SAFEXPLAIN

Safe and Explainable Critical Embedded Systems based on AI

Functional Safety on AI-based critical systems

Irune Yarza

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

AEiC 2024

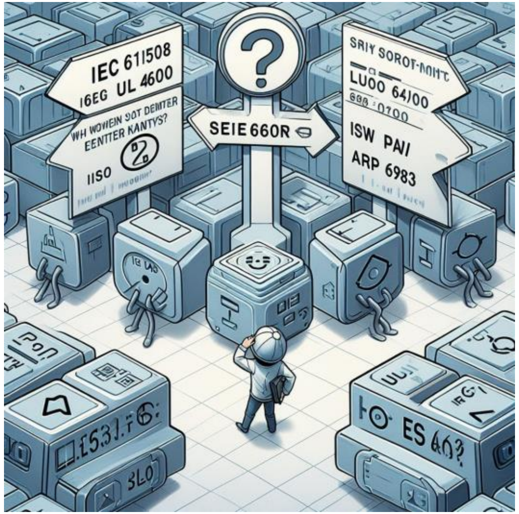
Enabling the use of AI in Safety-Critical Systems

14 June 2024, Barcelona, Spain

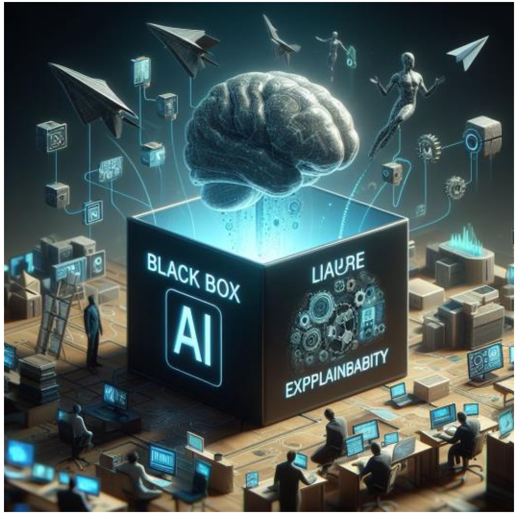
Funded by the European Union

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

AI Safety



Which standards
should we follow?



Can we make AI
explainable?

Can we make AI
safe?

NOTE: Images were generated usign Copilot

Safety Standards & Technical Reports

WIP 2023-06-26

Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI ARP6983

ISO/CD PAS 8800

Road Vehicles

Safety and artificial intelligence

Status : Under development

ISO/TR 22100-5:2021

Safety of machinery

Relationship with ISO 12100

Part 5: Implications of artificial intelligence machine learning

Status : Published

ISO/IEC TR 5469:2024

Artificial intelligence

Functional safety and AI systems

Status : Published

Transportation	Railway	IEC 62290, IEC 62267	
	DO-178C	ASTM F3269-21	(ARP6983)
		ISO/PAS 21448 / ISO 4804, ISO 5083	(ISO/AWI PAS 8800)
Industrial			ISO/TR 22100-5
		ISO 10975, ISO 14897	
General		(VDE-AR-E2842-61)	ISO TR 5469

AI standards for safety systems

Jon Perez-Cerrolaza et al, *"Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: a Survey", ACM Computing Surveys, 2023:* https://doi.org/10.1145/3626314

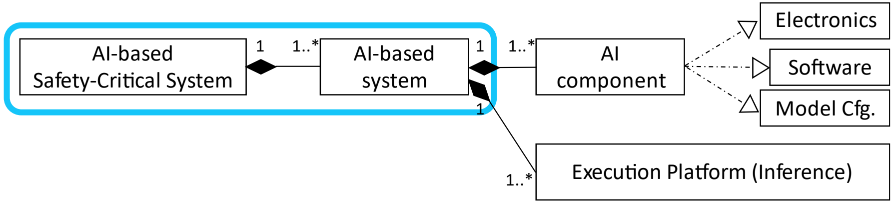
AI – Usage and Compliance

ISO TR 5469: Usage Level (UL) and Class

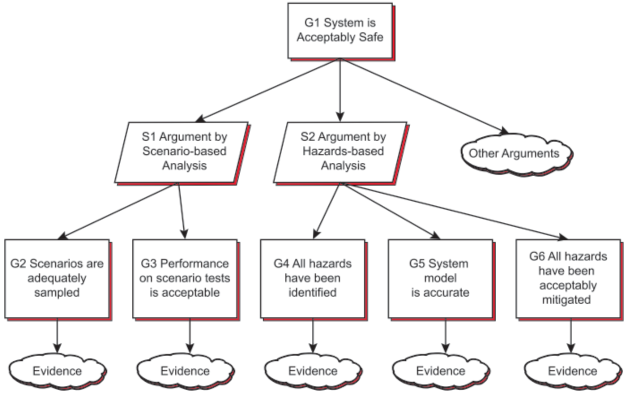
	Usage Level (UL)	Class I	Class II	Class III
PRODUCT	**A - Implements a safety function**	**Complies with safety standards**	**Does not comply with safety standards but compensation measures are sufficient**	Does not comply and compensation measures are not sufficient
PRODUCT	C - Implements a function 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



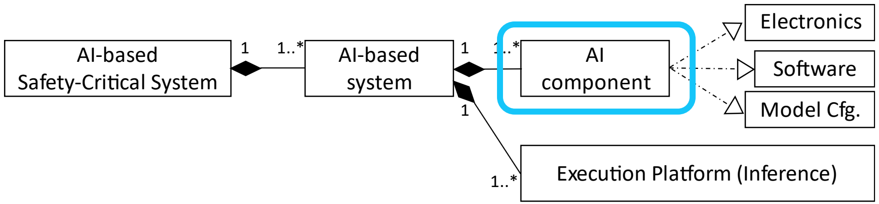
Safety assurance cases



UL4600 – Safety cases for autonomous systems

Product: AI-based Safety Critical-System

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



Automatic Systems (closed environment)

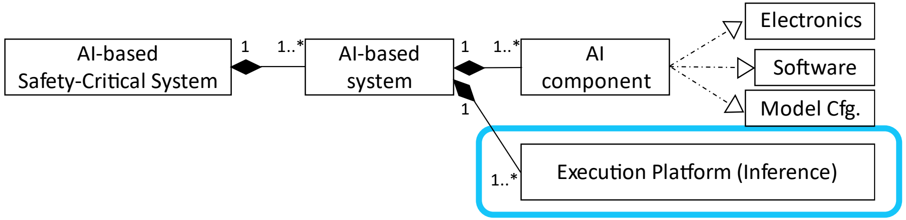
- Formal verification
- Safety Bag/Safety Net

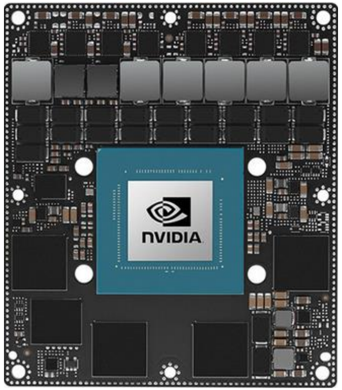
Heteronomous/Autonomous Systems (open/semi-open environment)

