Artificial Intelligence, Safety and Explainability

SAFEXPLAIN

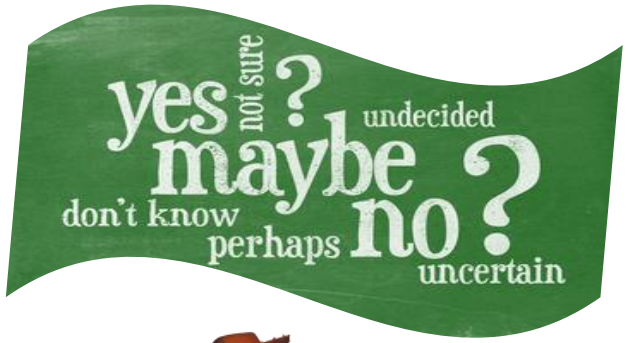
Safe and Explainable
Critical Embedded Systems based on AI

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Workshop CARS@EDCC2024
EDCC 2024, Leuven, Belgium
April 8th, 2024

www.ikerlan.es

AI, Safety and Explainability



"Who invited that kid?"

"Well, then, let's find out together!"

“To infinity and beyond!”

Ref.: [DISNEY24a][DISNEY24b]

AI & SAFETY

Pérez-Cerrolaza, J., et al. "**Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey**." 2023, ACM Computing Surveys

Agirre, et al. "**D2.2 DL safety architectural patterns and platform (PU)**." (dissemination level PU), April 2024

Ref.: [PER23]

Artificial Intelligence (AI)

- **AI:** "Set of methods or automated entities that together build, optimize and apply a model so that the system can, for a given set of predefined tasks, compute predictions, recommendations or decisions" [ISO 22989]

Types of AI:

- **Connectionists** (e.g., neural networks, Deep Neural Networks)
- **Bayesians.**
- **Symbolists.**
- **Analogizers.**
- **Optimizations.**

Ref.: [ISO22989][PER23]

Safety Standards & Technical Reports

Domain	Sector	Safety Standards			AI standards for safety systems
		Functional Safety	Heteronomous	Autonomous	
Transportation	Space	ECSS-Q-ST 30C/40C			
	Railway	EN 5012x	IEC 62290, IEC 62267		
	Avionics	ARP 4754, DO-178C DO-254	ASTM F3269-21		(ARP6983)
	Automotive	ISO 26262	ISO/PAS 21448	ISO 4804, ISO 5083, UL 4600	(ISO/AWI PAS 8800)
Industrial	Robotics	ISO 10218-1			
	Mining & earth moving machin.	ISO 19014	ISO 17757, ISO 16001, ISO 18758-2		
	Industrial Machinery	ISO 13849-1	(ISO TR 22100-5), (ISO 3691-4)		
	Agriculture	ISO 25119	ISO 10975, ISO 14897		
Generic		IEC 61508	(VDE-AR-E2842-61)		ISO TR 5469

