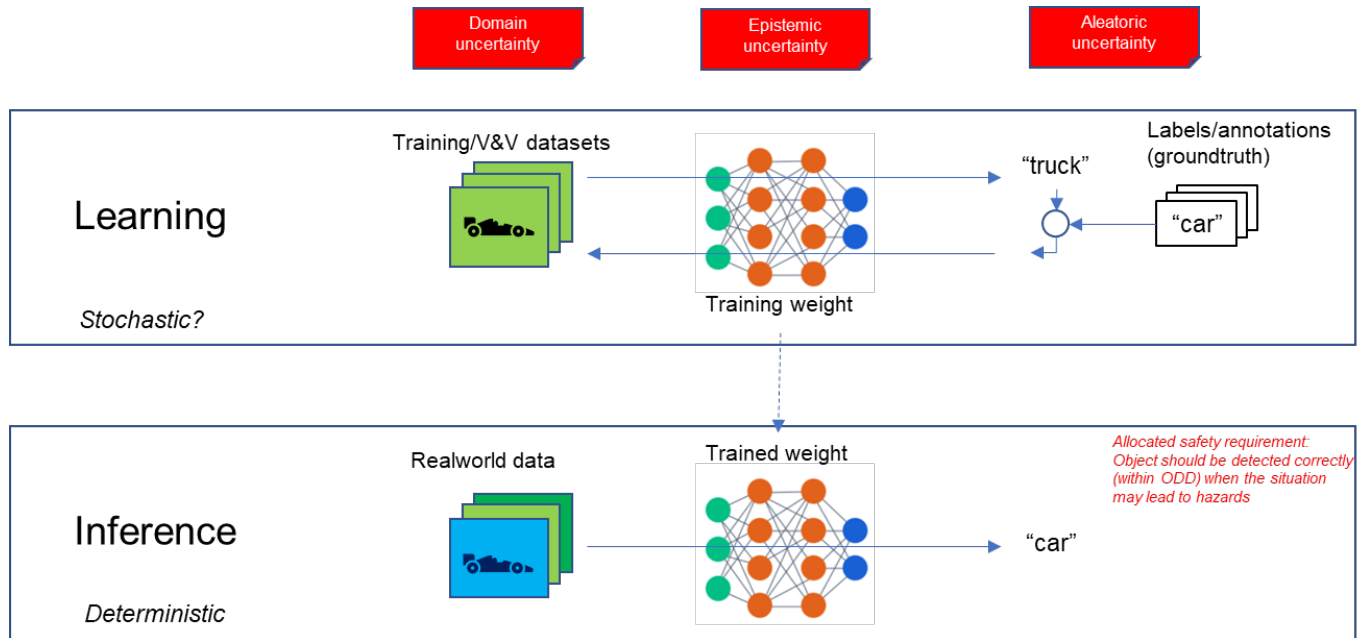


$$p(x^*, y^* | X, Y) = p(x^* | X) \int_{\mathbb{M}} p(M | X, Y, x^*) \cdot p(y^* | X, Y, M, x^*) dM$$

X	Predictive AI System Defined by $M \subset \mathbb{M}$	Y
Domain uncertainty	Epistemic uncertainty of the predictive system	Aleatoric uncertainty of the predictive system

Brando, A., Serra, I., Mezzetti, E., Cazorla Almeida, F. J., & Abella Ferrer, J. (2023). Standardizing the probabilistic sources of uncertainty for the sake of safety deep learning. In Proceedings of the Workshop on Artificial Intelligence Safety 2023 (SafeAI 2023) co-located with the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI 2023): Washington DC, USA, February 13-14, 2023. (Vol. 3381). CEUR Workshop Proceedings.