- (Formal verification)
- (Safety Bag/Safety Net)
- Safety Monitor, Safety Envelope (ODD)…

Product: AI-based Safety-Critical System

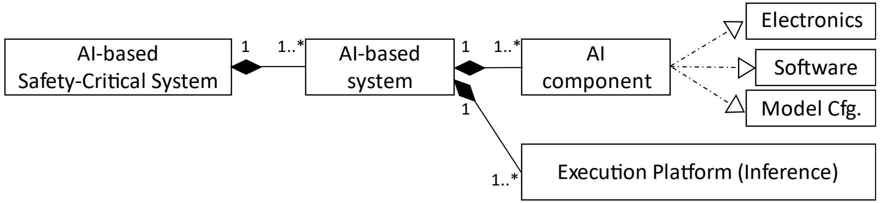
- AI-based system
- AI component
- **Execution platform**
- Training and tools



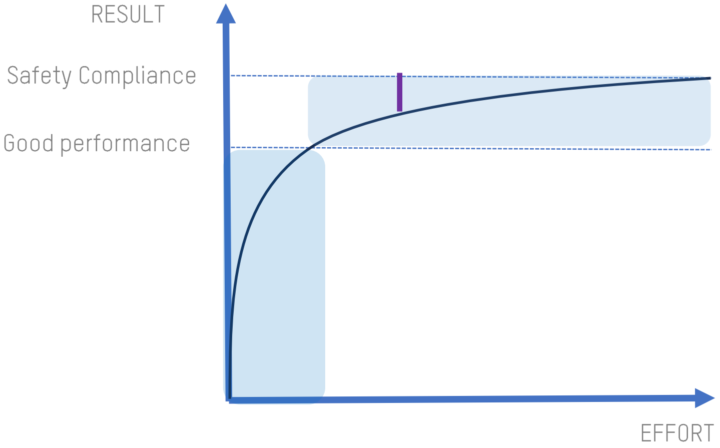


Product: AI-based Safety-critical System

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



SoA analysis conclusions

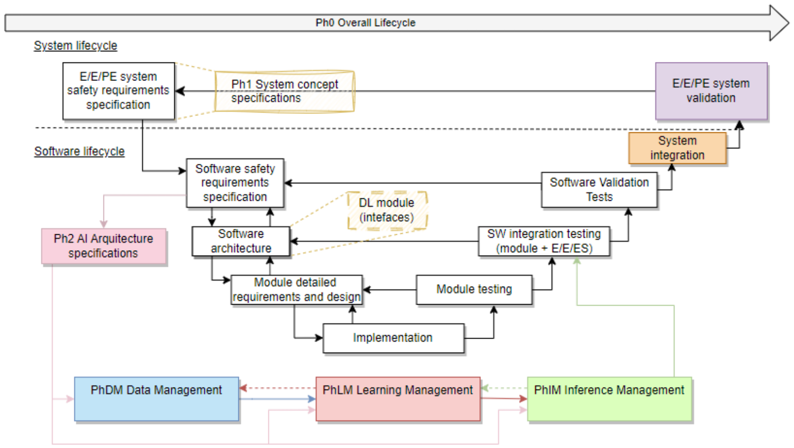


- Railway: The safety bag technique is already used to perform safe automated decision-making (A1) for SIL4 railway interlocking.
- Automotive: The latest ADAS systems already use decision-making safety functions that require human supervision (A2).
- But still a significant pending effort to:
  - Formalize AI and heteronomous/autonomous safety standards.
  - Define generic AI techniques and processes for developing safety-critical systems: "How things can be done" and "How things should be done"

Ref.: [DIM23][PER23]