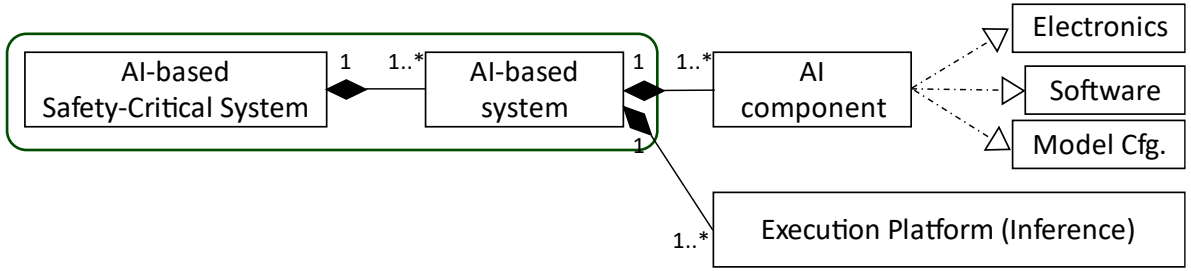
AI – Usage and Compliance

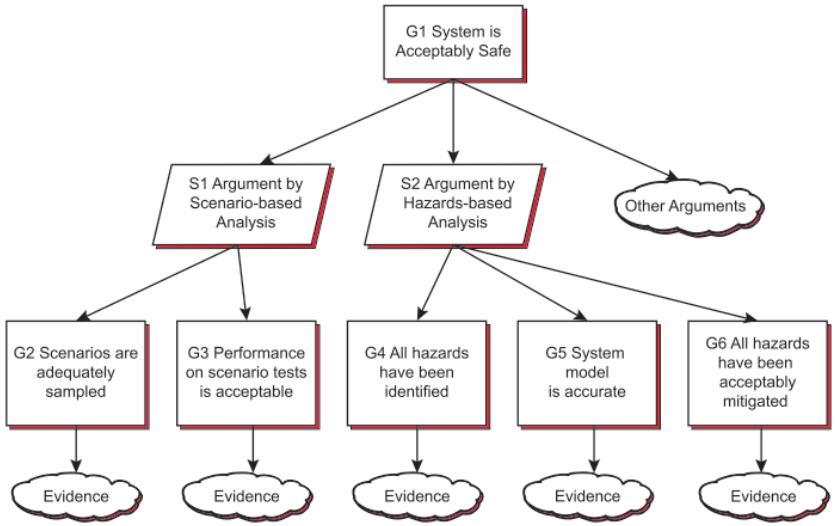
ISO TR 5469: Usage Level (UL) and Class

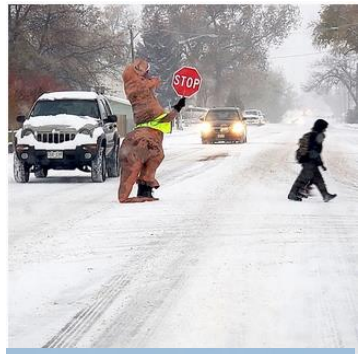
	Usage Level (UL)	Class I	Class II	Class III
PRODUCT	**A - Implements a safety function**	**Complies with safety standards**	**Does not comply with safety standards but compensation measures are sufficient**	Does not comply and compensation measures are not sufficient
PRODUCT	C - Implements a function that could interfere with safety functions			
PRODUCT	D - Implements a function that does not interfere with safety functions			
PROCESS	**B - Development process of a safety function**			

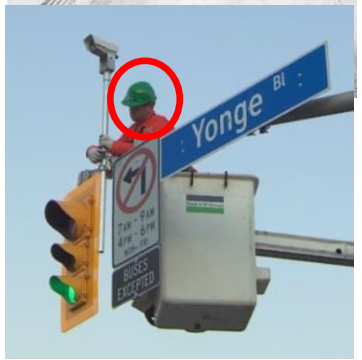
Product: AI-based Safety-Critical System

- **AI-based system**
- AI component
- Execution platform
- Training and tools





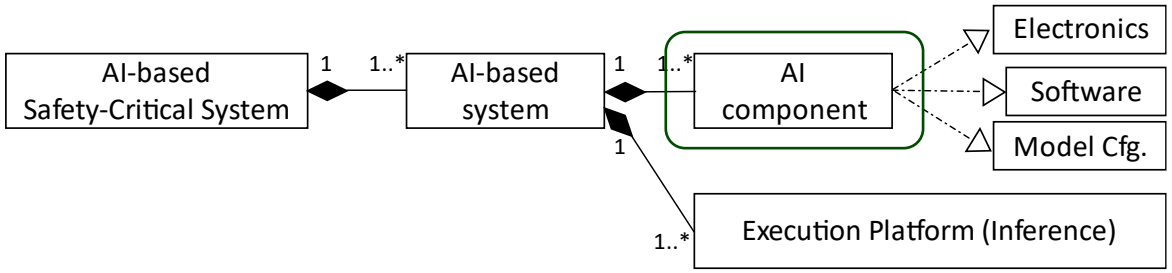




Product: AI-based Safety Critical-System

- AI-based system
- **AI component**
- Execution platform
- Training and tools

- **Connectionists**
- Symbolists.
- Optimizations.



Automatic Systems

- Formal verification
- Safety Bag/Safety Net

Heteronomous/Autonomous Systems

- (Formal verification)
- (Safety Bag/Safety Net)
- Safety Monitor, Safety Envelope, etc.

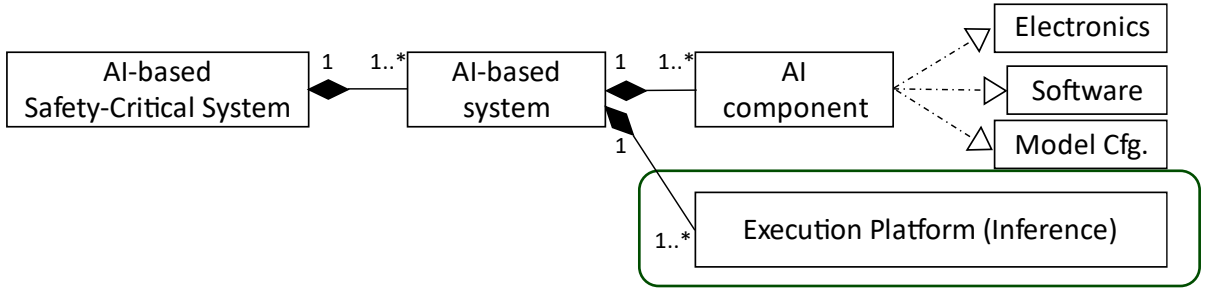
Explainability

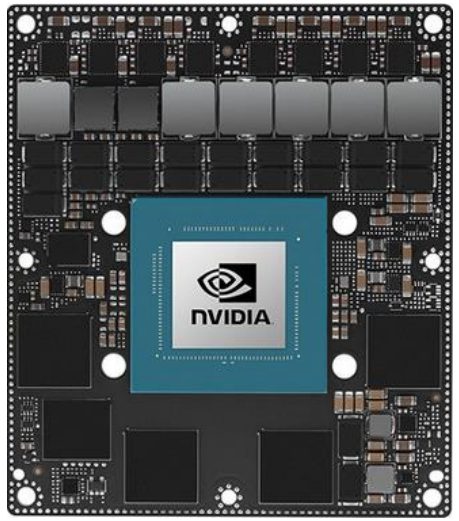
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

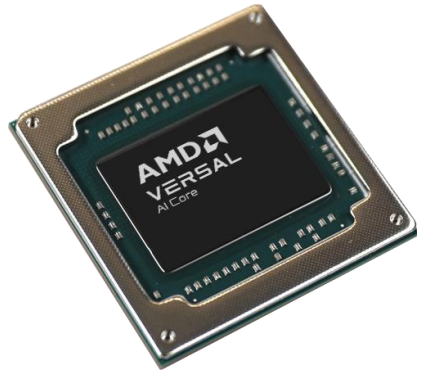
SAFEXPLAIN
Safe and Explainable Critical Embedded Systems based on AI