DL uncertainties





Strategy

```
def apply(self, e):  
    if (e == t.apply(e[0], e[1], e[2])): r = e[0]  
    ) else if (e):  
        for (i = 0; i < 1; i++)  
            if (e == t.call(e[1], 1, e[2]), r == 13)  
    ) else  
        for (i in e)  
            if (e == t.call(e[1], 1, e[2]), r == 13)
```



Strategy: Reduce and Control uncertainties

- Functional safety lifecycle (AI-FSM)
 - Reduce “reducible” uncertainties
 - “Know” residual uncertainties
- Safe operation architecture (OM)
 - Monitor uncertainties
 - Safe actions & diagnosis

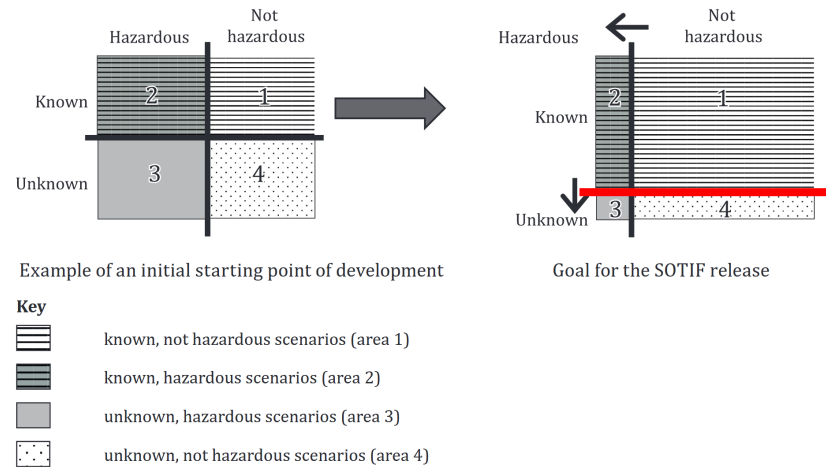
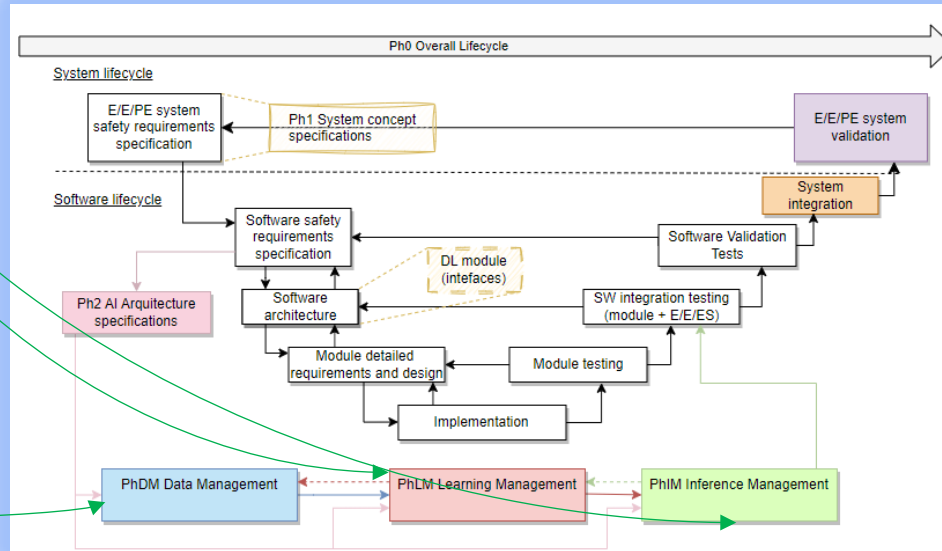


Figure 8 — Alternative evolution of the scenario categories resulting from the ISO 21448 activities

