

Explainable AI for Systems with Functional Safety Requirements

Robert Lowe



Overview

- **Research Institutes of Sweden (RISE)** work on SAFEXPLAIN in relation to Explainable AI used for systems with Functional Safety Requirements
- Robert Lowe, Senior Researcher at RISE AB, **Unit of Human-Centred AI (Dept. of Mobility and Systems)** - focus much on Driver monitoring systems and enabling AI systems (e.g. through use of XAI, modelling of driver behaviour and cognitive-affective states).
- **For more info** from this webinar please visit SAFEXPLAIN website (<https://safexplain.eu/deliverables/>) and check out **D2.1 (Safety lifecycle considerations)**, **D2.2 (DL safety architectural patterns and platform)**, **D3.1 (Specifiability, explainability, traceability, and robustness proof-of-concept and argumentation)**, also the previous HiPEAC webinar for info on AI-Functional Safety Management (<https://safexplain.eu/ikerlan-webinar-provides-key-insights-into-ai-functional-safety-management/>)



Agenda

1) Context:

- Functional Safety Management for Artificial Intelligence (AI-FSM)
- XAI for supporting AI-FSM:
 - Data Explainers
 - Model Explainers

2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- Libraries

Agenda

1) Context:

- **Functional Safety Management for Artificial Intelligence (AI-FSM)**
- XAI for supporting AI-FSM:
 - Data Explainers
 - Model Explainers

2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- Libraries

Context: AI-FSM lifecycle

Safety lifecycle guiding the development process of dependable DL

“AI-FSM refers to all essential activities to be performed throughout the functional safety lifecycle phases to avoid systematic errors in the development of AI constituents. It is an annex to traditional FSM to be employed when a safety-critical system involves the use of AI. AI-FSM maps the content of the AI development process with the traditional safety development process.” (from: <https://safexplain.eu/ikerlan-webinar-provides-key-insights-into-ai-functional-safety-management/>)

Context: AI-FSM lifecycle

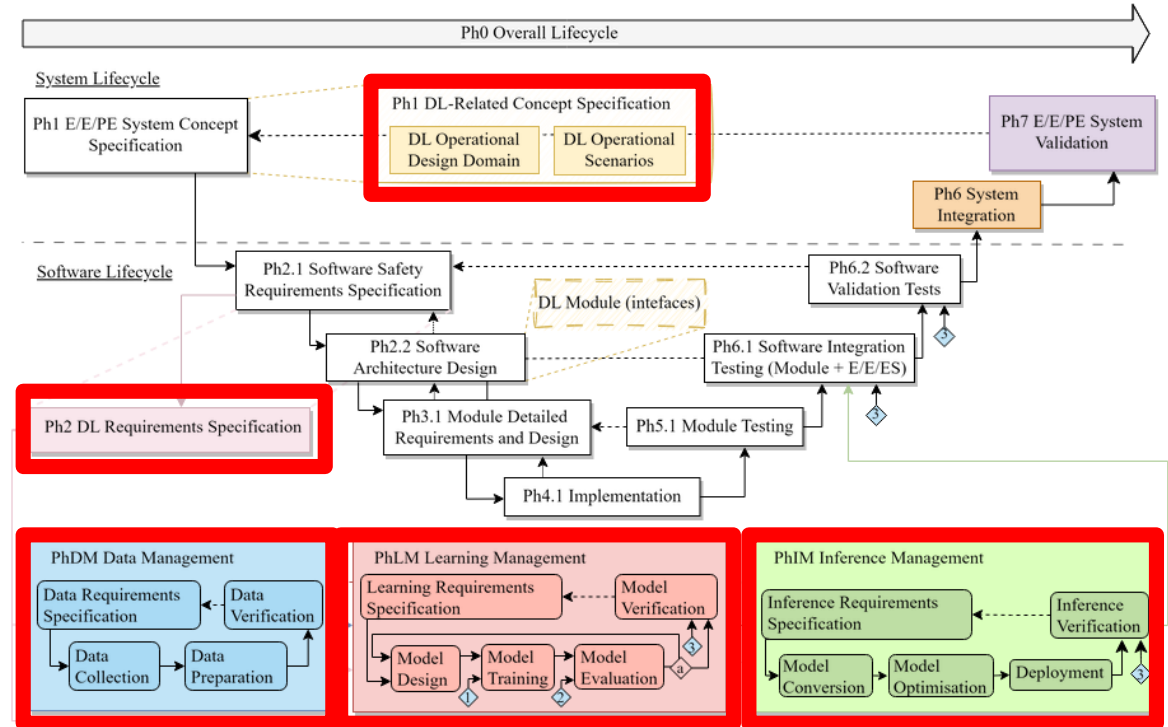
Safety lifecycle guiding the development process of dependable DL

“AI-FSM refers to all essential activities to be performed throughout the functional safety lifecycle phases to avoid systematic errors in the development of AI constituents. It is **an annex to traditional FSM to be employed when a safety-critical system involves the use of AI.** AI-FSM maps the content of the AI development process with the traditional safety development process.” (from: <https://safexplain.eu/ikerlan-webinar-provides-key-insights-into-ai-functional-safety-management/>)

Context: XAI for supporting AI-FSM

AI-FSM lifecycle

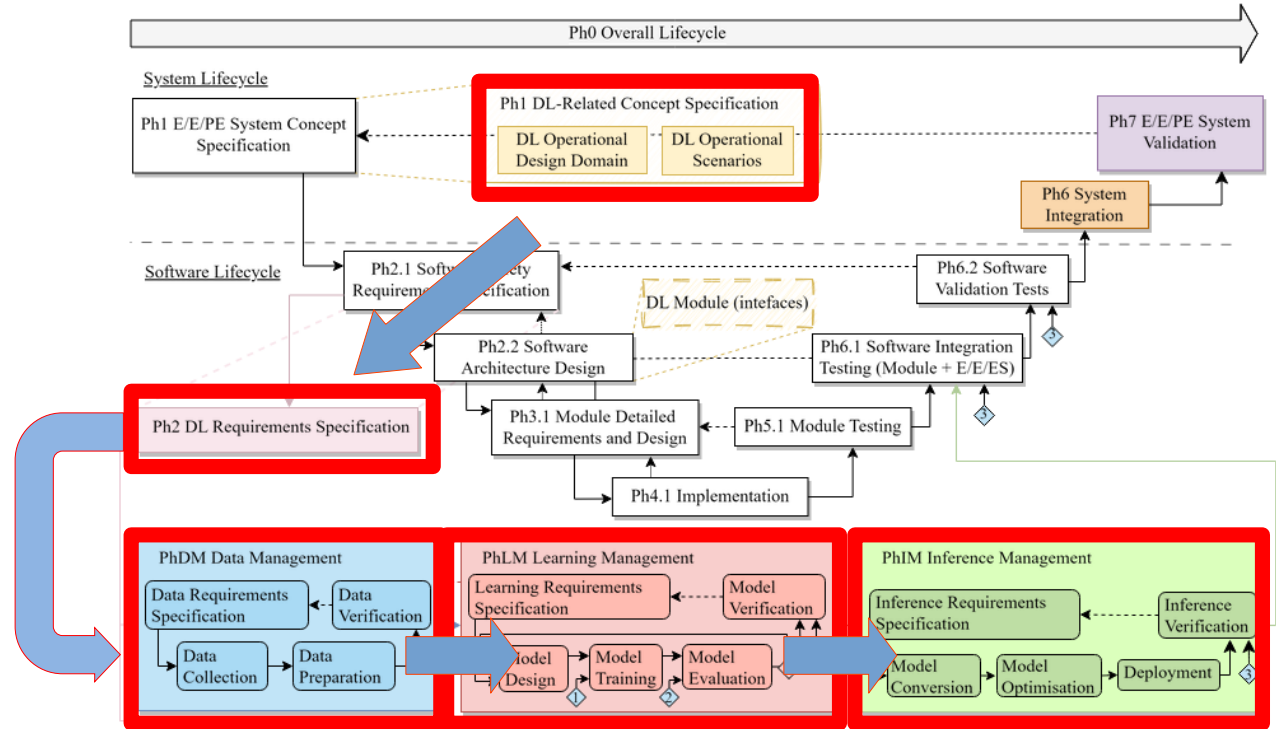
- Extension of FSM that uses V-model
- AI lifecycle is embedded within the overall lifecycle according to:
 - (Ph1) DL-related concept specification
 - (Ph2) DL Requirements Specification
 - (PhDM) Data Management
 - (PhLM) Learning Management
 - (PhIM) Inference Management