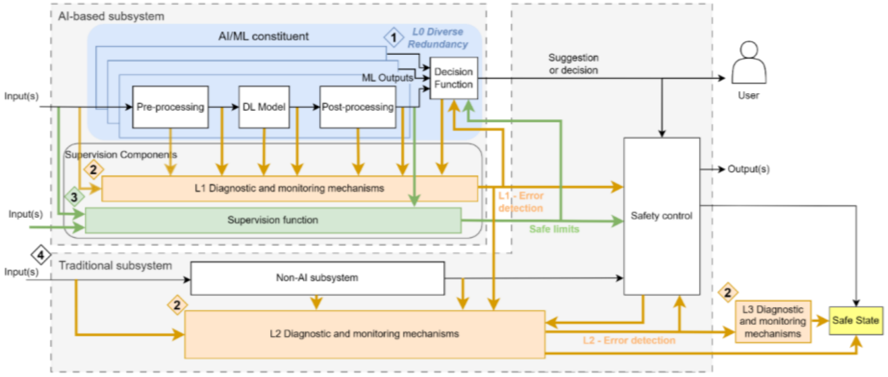
SAFEXPLAIN contributions

AI-FSM



Reference software development process

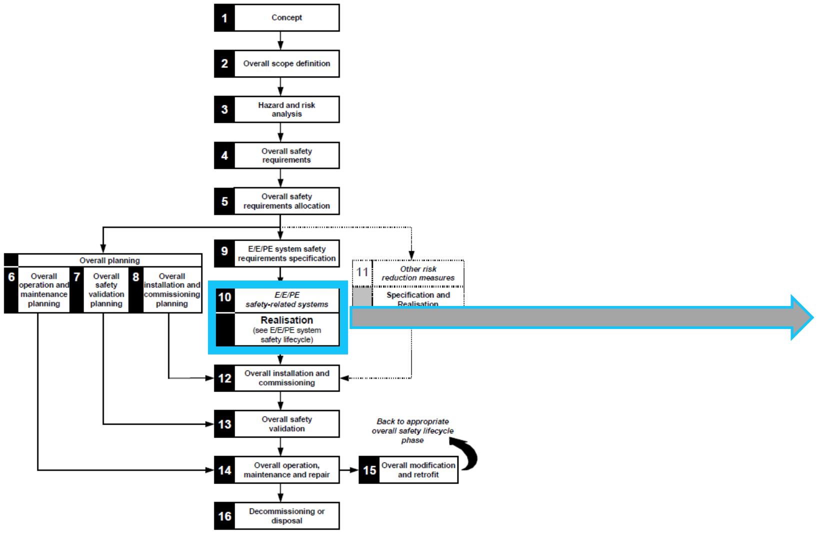
AI Safety Architecture Pattern

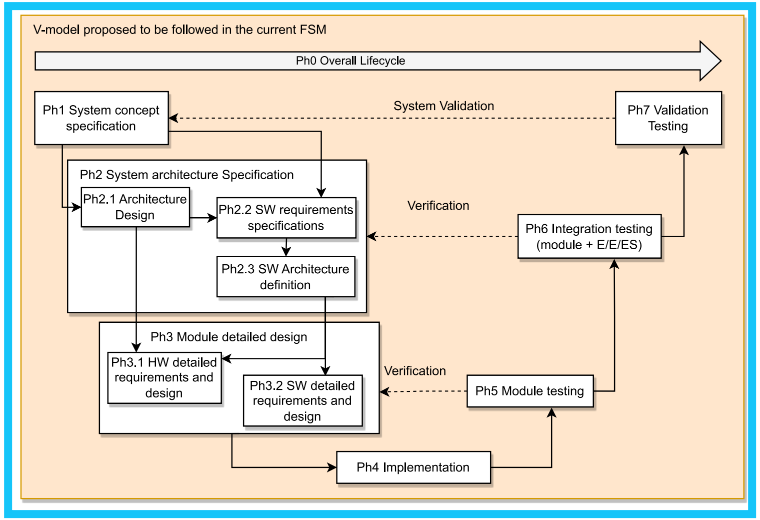


Reference safety architecture

AI-FSM – Context

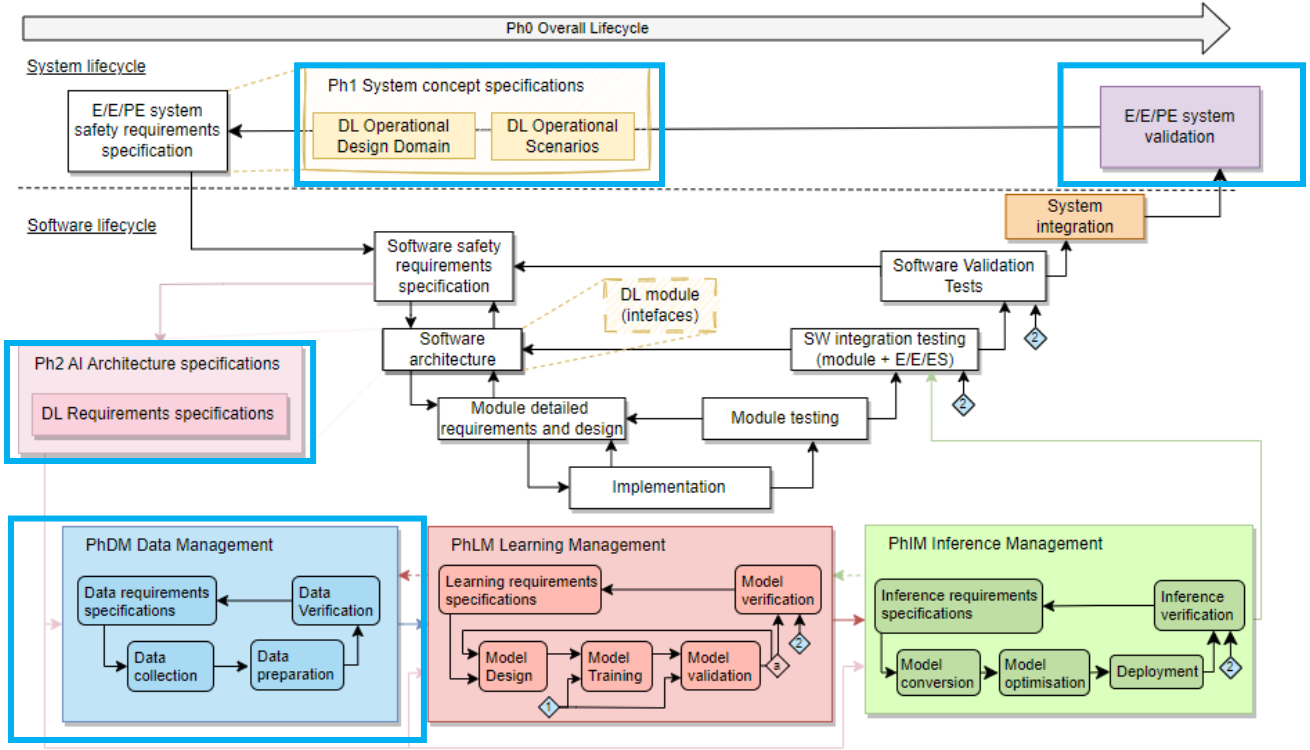
Functional Safety Management (FSM): encompasses all essential activities throughout the Functional Safety lifecycle phases, as mandated by IEC 61508-1. FSM is designed to **prevent errors during specification, design, development, manufacturing, and commissioning**.

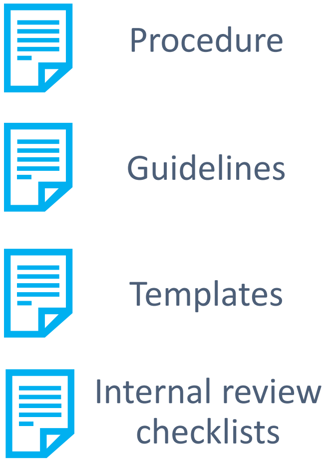




Functional safety lifecycle including AI (AI-FSM)