AI-FSM XAI solution mappings

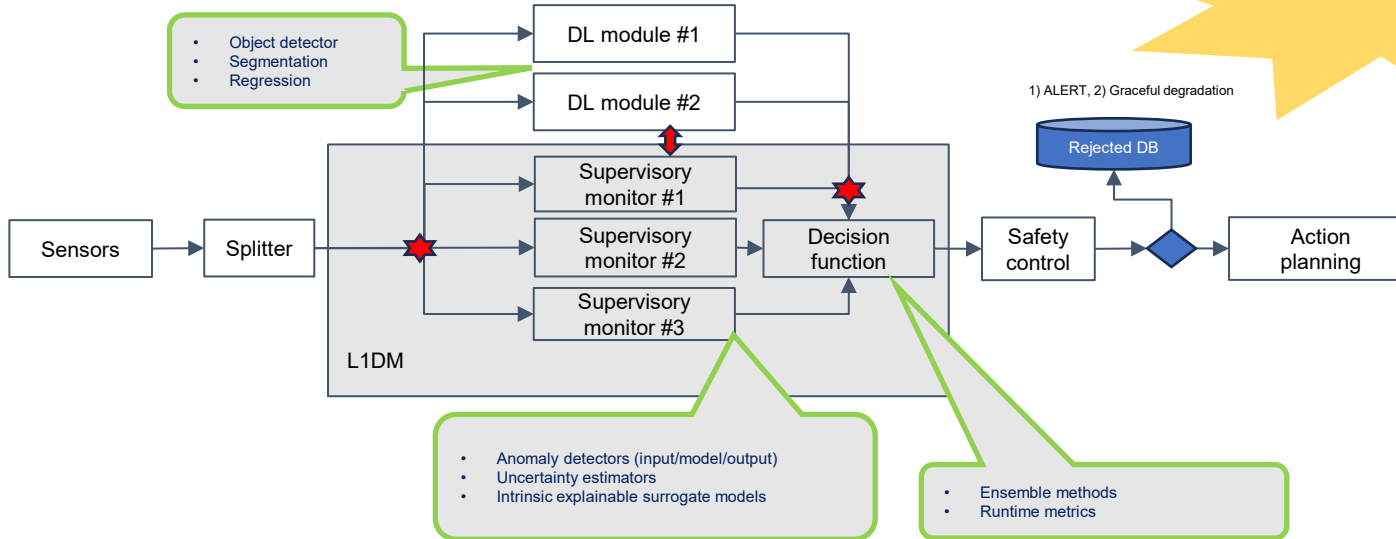
Subject	Transparency approach	Representation	Scope	Audience	Technique
Data	Decomposability	Text/tabular	Local	AI dev	Intrinsic
Model	Simulatability	Graph	Global	Safety expert	Posthoc
	Algorithmic	Image		End user	



**XAI as
enabler to
reduce
uncertainties**

Safety architecture (Operation & Monitoring)

XAI as enabler to manage residual uncertainties



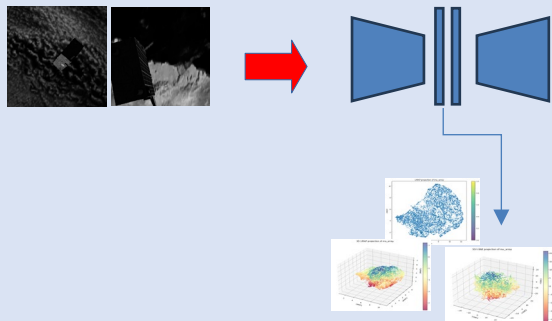


XAI enabled compliance

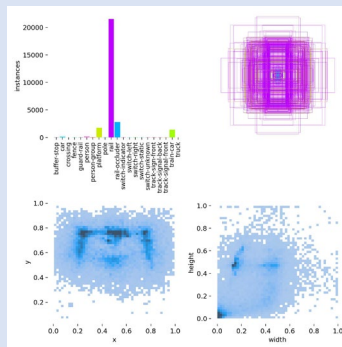


Data Management

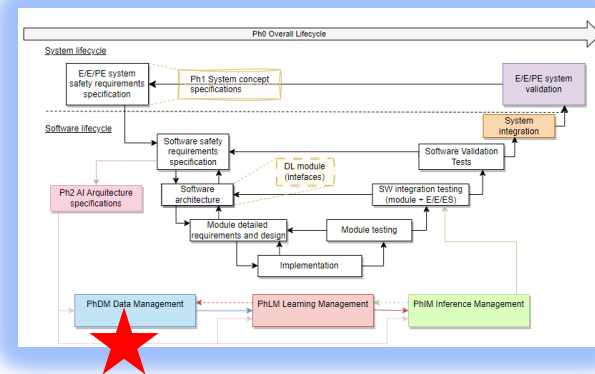
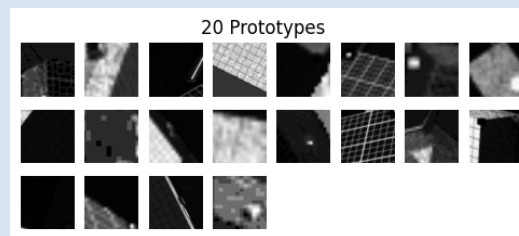
Data descriptor