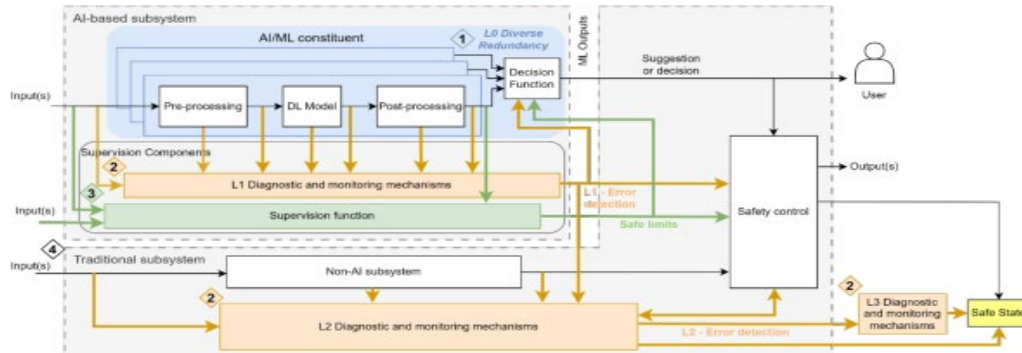
Context: XAI for supporting AI-FSM

Aim:

- Leverage XAI to ensure DL components/models are AI-FSM compliant
- Subset of key AI-FSM components are highlighted
- PhDM, PhLM and PhIM focus of XAI algorithms use in SAFEXPLAIN to ensure AI-FSM compliance

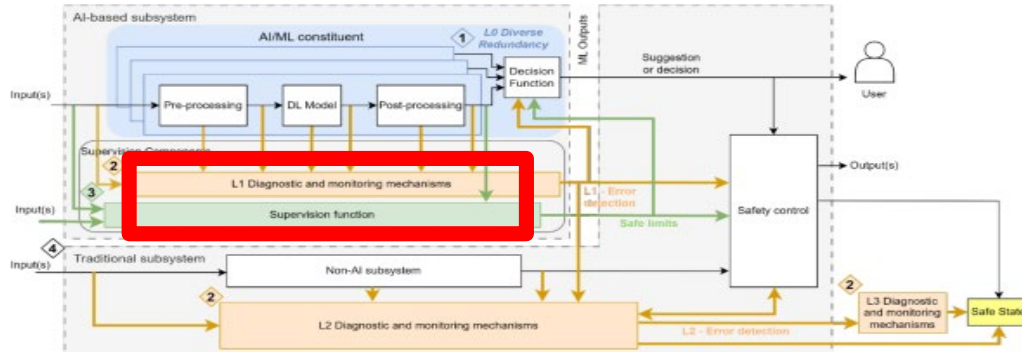


Context: Safety Patterns



Ensuring the DL is operating within the safe boundaries

Context: XAI for supporting Safety Patterns



Leverage XAI tools to build safety components

- Supervision function
- Decision function

Agenda

1) Context:

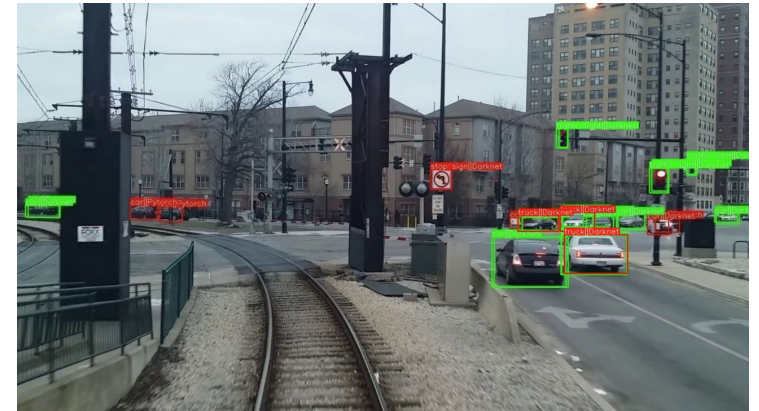
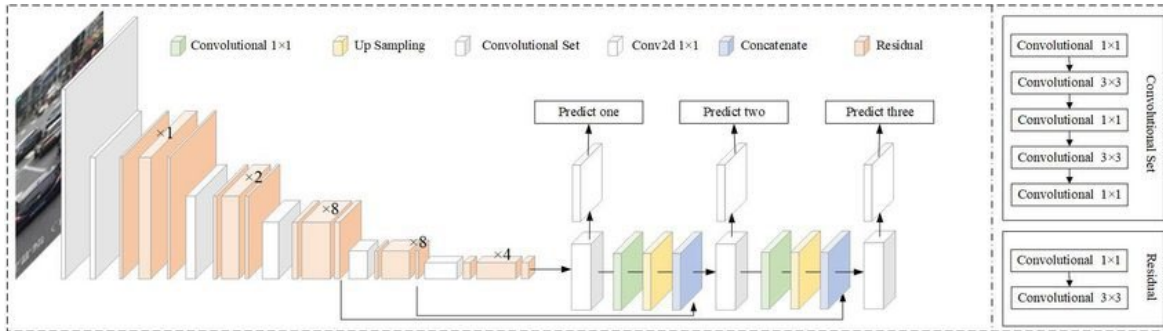
- Functional Safety Management for Artificial Intelligence (AI-FSM)
- **XAI for supporting AI-FSM:**
 - Data Explainers
 - Model Explainers

2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- Libraries

Context: AI-FSM lifecycle

- In SAFEXPLAIN the A.I approach we focus on is Deep Learning:
 - Deep Learning based Object Detection algorithms (e.g. versions of Yolo, Single Shot Detectors)
 - Many (millions of) parameters, many successive non-linear computations → complex, black boxes!