Product: AI-based Safety-Critical System

- AI-based system
- AI component
- **Execution platform**
- Training and tools



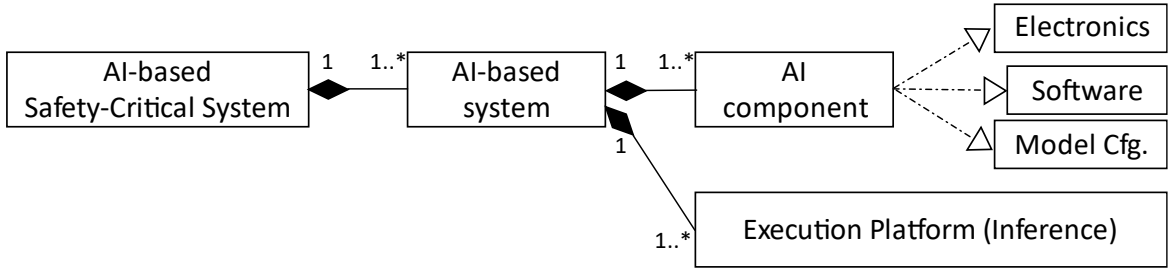




Ref.: [PER23][PER22]

Product: AI-based Safety-critical System

- AI-based system
- AI component
- Execution platform
- **Training and tools**



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

SAFEXPLAIN
Safe and Explainable Critical Embedded Systems based on AI

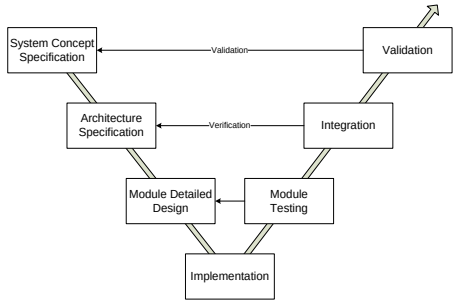
Product with online/learning adaptation

- **Techniques:**
  - Safety bag.
  - Safe adaptation.
  - Limited adaptation/actuation.
  - Library-based offline.

Usage Level (UL)	Domain	Description	Class	Type of AI	Technique
C	Avionics	Intelligent Flight Control System	II	Connectionist	Safety bag
A	Optimization	Gas turbine Generic adaptative control	I I, II	Generic (all)	Safe adaptation
A	Aerospace	Adaptative guidance	I, II	Connectionist	Limited adaptation
A, C	Industrial	ILC-based hydraulic machinery	I, II	Optimization	Limited actuation

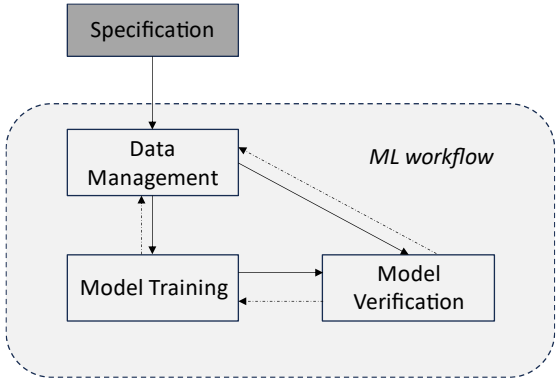
Process (UL B)

(a) Safety Engineering



Lifecycle (phase)	Usage Purpose	Type of AI
Specification	Analysis	Connectionist, Bayesian, Symbolic, Analogizer
Design	Design Optimization	Optimization
Test	MC/DC coverage	Connectionist, Optimization, Symbolic