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

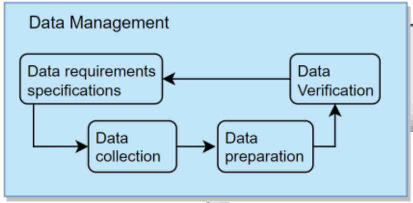




SAFEXPLAIN

AI-FSM – Overview

- PhDM Data Management



SAFEXPLAIN
Safe and Explainable Critical Embedded Systems based on AI

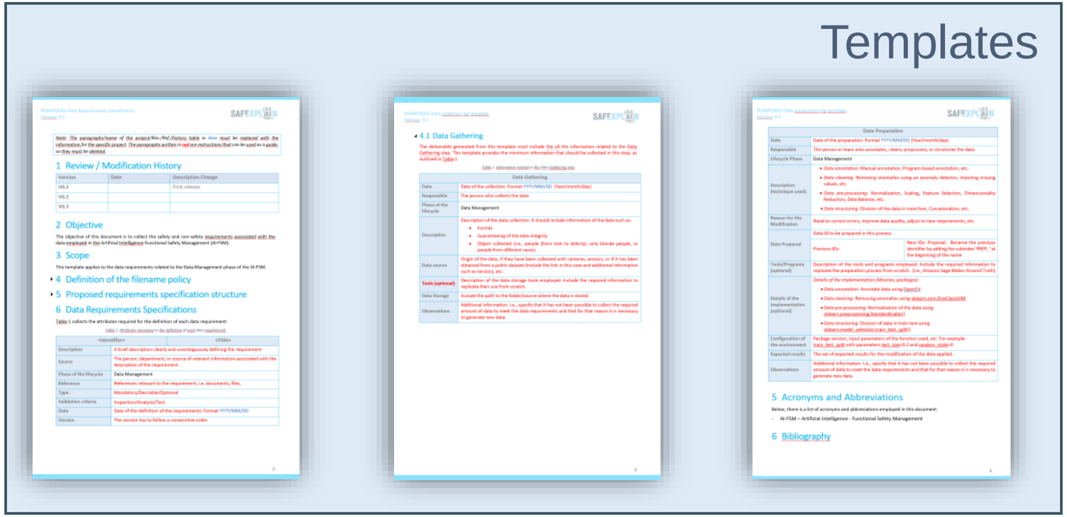
PhDMG001 Data Management Guideline

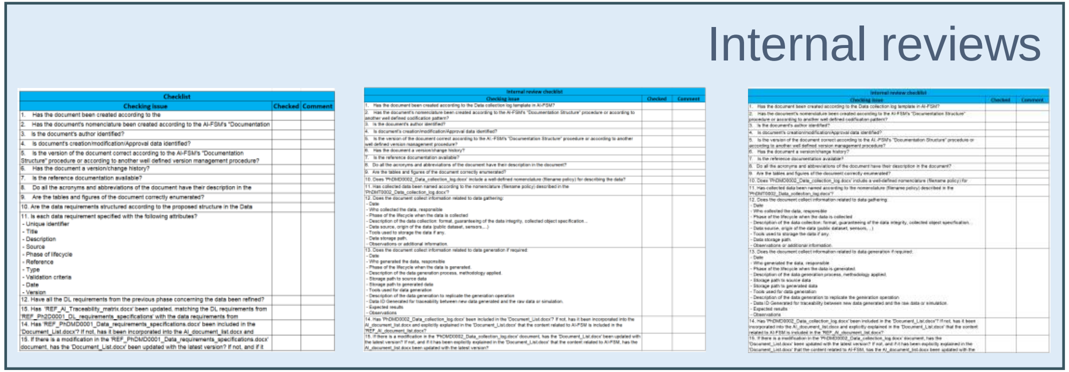
Version 0.1

Documentation Information

Contract Number	101069595
Project Website	www.safexplain.eu
Contractual Deadline	DD.MM.YYY
Dissemination Level	PU or SEN -see DoA
Nature	R or OTHER -see DoA
Author	Irune Agirre, Javier Fernández
Modified by	Lorea Belategi
Reviewed by	Name (Partners' short name)
Approved by	Name (Partners' short name)
Keywords	AI, Functional safety, FSM, Explainability

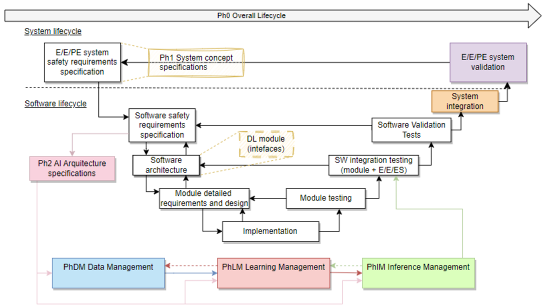
This project has received funding from the European Union's Horizon Europe programme under grant agreement number 101069595.





Safety architecture patterns

Reference software development process



Need for runtime safety mechanisms to deal with:

- *Random and residual systematic faults*
- *HW / SW platform complexity: integration problems (e.g., determinism, interferences on mixed-criticality approaches, use of resources…)*
- *DL model insufficiencies*
- *Support DL explainability*

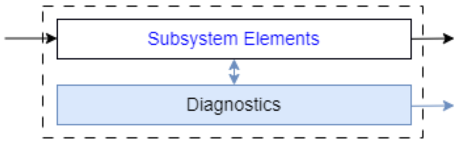
…

GOAL: To provide reference safety architecture patterns for the adoption of DL in safety-critical systems with varying safety requirements

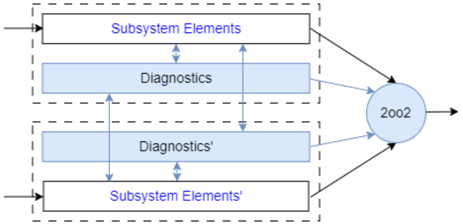
Safety architecture patterns – Overview

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

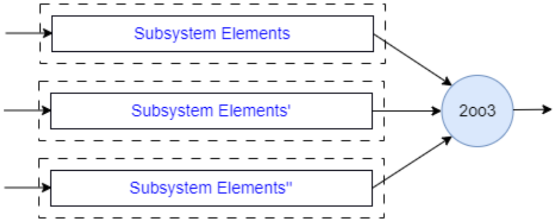
Common examples:



Single channel with diagnostics (1oo1D)



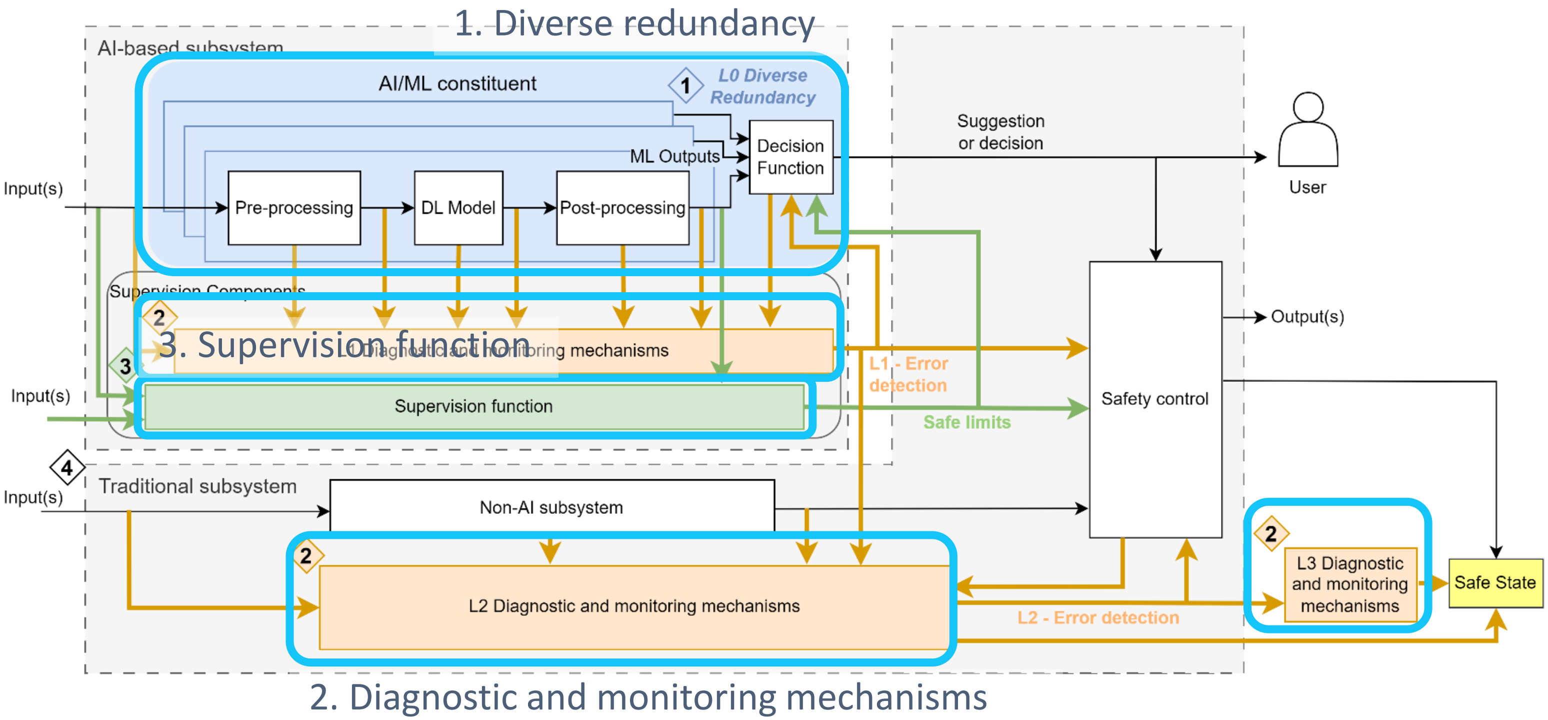
Dual channel with diagnostics (2oo2D)