XAI for supporting AI-FSM

What is explainable AI? Algorithmic approaches to make understandable the predictions of an AI model to selected stakeholders.

Relevant stakeholders: *AI developers, data analysts, domain and FuSa experts.*

- **Comprehensibility?** can the model represent its predictions in an understandable way (to the given stakeholder)?
- **Interpretability?** does the model process 'make sense' at a domain-relevant level of explainability?

Agenda

1) Context:

- Functional Safety Management for Artificial Intelligence (AI-FSM)
- XAI for supporting AI-FSM:
 - Data Explainers
 - Model Explainers

2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- Libraries

Data Explainers

What are data explainers?

- Methods for providing insights into the data and datasets.
 - Data Profiling
 - Data Prototyping
 - Data Descriptors

Data Explainers

What are data explainers?

- Methods for providing insights into the data and datasets.
 - Data Profiling
 - Data Prototyping
 - Data Descriptors

Data Explainers: Profiling

Data Profiling:

In SAFEXPLAIN we are utilizing a number of data profiling tools according to (examples):

- i) Annotation distributions: class, bounding boxes,
- ii) image statistics -> size (max, min), resolution, histograms of various properties

Data Explainers

What are data explainers?

- Methods for providing insights into the data and datasets.
 - Data Profiling
 - Data Prototyping
 - Data Descriptors

Data Explainers: Prototyping

Data Prototyping:

Prototype selections belong to a class of algorithms to verify whether key information can be extracted from the dataset. We include within this data “prototypes/criticism” methods

Example – from ImageNet dataset



Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. Advances in neural information processing systems, 29.

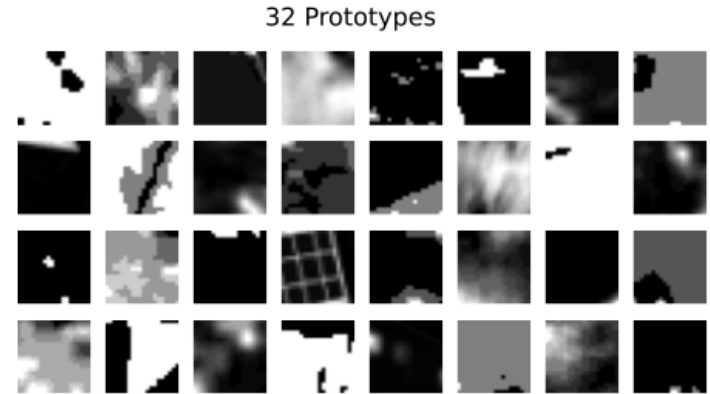
Data Explainers: Prototyping

Data Prototyping:

Prototype selections belong to a class of algorithms to verify whether key information can be extracted from the dataset. We include within this data “prototypes/criticism” methods

Example – from Satellite image dataset (used in SAFEXPLAIN)

- Top Right: 32 Prototypes; Bottom Right: 10 Criticisms



Data Explainers

What are data explainers?

- Methods for providing insights into the data and datasets.
 - Data Profiling
 - Data Prototyping
 - Data Descriptors

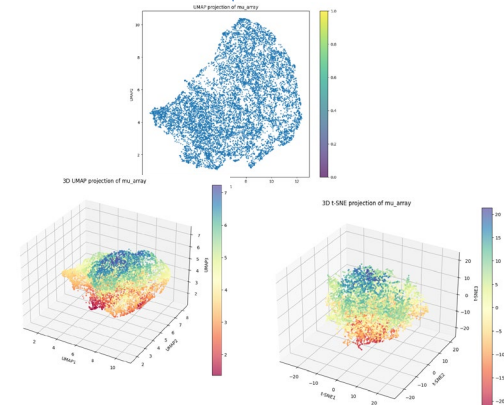
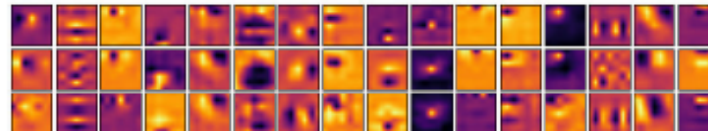
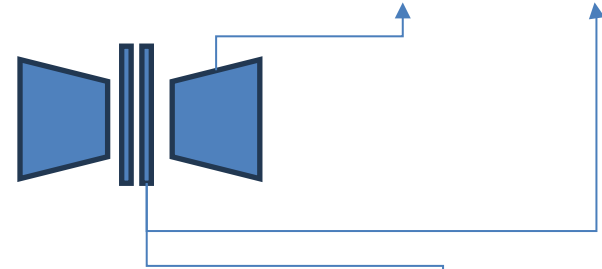
Data Descriptors

What are data descriptors?

- Used to find patterns in the dataset
- Data descriptors can help identify features/dimensions (e.g. through representations of filter vectors, or dimension reduction)
- The latent space (feature space) can also help understand what are the core features of the data, e.g. to enable reconstruction of an image



Data descriptor = (Conv basis, latent/feature)



Agenda

1) Context:

- Functional Safety Management for Artificial Intelligence (AI-FSM)
- XAI for supporting AI-FSM:
 - Data Explainers
 - **Model Explainers**

2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- Libraries

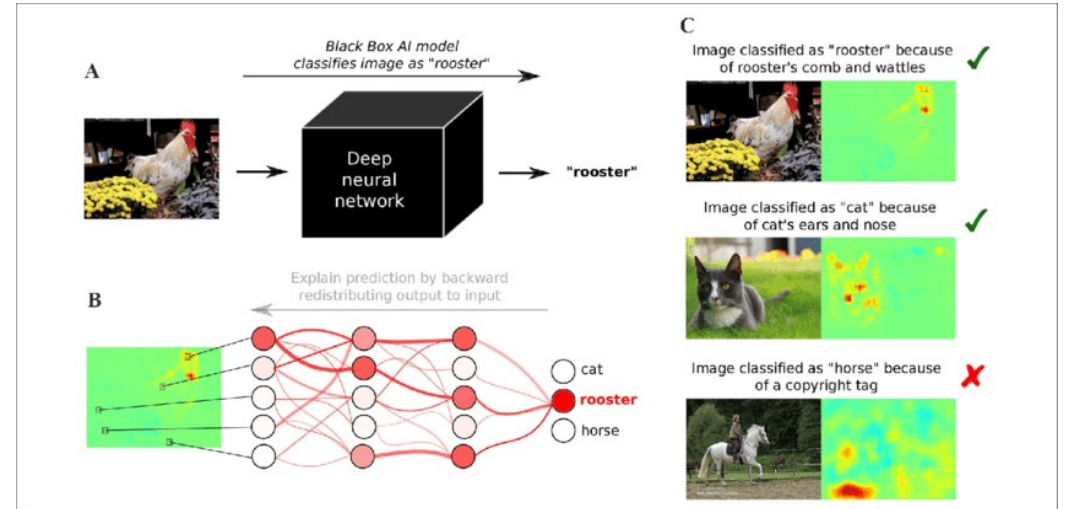
Model Explainers: Concepts

What are model explainers?