Data profiling

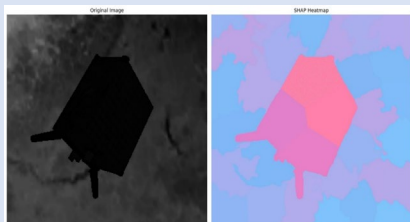


Data prototypes

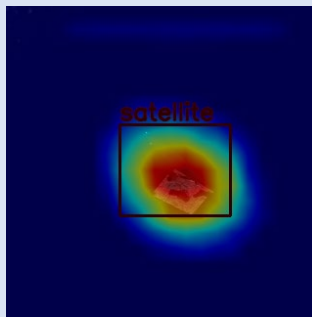


Learning/Inference Management

Feature importance

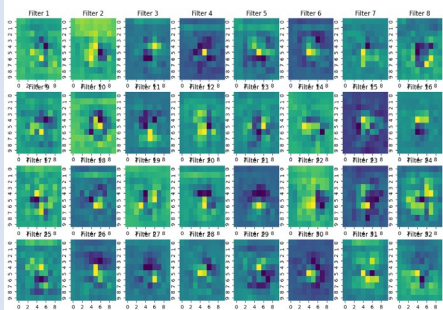


Saliency maps

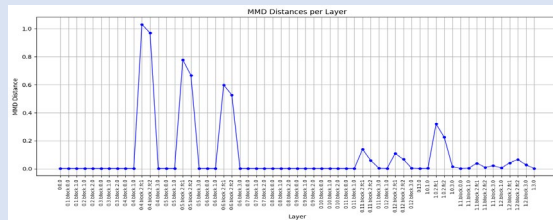


Model local behaviour

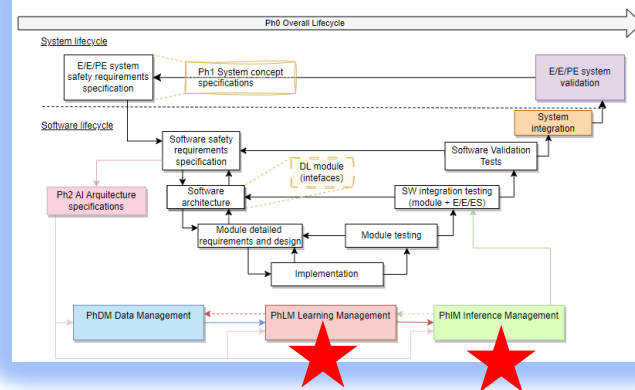
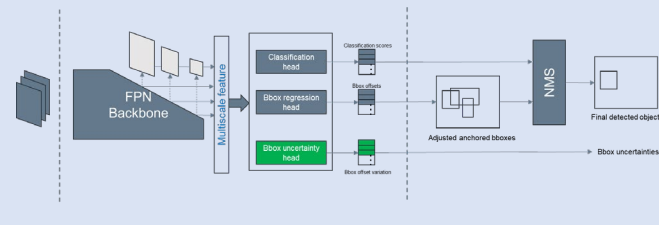
Activation Heatmaps