Triple Module Redundancy (TMR) with majority voter (2oo3)

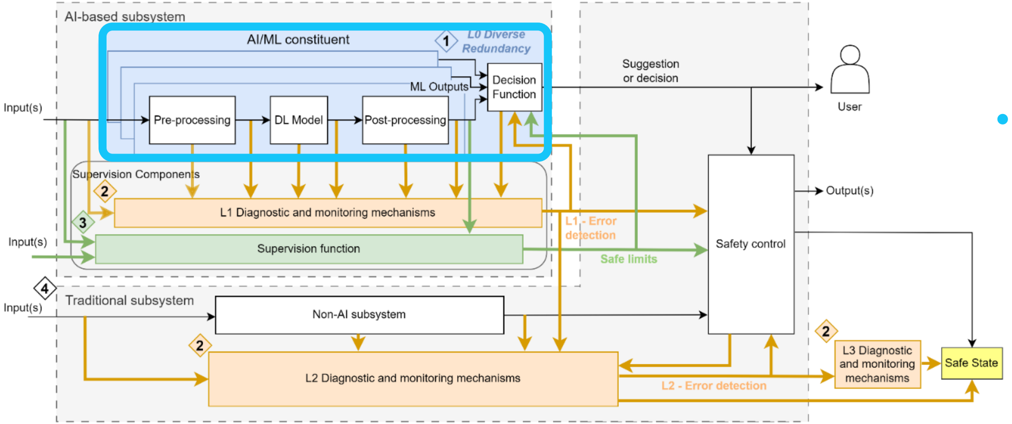
Extend and combine common patterns from traditional Functional Safety (FUSA) to address the new challenges brought by DL-based approaches in complex HW/SW platforms

Reference safety architecture



Reference safety architecture

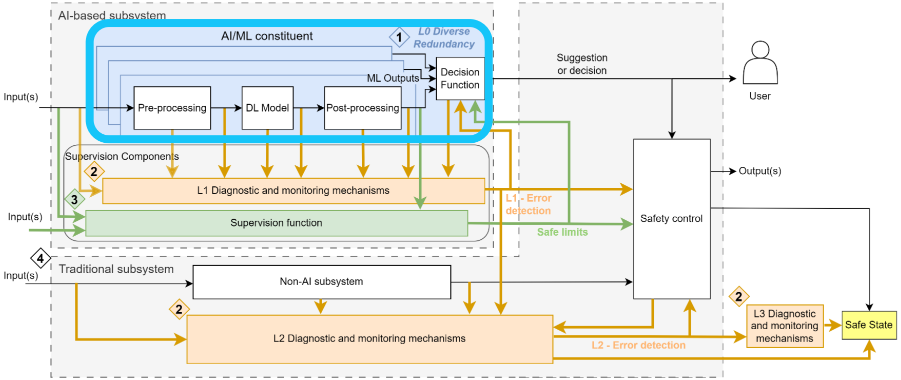
- L0 Diverse Redundancy

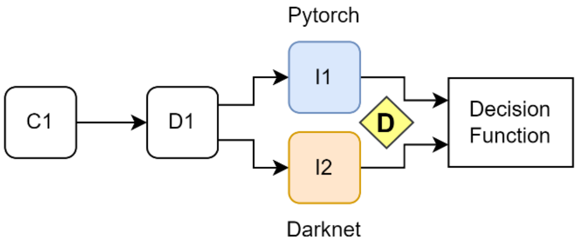


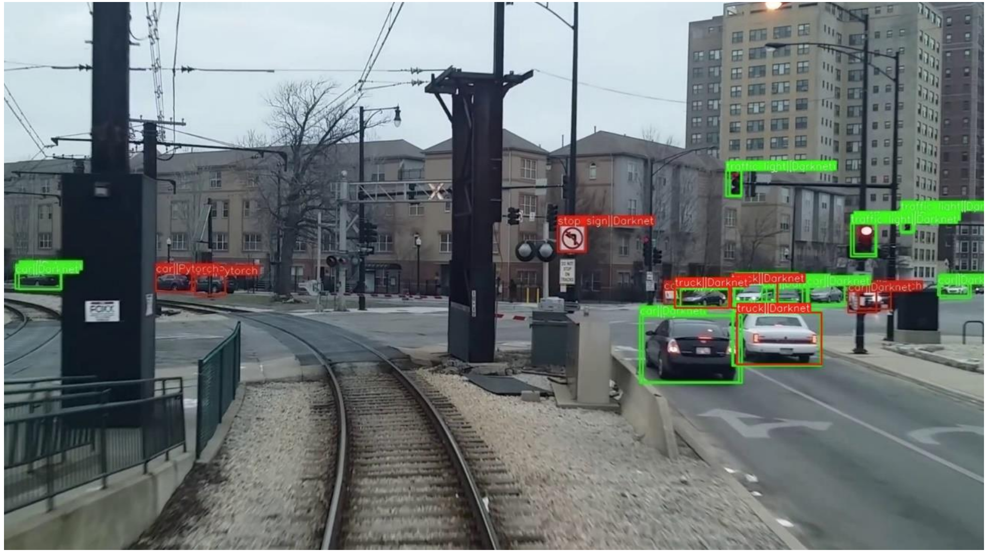
- Inference Platform diversity
  - Inputs (diverse cameras, sensors, input image flips…)
  - Processing resources (accelerators, CPUs…)
  - OS, Execution framework (e.g., TF lite, pytorch, darknet…)
  - …
- DL model Development diversity
  - Model Architecture
  - Hyperparameters
  - Training datasets
  - 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
  - …

Reference safety architecture

- L0 Diverse Redundancy – Inference platform diversity using diverse redundant frameworks (i.e., Pytorch and Darknet).