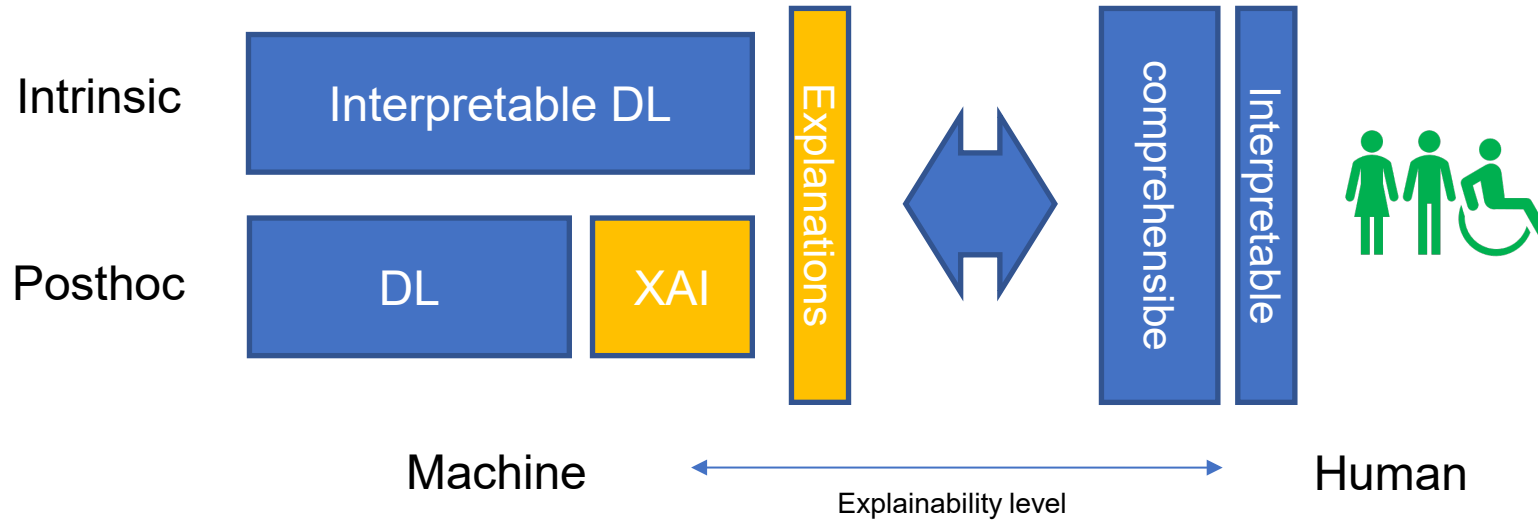
Deep learning models are often of a black box nature

Question: How can we explain the workings of black boxes to make them **interpretable**, **comprehensible**?

Answer: We use model explainability techniques.



Model explainers: Concepts



Model Explainers: Concepts

Key terms for model explainers:

- Intrinsic/Post hoc:

- **Intrinsic** : The (DL) model is optimized for a combination of performance and explainability of model predictions
- **Post hoc** : The (DL) model is trained, then XAI algorithms are applied to explain model predictions

- Global/Local:

- **Global** : Explainability of the model is achieved with respect to the entire dataset
- **Local** : Explainability of the model is achieved with respect to specific data sample

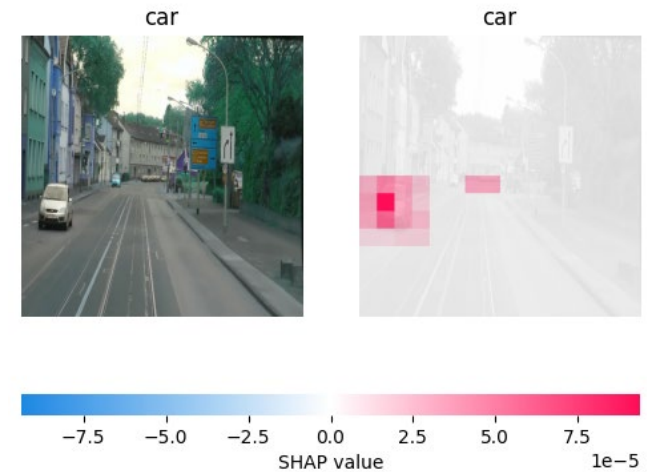
- Model agnostic/specific:

- **Model agnostic** : Explainability is applied only to the input-output mappings
- **Model specific** : Explainability is applied to the inner (featural) representations of the model

Model Explainers: Examples

Example 1 : SHAP (SHapley Additive exPlanations)

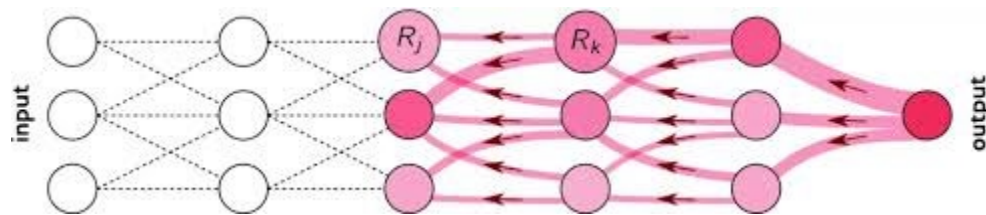
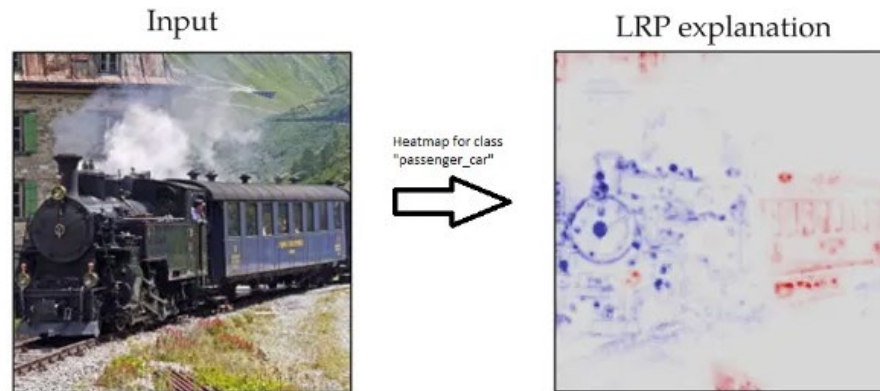
- **post hoc**: evaluates feature importance (shap values) of data for a trained model
- **local and global**: provides shap values of an data sample but also distributions of values for features across the dataset
- **model-agnostic**: evaluates effects of features (e.g. pixels in images) to model predictions, does not depend on model architecture



Model Explainers: Examples

Example 2: Layer-wise Relevance Propagation (LRP)