(b) AI/ML Safety Engineering

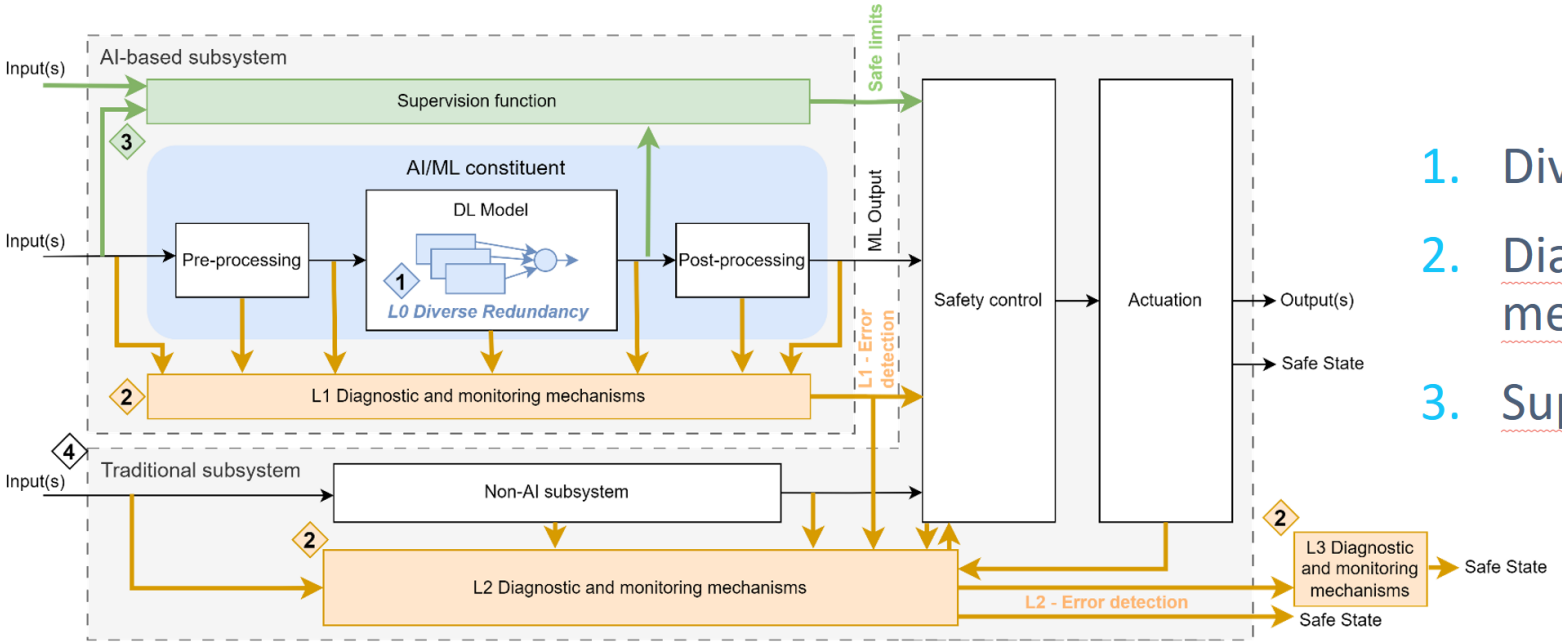


Lifecycle (phase)	Usage Purpose	Type of AI
Data Mgmt.	See model verification	
Model Training	AutoML	Reinforcement learning, Bayesian
Model Verification	Test definition automation	All
	Test classification automation	Connectionist, Symbolists
	Fault injection	Bayesian
	Rule extraction	Symbolist
	Quantify Uncertainty	Bayesian

T2.3 DL Safety Architectural design Patterns

Safety pattern: Generic solutions for commonly recurring design problems with the aim of simplifying and standardizing the design process

- Extend and combine common patterns from traditional Functional Safety (FUSA) to address the new challenges brought by DL-based approaches in complex HW/SW platforms



1. Diverse Redundancy
2. Diagnostic and monitoring mechanisms
3. Supervision function

SAFEXPLAIN

5 Feb 2024

13

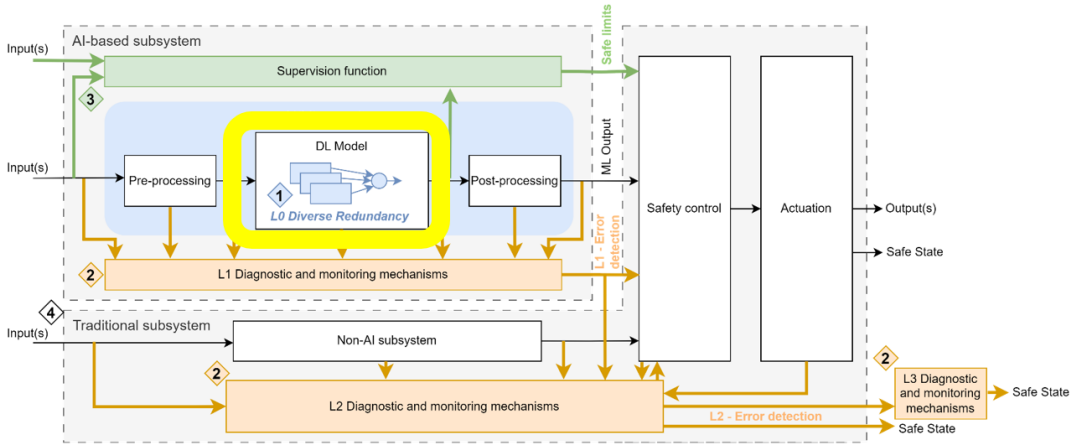
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

SAFEXPLAIN
Safe and Explainable Critical Embedded Systems based on AI

Ref.: [SAF24]

T2.3 DL Safety Architectural design Patterns

- L0 Diverse Redundancy



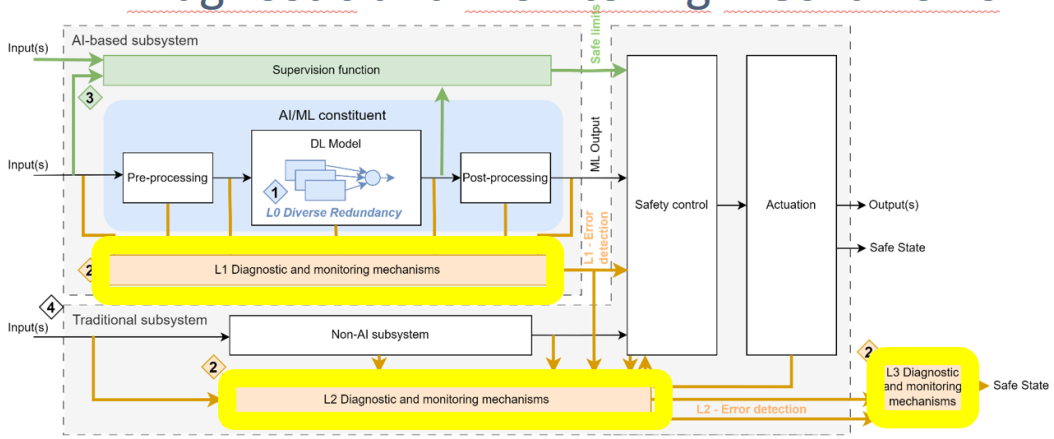
On Neural Networks Redundancy and Diversity for Their Use in Safety-Critical Systems