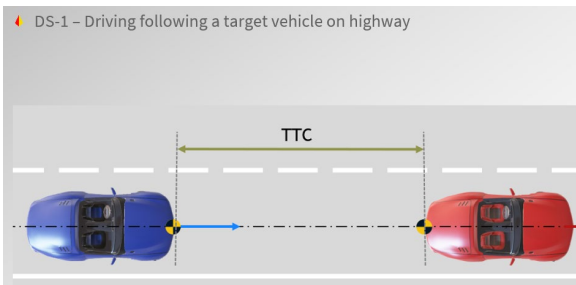
Model sensitivity to distribution shift



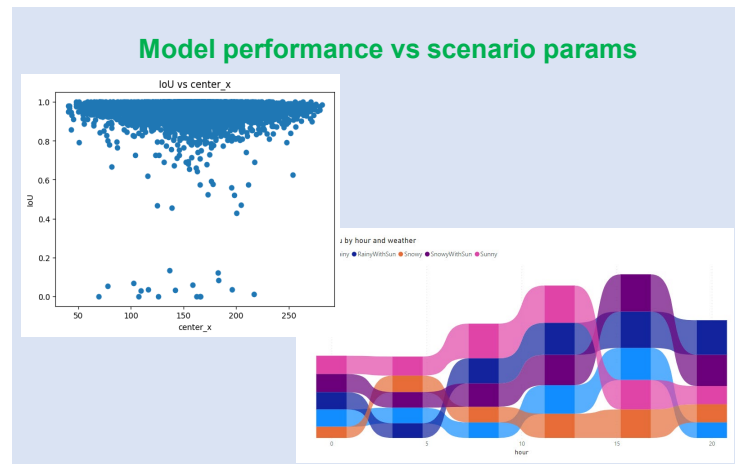
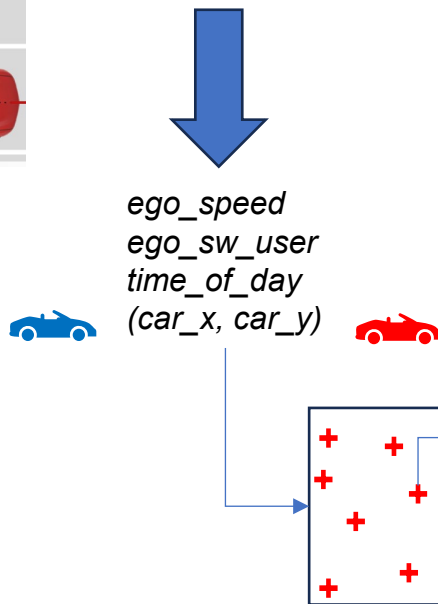
Uncertainty aware models



Support V&V strategy



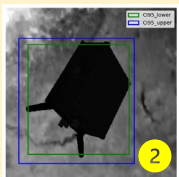
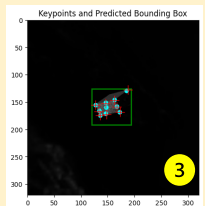
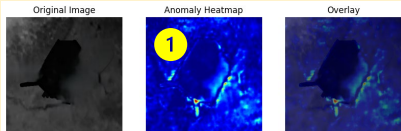
Scenario ID	Input parameters				Driver input				Target vehicle parameters				Impact				Discomfort parameters				Evaluation criteria			
	in. coord	in. coord	in. coord	in. coord	steering	brake	accel	other	in. coord	in. coord	in. coord	in. coord	impact	offset	count	error	light	right	condition	count	total	total	total	total
DS-1	1000.0	1000.0	1000.0	1000.0	0.0	0.0	0.0	0.0	1000.0	1000.0	1000.0	1000.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
DS-2	1000.0	1000.0	1000.0	1000.0	0.0	0.0	0.0	0.0	1000.0	1000.0	1000.0	1000.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
DS-3	1000.0	1000.0	1000.0	1000.0	0.0	0.0	0.0	0.0	1000.0	1000.0	1000.0	1000.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
DS-4	1000.0	1000.0	1000.0	1000.0	0.0	0.0	0.0	0.0	1000.0	1000.0	1000.0	1000.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
DS-5	1000.0	1000.0	1000.0	1000.0	0.0	0.0	0.0	0.0	1000.0	1000.0	1000.0	1000.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0



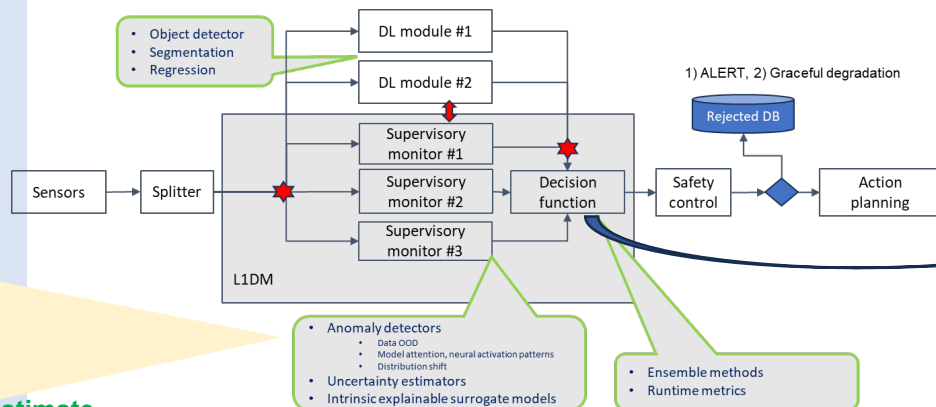
Scenario-based verification and validation

Operational and Monitoring stage

Anomaly detection



Safe surrogate model



Consolidated prediction & trustworthiness





Q & A





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.