- **post hoc**: evaluates feature *relevance* of a trained model
- **local**: provides feature-relevance explanations of an individual image using a heatmap
- **model-specific**: back chains relevance values from output nodes (of an ANN) to input nodes (for assigning relevance to a given output/prediction)



Agenda

1) Context:

- Functional Safety Management for Artificial Intelligence (AI-FSM)
- XAI for supporting AI-FSM:
 - Data Explainers
 - Model Explainers

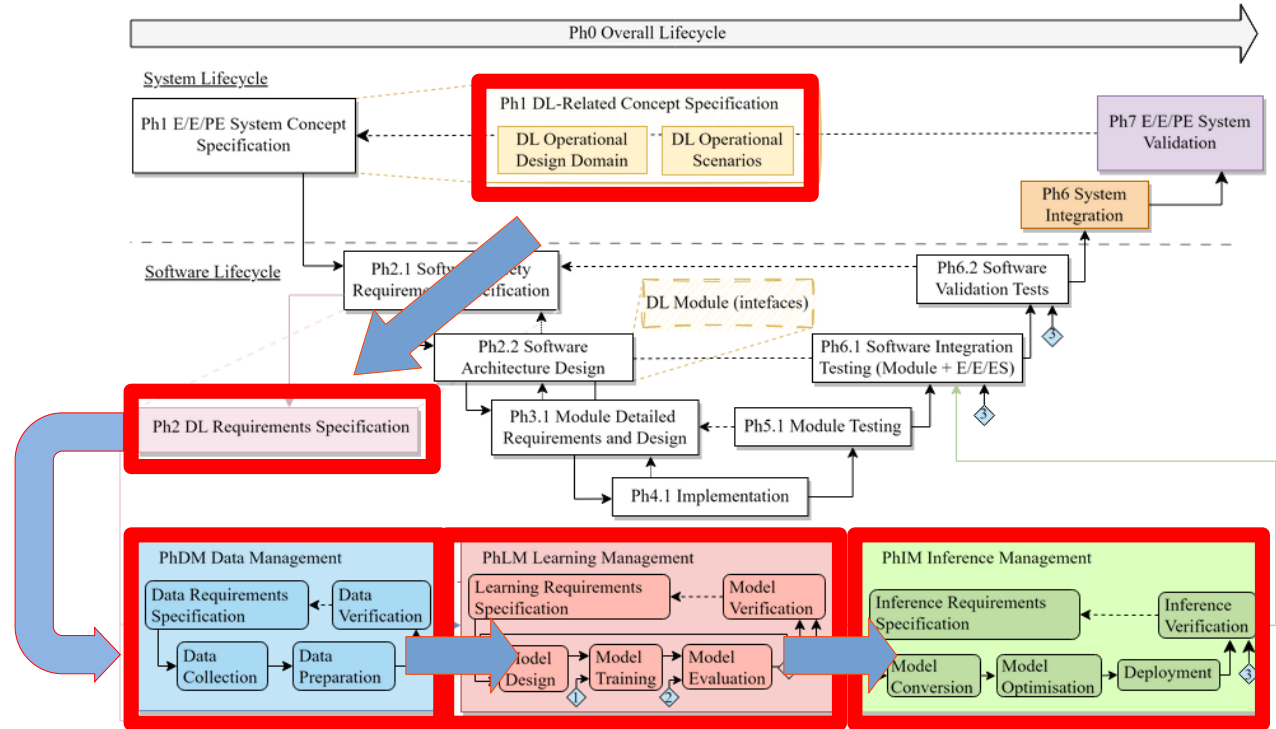
2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- Libraries

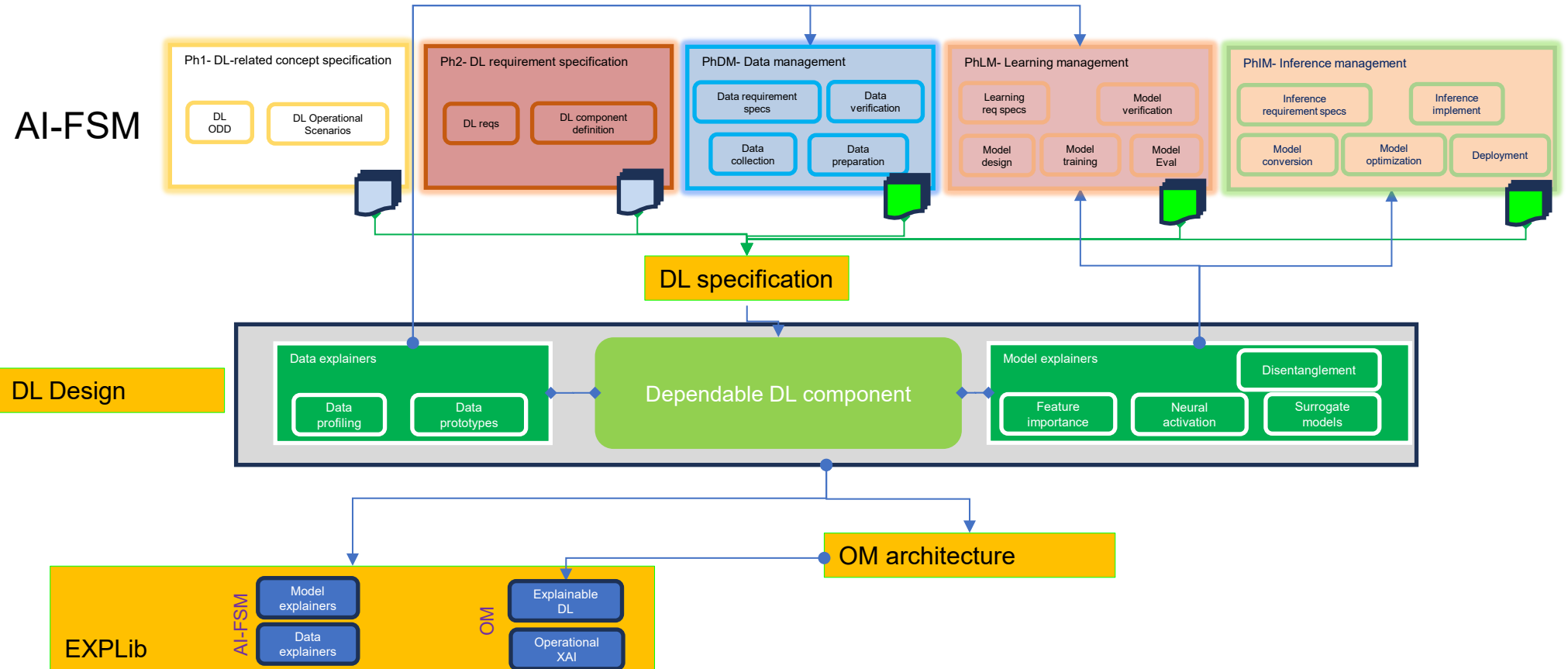
Walkthrough

Aim:

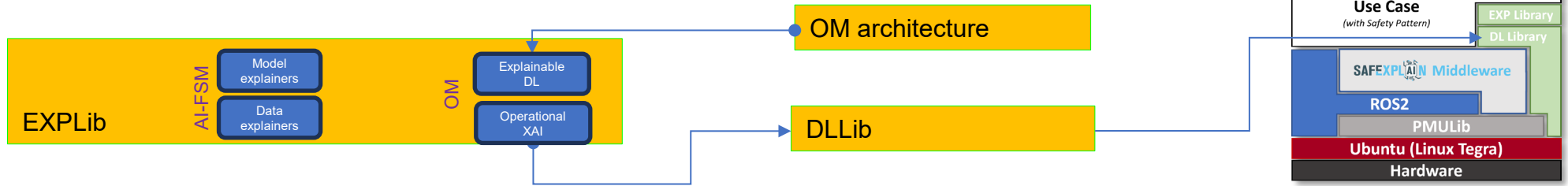
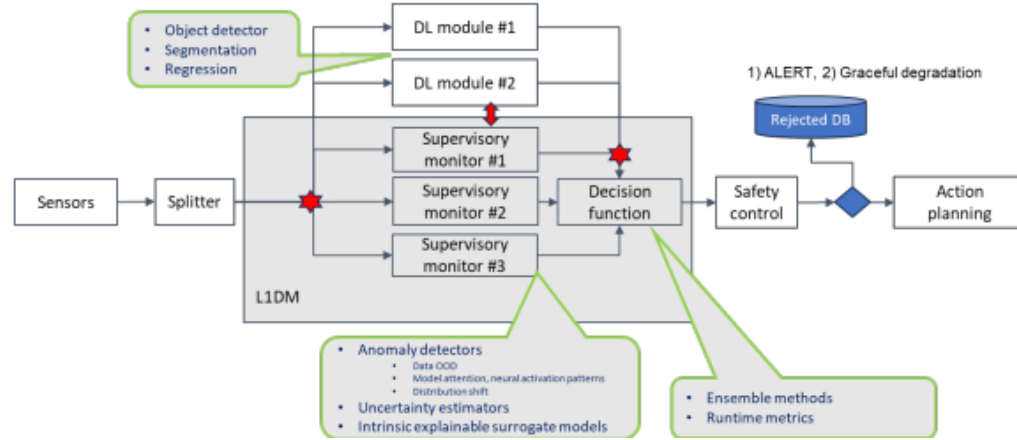
- Leverage XAI to ensure DL components/models are AI-FSM compliant
- Subset of key AI-FSM components are highlighted
- PhDM, PhLM and PhIM focus of XAI algorithms use in SAFEXPLAIN to ensure AI-FSM compliance



Walkthrough: Development Stage



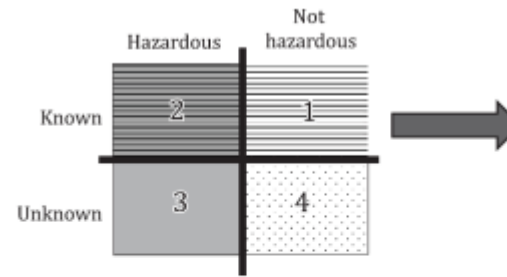
Walkthrough: Deployment Stage



XAI and Dealing with Uncertainty

- **Pre-Development:**
- Scenarios in which the DL component may operate can be defined according to 4 areas (SOTIF standard: ISO 21448):

1. Known, not hazardous;
2. Known, hazardous;
3. Unknown, hazardous;
4. Unknown, not hazardous.



Example of an initial starting point of development

Key



known, not hazardous scenarios (area 1)



known, hazardous scenarios (area 2)



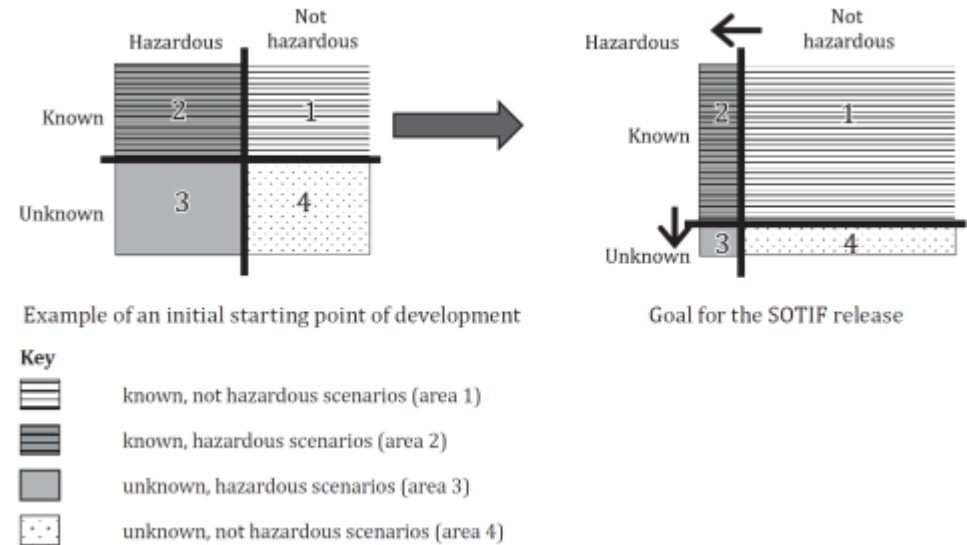
unknown, hazardous scenarios (area 3)



unknown, not hazardous scenarios (area 4)

XAI and Dealing with Uncertainty

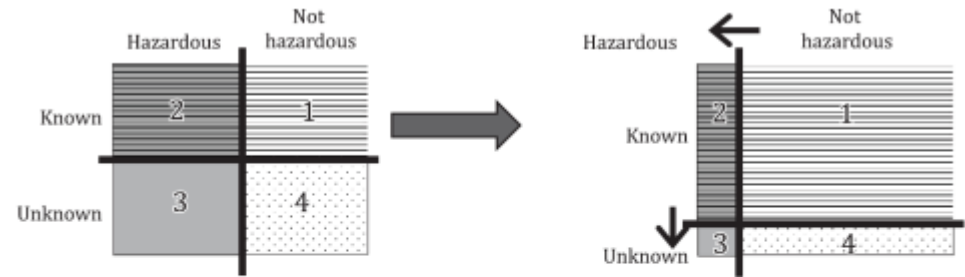
- **Post-Development:**
- We would like to reduce the areas of 3. (Unknown hazardous) and 2. (Known hazardous) as much as possible
- Perform a risk acceptance evaluation
- Acceptable residual uncertainty



XAI and Dealing with Uncertainty

- **Deployment:**
- We need the deployed system to recognize the boundaries of safe use, i.e. if areas 2 and 3 are identified, the DL component is operating outside its ODD





During deployment, the AI-FSM compliant solution should operate outside the reduced areas 2 & 3.