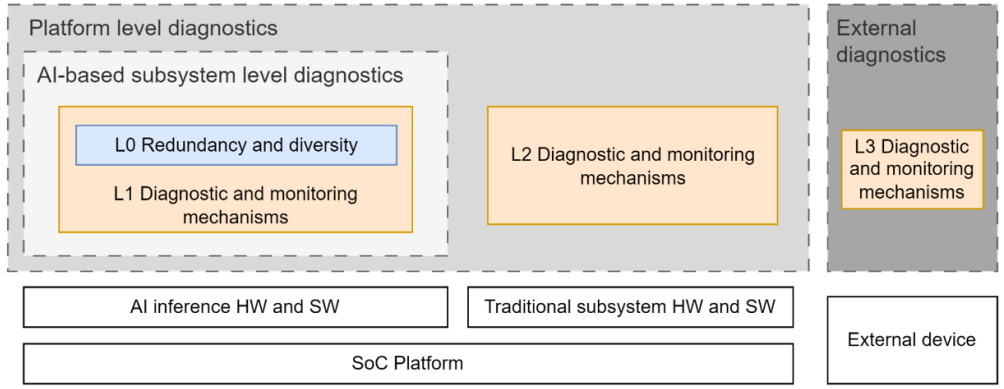
- Inference Platform diversity
  - Inputs (diverse cameras, sensors, input image flips…)
  - Processing resources (accelerators, CPUs…)
  - OS, Execution framework (e.g., TF lite, pytorch, darknet…)
- DL model Development diversity
  - Model Architecture
  - Hyperparameters
  - Training datasets
  - Training process
  - Training platform
- Concept diversity: different problem formulation with same final goal
  - Object detection vs object part detection
  - Object detection vs obstacle free path detection

Ref.: [BRA23]

T2.3 DL Safety Architectural design Patterns

- Diagnostic and monitoring mechanisms





L0 – Diverse Redundancy

L1 – AI-based subsystem level diagnostics: runtime errors or model insufficiencies and anomalies on the AI subsystem and the elements required for its execution (e.g., accelerators, AI frameworks, etc.)

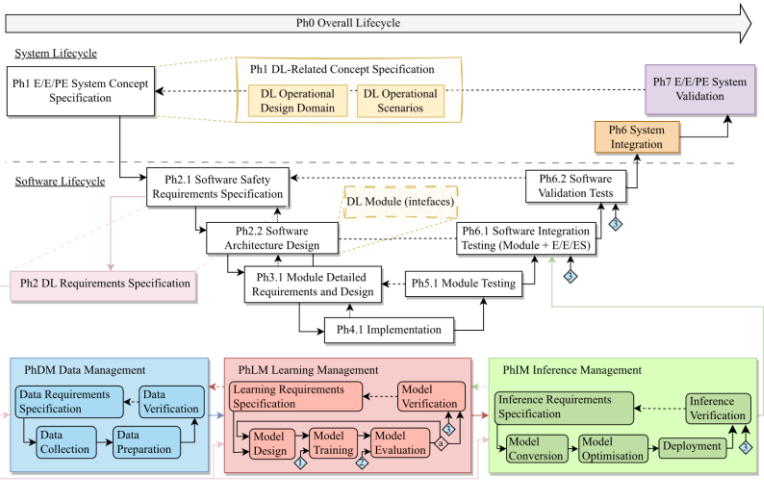
L2 – Platform level diagnostics: runtime errors on additional platform HW and SW components following traditional functional safety practices and diagnostics techniques (e.g., memory self-tests, freedom from interference at platform level…)

L3 – External diagnostics

Based on the Standardized E-Gas Monitoring concept (Automotive domain)

T2.1 Safety lifecycle: DL spec., design and implementation

- IEC 61508 traditional functional safety lifecycle (Software V-model) + AI lifecycle