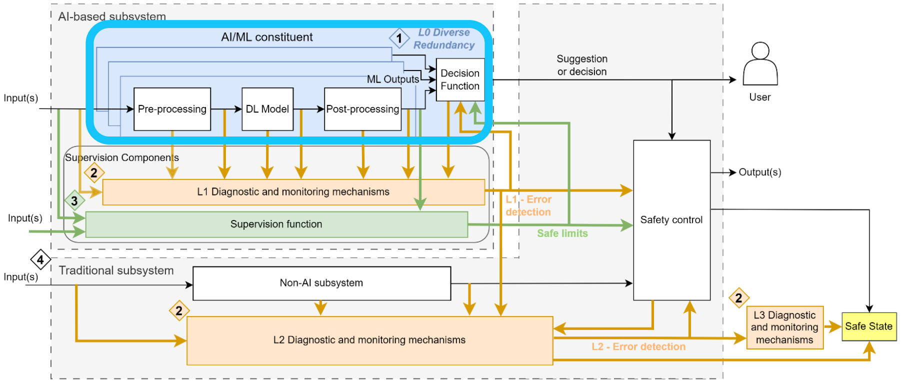




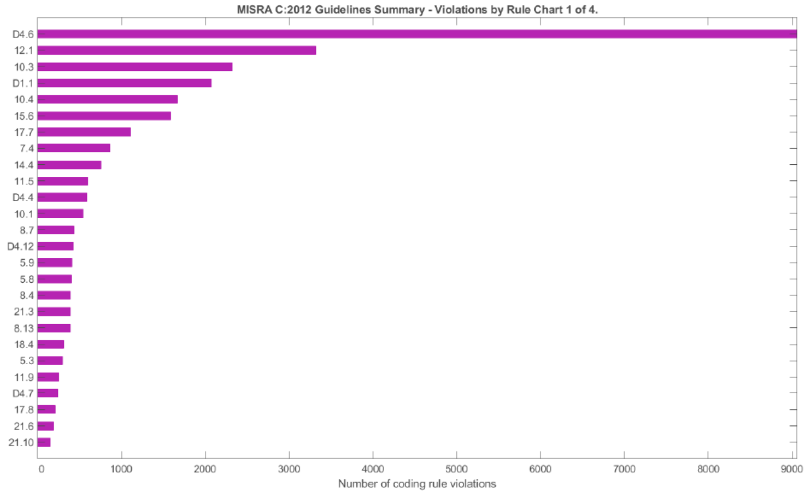
SAFEXPLAIN

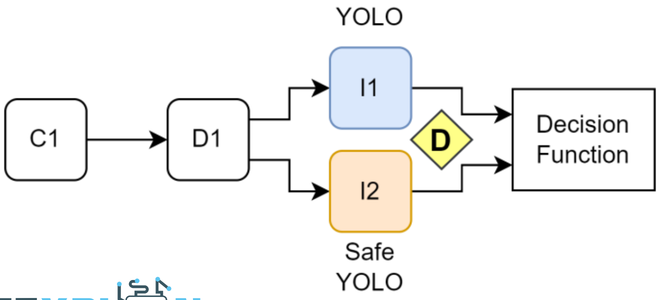
Reference safety architecture

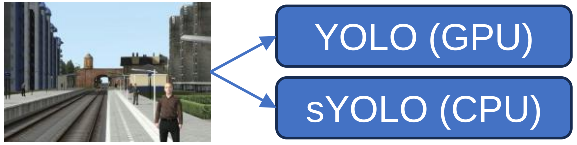
- L0 Diverse Redundancy – Inference platform diversity using diverse redundant frameworks (i.e., YOLO and SafeYOLO).



MISRA C:2012 Guidelines Summary - Violations by Rule

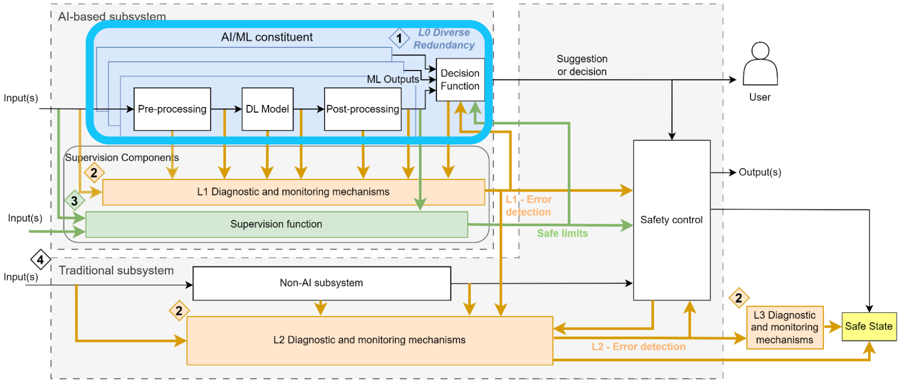


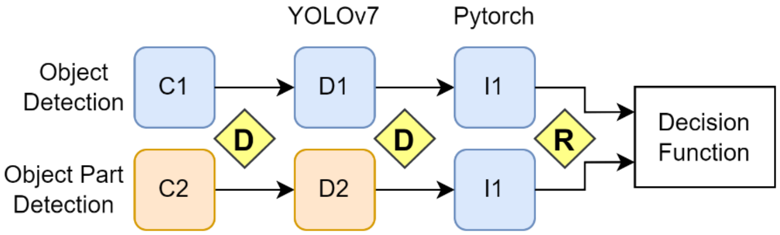


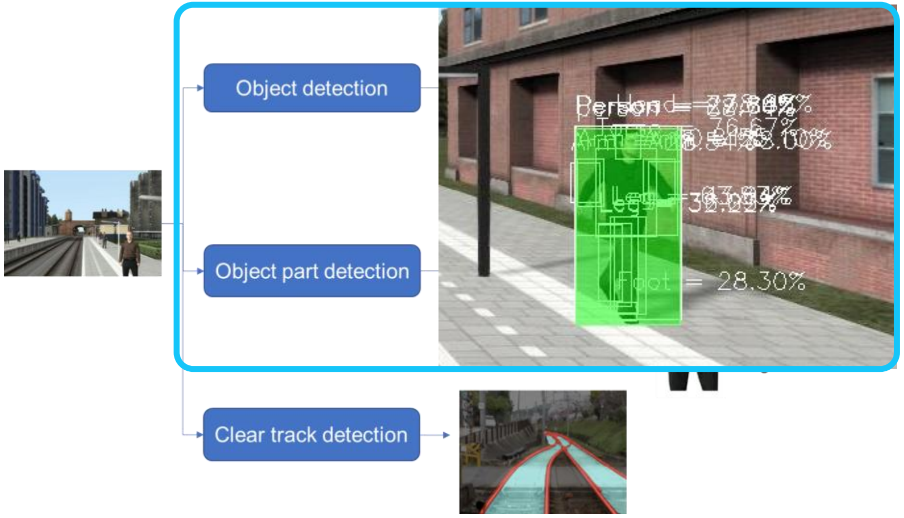


Reference safety architecture

- L0 Diverse Redundancy – Concept diversity using diverse concepts (i.e., Object Detection and Object Part Detection).