Example of an initial starting point of development

Goal for the SOTIF release

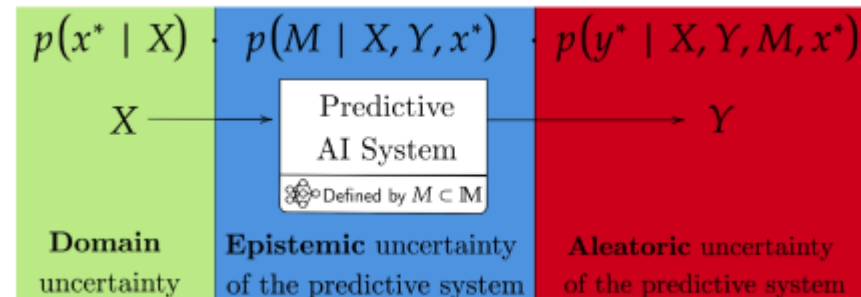
Key

-  known, not hazardous scenarios (area 1)
-  known, hazardous scenarios (area 2)
-  unknown, hazardous scenarios (area 3)
-  unknown, not hazardous scenarios (area 4)

XAI and Dealing with Uncertainty

Sources of uncertainty:

- 1) **Domain uncertainty** → Uncertainty that the dataset represents the stated problem
- 2) **Epistemic uncertainty** → Uncertainty that the model correctly models the stated problem
- 3) **Alleatoric uncertainty** → Irreducible uncertainty concerned with the nature of the problem



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.

XAI and Dealing with Uncertainty

- **Managing uncertainty:**

Domain uncertainty:

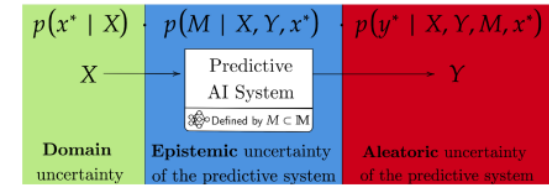
- Ensure that the input data belongs to the “known” datasets as managed by PhDM
- Realized by Out of Distribution detectors (Input anomaly detectors)

Epistemic uncertainty:

- Ensure that the model behaviours are belonging to the “known” behaviours
- Realized by model global explainers, such as interpretable surrogate models

Aleatoric uncertainty:

- Ensure that the uncertainties are within the safe boundaries
- Realized by modifications of the target DL component to provide uncertainty estimates



Agenda

1) Context:

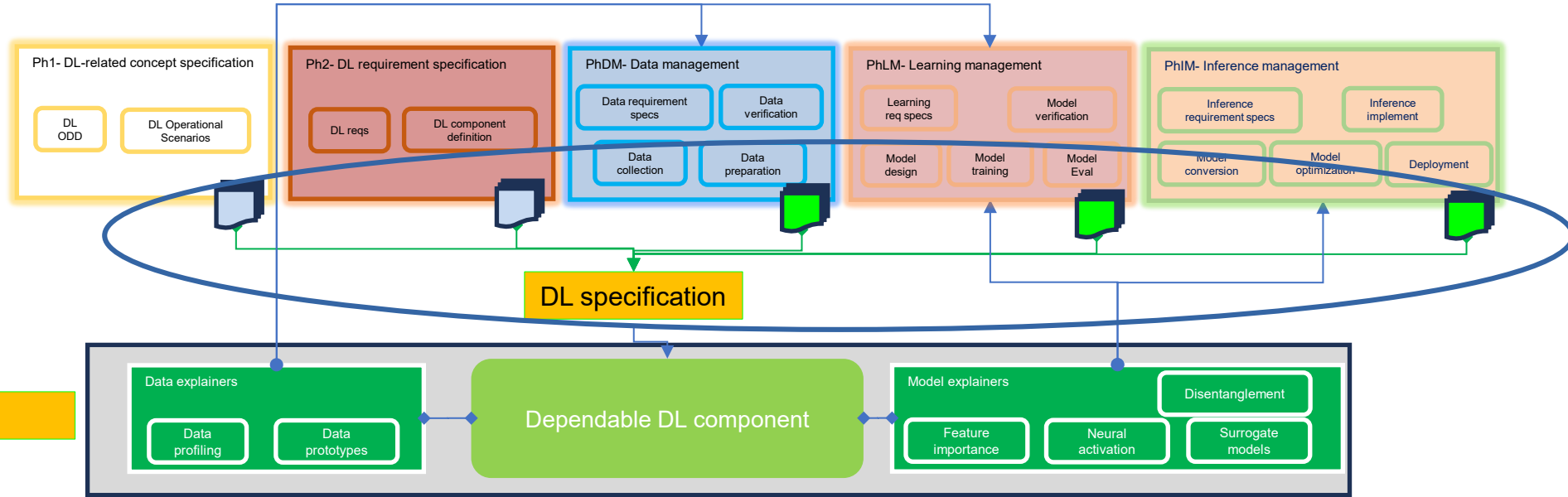
- Functional Safety Management for Artificial Intelligence (AI-FSM)
- XAI for supporting AI-FSM:
 - Data Explainers
 - Model Explainers

2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- Libraries

Walkthrough: Making DL component compliant with AI-FSM

AI-FSM



- Specification of Dependable DL components derived from AI-FSM artifacts
- XAI supports activities and artifacts within PhDM, PhLM, PhIM

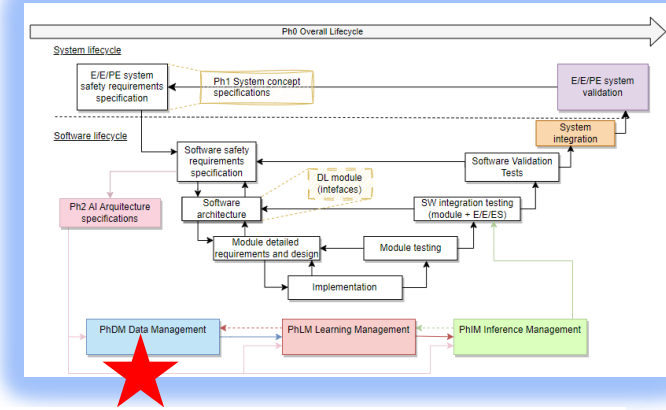
PhDM - Data Management

- Purpose

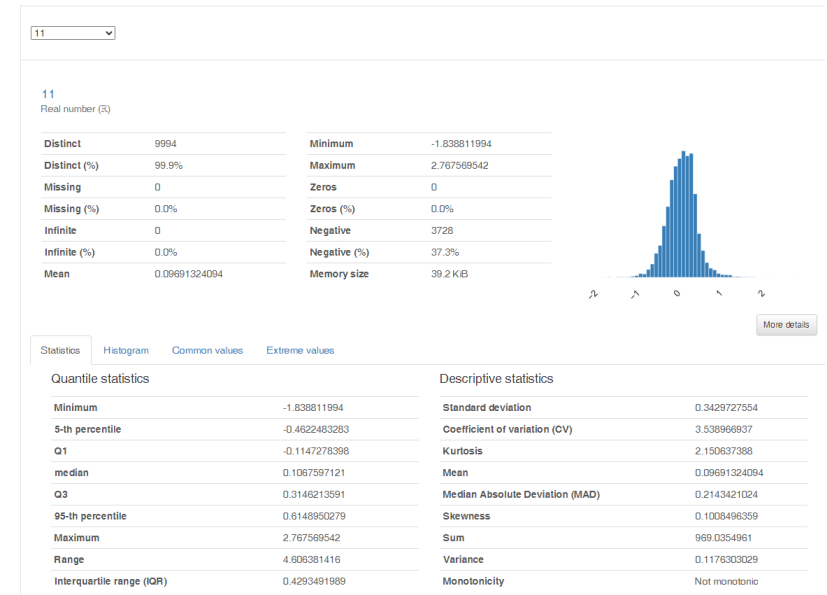
- Generate data specs
- Assess/Verify if dataset meets data requirements
- Baseline “Known” conditions for OM