Procedure

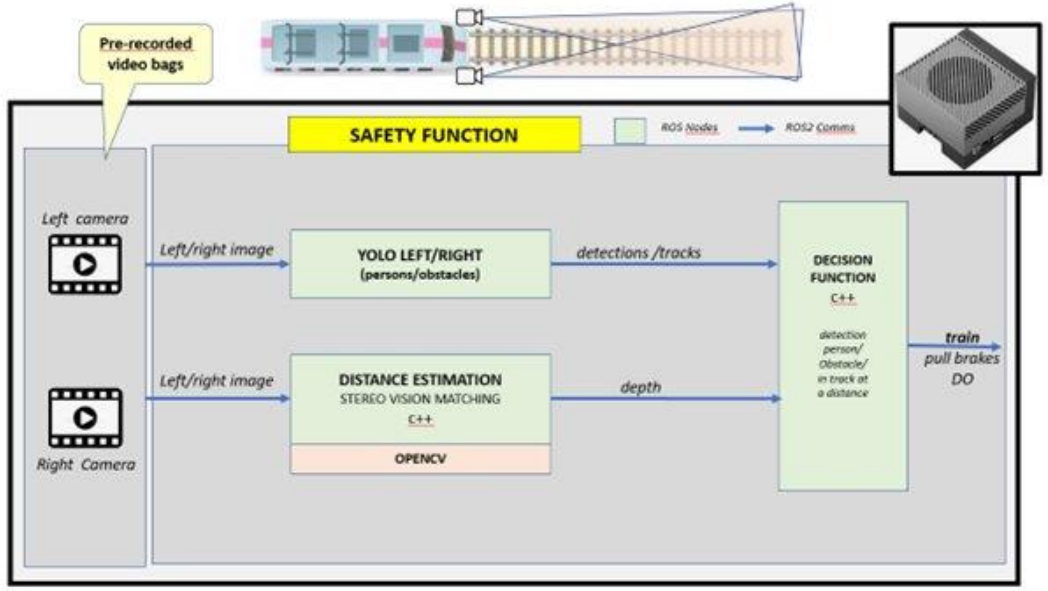
Guidelines

Templates

Internal review checklists

WP2 ⟷ WP3

Railway case study



AI & SAFETY & EXPLAINABILITY

Thanh Hai Bui, et al. "**D3.1 Specifiability, explainability, traceability, and robustness proof-of-concept and argumentation**." (dissemination level PU), April 2024

Ref.: [BUI24]

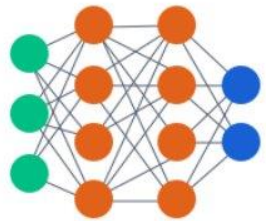
Explainability

(Traditional) Safety Engineering

IA/ML Safety Engineering

Explainability:

- "Extent to which an ML system can provide an explanation about a decision in a form understandable by a human" [JEN20]
- "the ability to generate human understandable reasons for the model predictions along with the internal workings of the system" [BUI24]



Ref.: [BUI24][IEC61508][IET06][JEN20]

Trustworthiness - Explainability

Engineering dimension

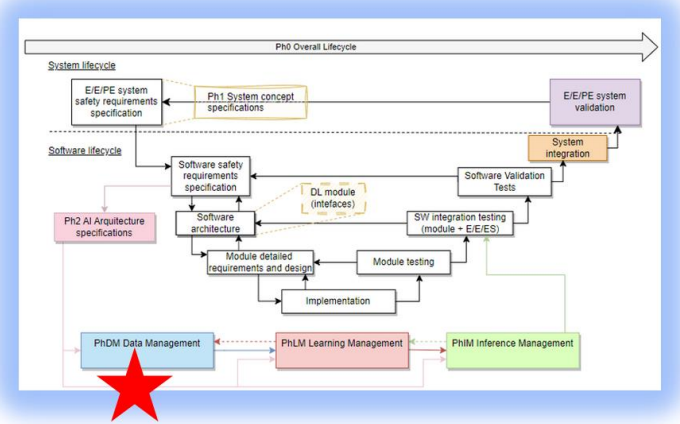
Ethical dimension

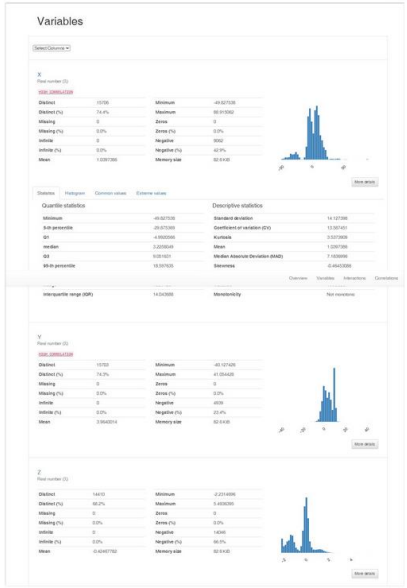
Legal dimension

T3.1 DL specification

AI-FSM Data Management

- Purposes
  - Assess/Verify if data is representative of ODD/scenarios
  - Safety driven data metrics
  - Baseline “Known” data
  - Data reports
- Data explainer techniques
  - Data profiling, statistical info
  - Important instances/prototypes



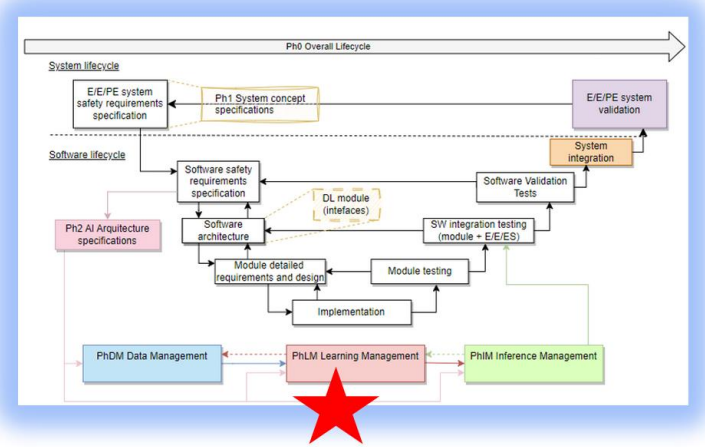


Ref.: [BUI24]

T3.1 DL specification

AI-FSM Learning Management

- Purposes
  - Safety driven performance/robustness metrics
  - Explainable by design, explanations as evidence
  - Reduce epistemic uncertainty
  - Baseline model “normal” behaviours
- Model explainers
  - Disentanglement, e.g. correlation performance/ ODD parameters
  - Decomposition
  - Feature relevance
  - Activation patterns
  - Local explanations to support diagnosis