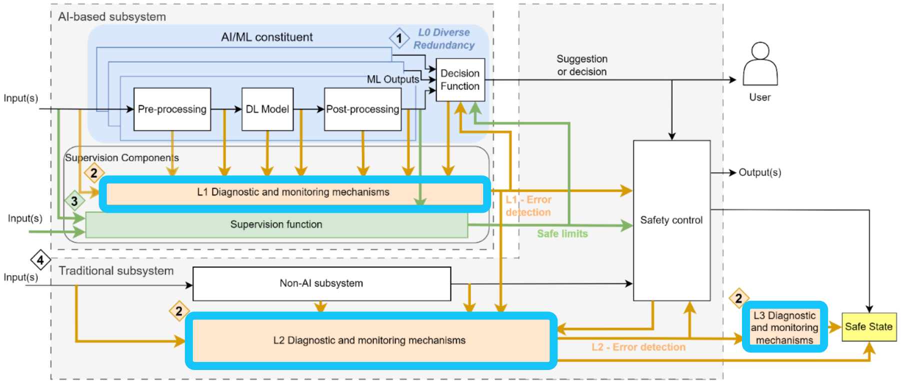


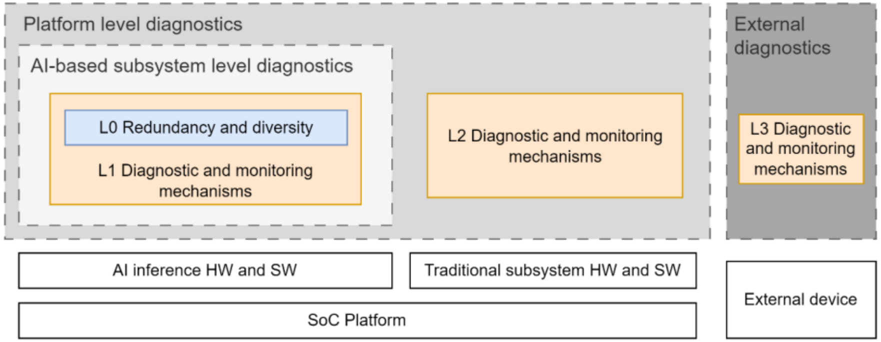




Reference safety architecture

- 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)

Reference safety architecture

- Diagnostic and monitoring mechanisms – L1 – AI-based subsystem level diagnostics

AI / ML Constituent

Input | Logic | Output

Input correctness
Data quality
Data redundancy
Temporal consistency
…

Execution errors
Timing
Program sequence
Neuron activation patterns
…

Plausibility checks
Input-output correlation
Temporal consistency
…

Reference safety architecture

- Diagnostic and monitoring mechanisms – L1 – AI-based subsystem level diagnostics

  - Input temporal consistency

Black frame

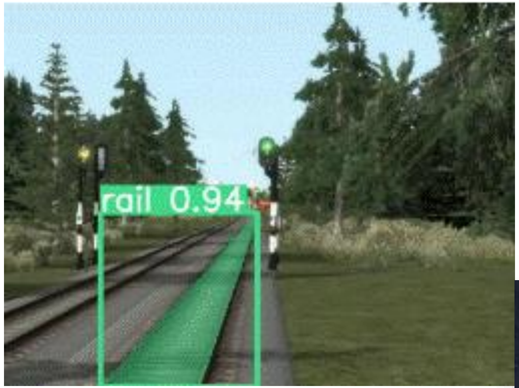
Lost frames

- Compute a metric that determines the difference among two consecutive frames
- Define a threshold

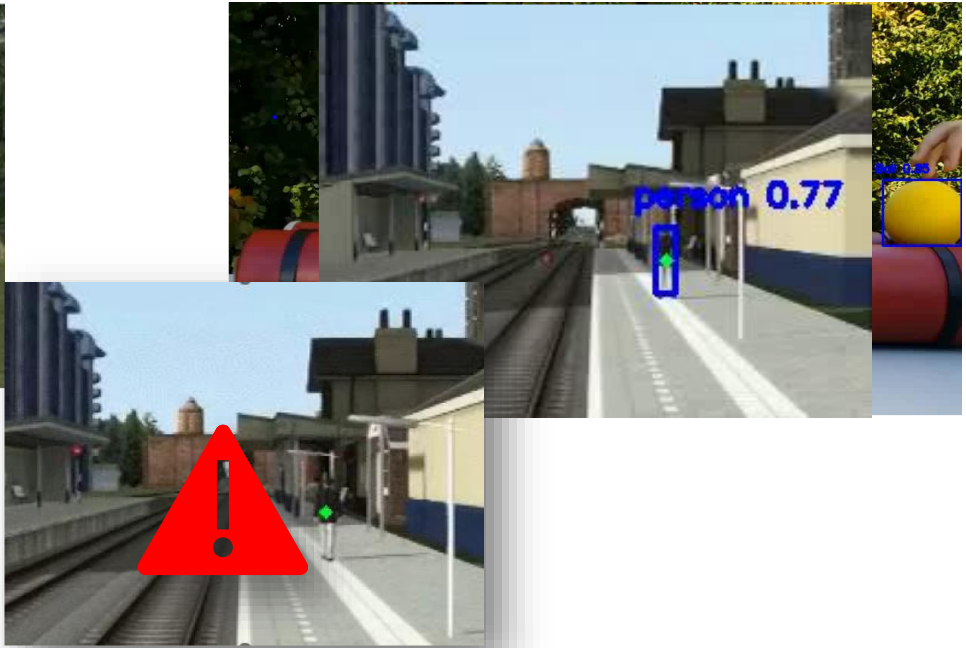
Reference safety architecture

- Diagnostic and monitoring mechanisms – L1 – AI-based subsystem level diagnostics
  - Output temporal consistency