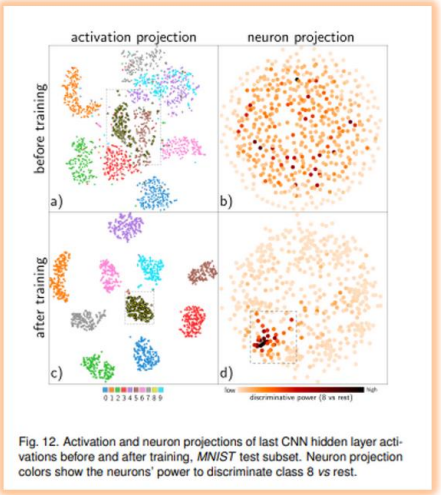


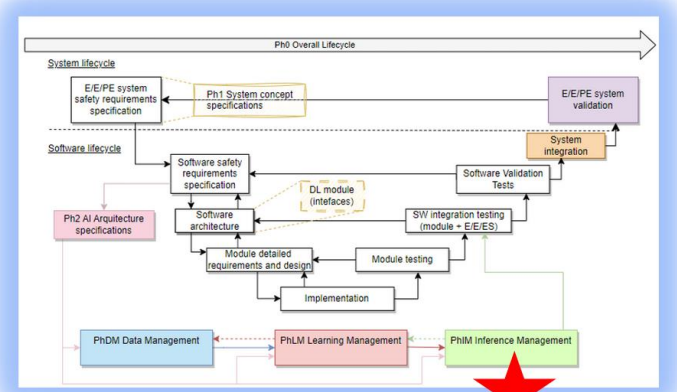
Fig. 12. Activation and neuron projections of last CNN hidden layer activations before and after training, *MNIST* test subset. Neuron projection colors show the neurons' power to discriminate class 8 *vs* rest.

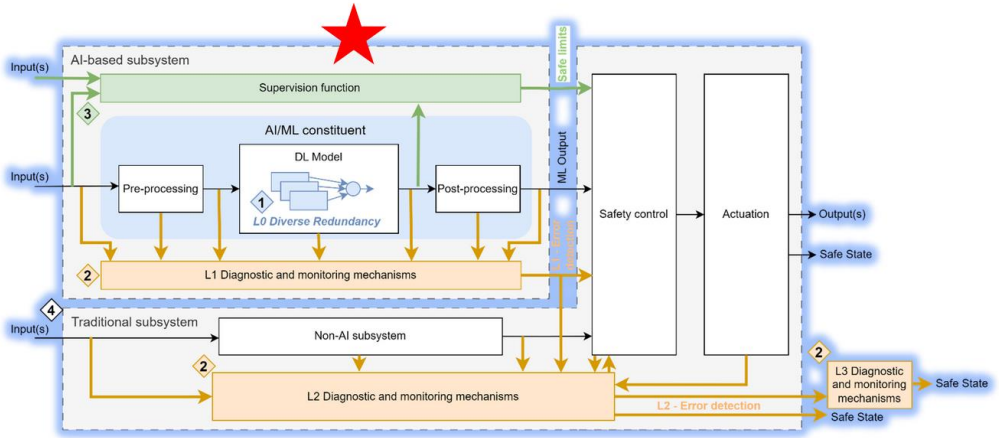
Ref.: [BUI24]

T3.1-3.2 DL specification and design

AI-FSM Inference Management & Operational XAI

- Purposes
  - V&V: Scenario-based tests, evidences, result interpretation
  - Monitor model insufficiency & anomalies
  - Extract runtime metrics
  - Diagnose rejected cases
- Model/Data explainers
  - Disentanglement/surrogate
  - Runtime supervisor
    - Out Of Distribution detections
    - Higher SIL surrogate models
    - Uncertainty estimators