Kalman filter

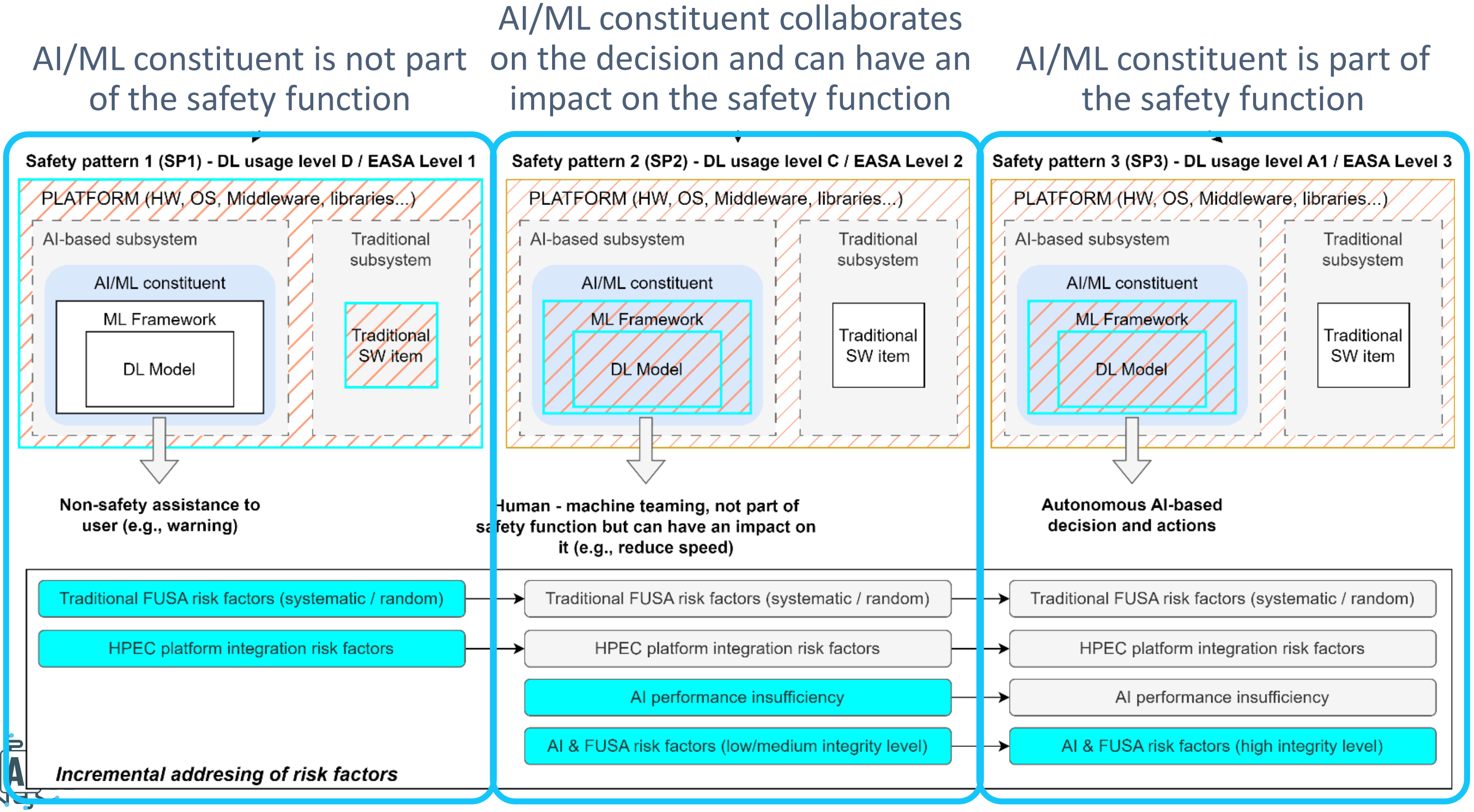


Glitches in railway track detection



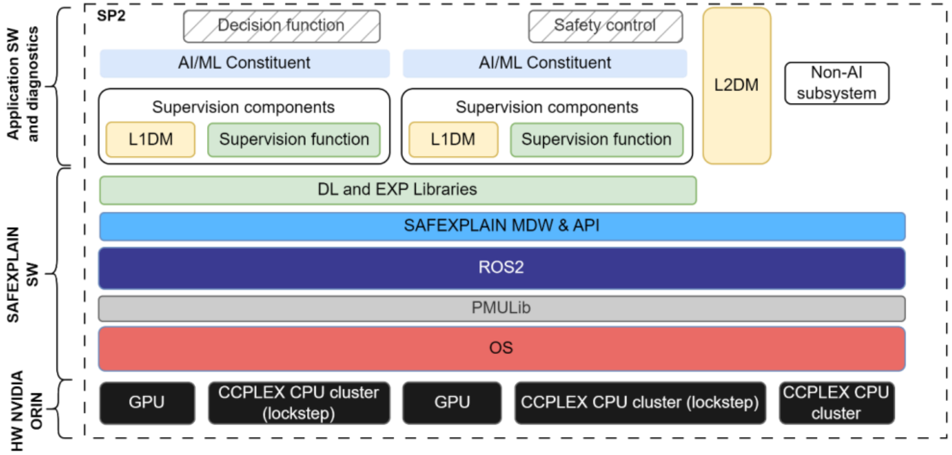
Reference safety architecture

- Incremental strategy for AI adoption in safety critical systems



Reference safety architecture

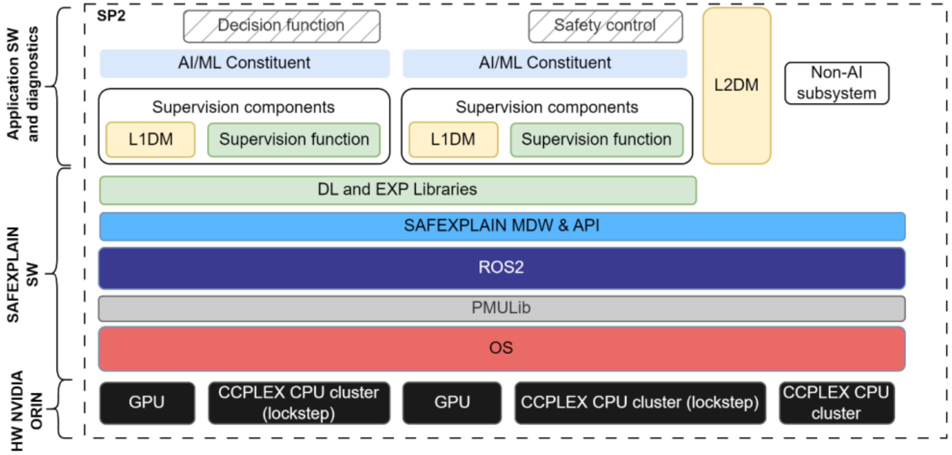
- SP2 to NVIDIA Orin resource allocation and configuration option



SP2 Element	Safety / non-safety	SP2 - A NVIDIA Orin resources and configuration
AI/ML constituent	AI based safety SW	Two instances, each in one separate CCPLEX CPU Cluster (Cortex A78) in lockstep configuration GPU for AI inference (depending on the DRS CPU or other computing resources could also be used to improve diversity) Memory controller fabric and traffic from CPU cluster to GPU MMUs for spatial independence SAFEXPLAIN SW Stack
Supervision components	Traditional or AI based safety SW	Each AI/ML constituent has each own L1DM and optionally each own supervisor function (depends on user application). Depending on the implementation of the supervision component, it may need GPUs for improved performance (e.g., AI based supervision function). The supervision components can share same CCPLEX CPU Cluster (Cortex A78) in lockstep configuration as the AI/ML constituent. MMUs for spatial independence SAFEXPLAIN SW Stack

Reference safety architecture

- SP2 to NVIDIA Orin resource allocation and configuration option



SP2 Element	Safety / non-safety	SP2 - A NVIDIA Orin resources and configuration
Decision function	Safety traditional SW	These SW components can run on any of the CCPLEX CPU Cluster (Cortex A78) in lockstep configuration used for the AI/ML constituent with the same configuration assuming they have the same integrity level.
Safety control	Safety traditional SW	
L2DM	Safety traditional SW	
Non-AI subsystem	Non-safety traditional SW	CCPLEX CPU Cluster (Cortex A78) or SPE (no need for lockstep configuration). MMUs for spatial independence L4 cache partitioning or disabled SAFEXPLAIN SW Stack or different OS on top of SPEs or hypervisor

Conclusions

- Open challenges:
  - Formalize AI and heteronomous/autonomous safety standards.
  - Define generic AI techniques and processes for developing safety-critical systems: “How things can be done” and “How things should be done”
- SAFEXPLAIN
  - Safety standards
    - Continuous follow-up of emerging initiatives and standards.
    - Define guidelines and/or adaptations to existing and ongoing standards.
  - AI processes
    - AI-FSM, to ease the development of AI-based systems while preserving safety.
  - AI techniques
    - Ongoing definition of safety architectural patterns and diagnostic mechanisms

Project Consortium

- **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/

THANK YOU!

SAFEXPLAIN
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.

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).