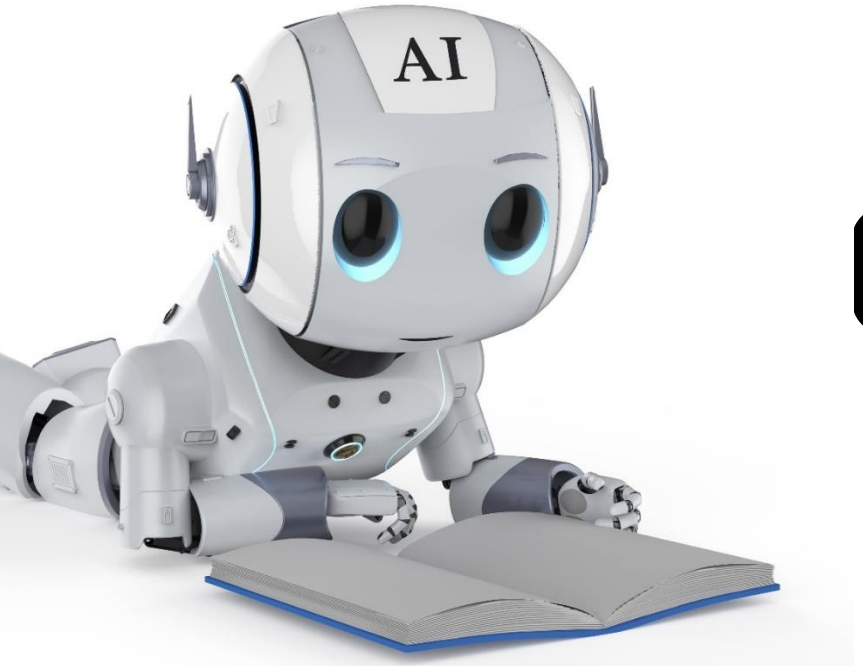


SAFEXPLAIN

5 Feb 2024

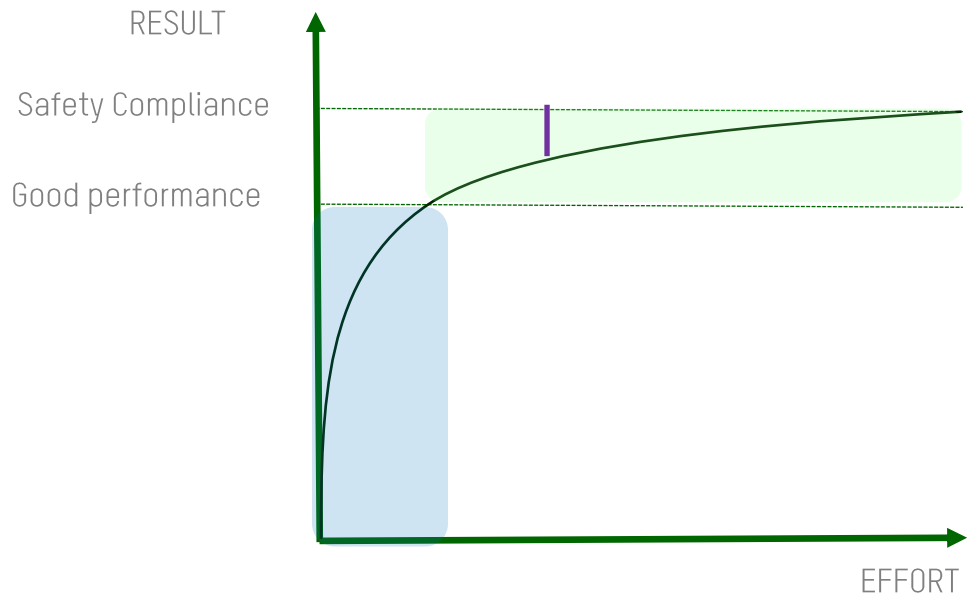
17

Ref.: [BUI24]



CONCLUSIONS

Conclusions



- Railway: The safety bag technique is already used to perform safe automated decision-making (A1) for SIL4 railway interlocking.
- Automotive: The latest ADAS systems already use decision decision-making safety functions that require human supervision (A2).
- Avionics: Research prototype DAL C runaway sign classifier [DIM23]
- But still a significant pending effort to:
  - Define generic AI techniques and processes for developing safety-critical systems: "How things can be done" and "How things should be done"
  - Explainability is a critical attribute
  - Formalize AI and heteronomous/autonomous safety standards.
  - Settle industry best practices.

"Move fast and fix things"

Ref.: [DIM23][PER23]

THANK YOU!

Jon Perez Cerrolaza
jmperez@ikerlan.es

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

References

[BRA23] Brando, A., I. Serra, E. Mezzetti, F.J. Cazorla, J. Perez-Cerrolaza, and J. Abella, On Neural Networks Redundancy and Diversity for Their Use in Safety-Critical Systems. Computer, 2023. 56(5): p. 41-50.

[BUI24] Thanh Hai Bui, et al. "D3.1 Specifiability, explainability, traceability, and robustness proof-of-concept and argumentation." (dissemination level PU), April 2024

[CARS24] https://conf.laas.fr/cars/cars2024.html

[CBC] https://www.cbc.ca/news/canada/toronto/smart-traffic-signals-1.4417573

[DIM23] Dmitriev, K., J. Schumann, I. Bostanov, M. Abdelhamid, and F. Holzapfel, Runway Sign Classifier: A DAL C Certifiable Machine Learning System. 2023. 1-8.

[DISNEY24a] https://news.disney.com/ultimate-list-toy-story-quotes

[DISNEY24b] https://toystory.disney.com/

[HER19] https://www.reporterherald.com/2019/10/31/dinosaur-big-brown-bear-help-children-cross-the-street-at-berthoud-elementary/

[IEC61508] [1] IEC 61508-1: Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 1: General requirements, IEC, 2010

[IET06] Competence Criteria for Safety-Related System Practitioners, IET, 2006

[JEN20] Eric Jenn et al. 2020. Identifying challenges to the certification of machine learning for safety critical systems. In 10th European Congr. on Embedded Real-Time Syst. (ERTS).

[ISO22989] ISO/IEC 22989 - Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

[PER22] Perez-Cerrolaza, J., J. Abella, L. Kosmidis, A.J. Calderon, F.J. Cazorla, and J.L. Flores, GPU Devices for Safety-Critical Systems: A Survey. ACM Comput. Surv., 2022. 55(7): p. 1-37.

[PER23] Perez-Cerrolaza, J., J. Abella, M. Borg , C. Donzella, J. Cerquides, F. Cazorla, C. Englund, M. Tauber, G. Nikolakopoulos, and J.L. Flores, Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. ACM Comput. Surv., 2023.

[SAF24] https://safexplain.eu/

[THO19] Thomas, S. and D. Vandenberg (2019). "Harnessing Uncertainty in Autonomous Vehicle Safety." Journal of System Safety 55(2).