- Data explainers

- **Data profiling, statistical inference**
- Data protoypes/descriptors
 - Prototypical instances/patches
 - Generative AI based descriptors

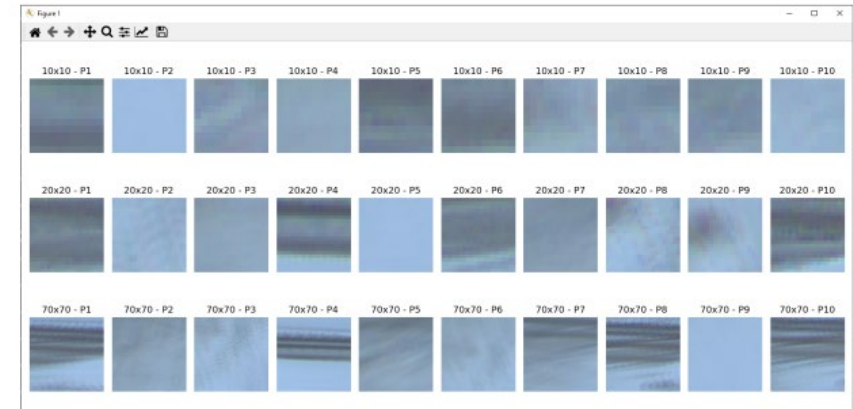
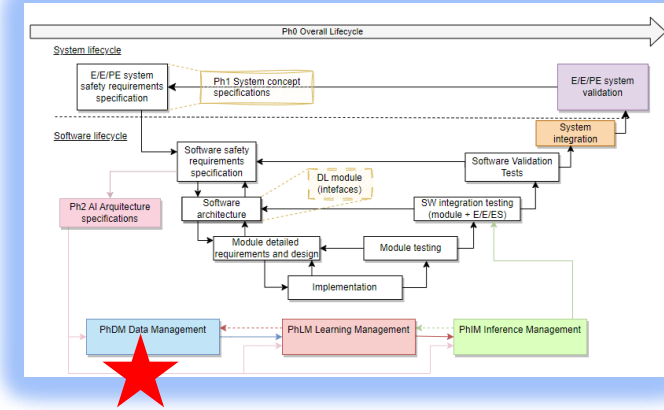


Variables



PhDM - Data Management

- Purpose
 - Generate data specs
 - Assess/Verify if dataset meets data requirements
 - Baseline “Known” conditions for OM
- Data explainers
 - Data profiling, statistical inference
 - **Data prototypes/descriptors**
 - **Prototypical instances/patches**
 - Generative AI based descriptors



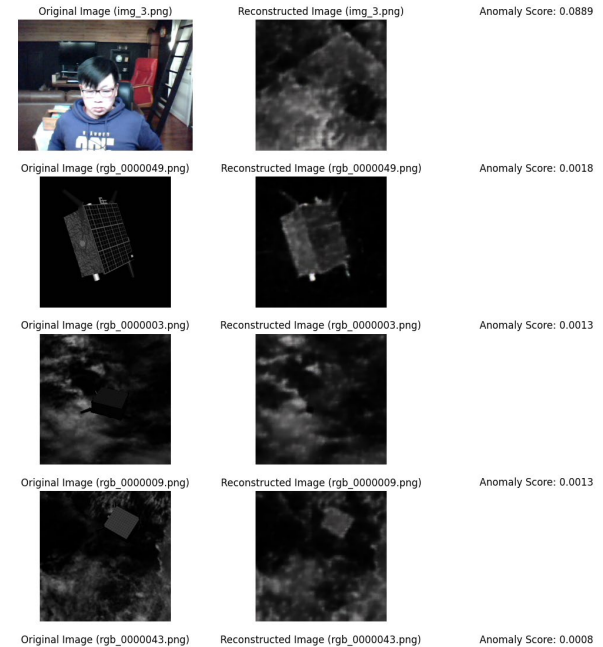
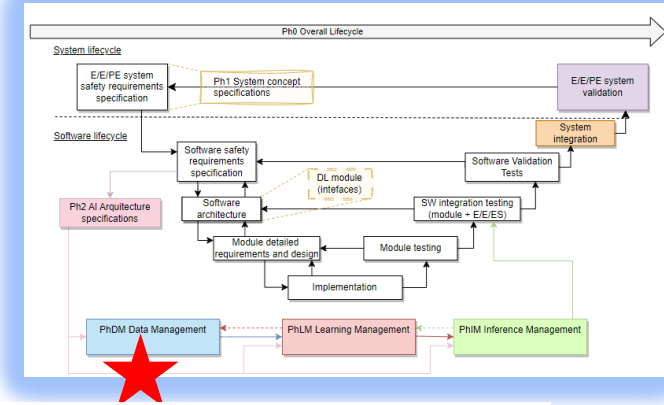
PhDM - Data Management

- Purpose

- Generate data specs
- Assess/Verify if dataset meets data requirements
- Baseline “Known” conditions for OM

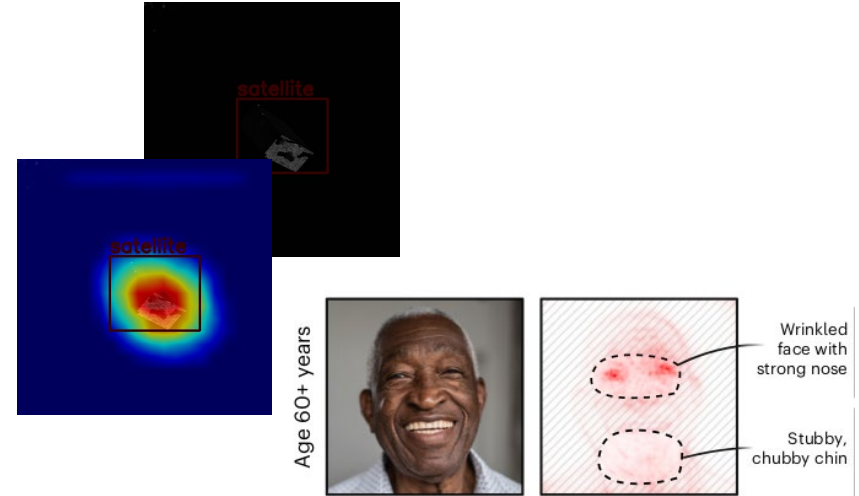
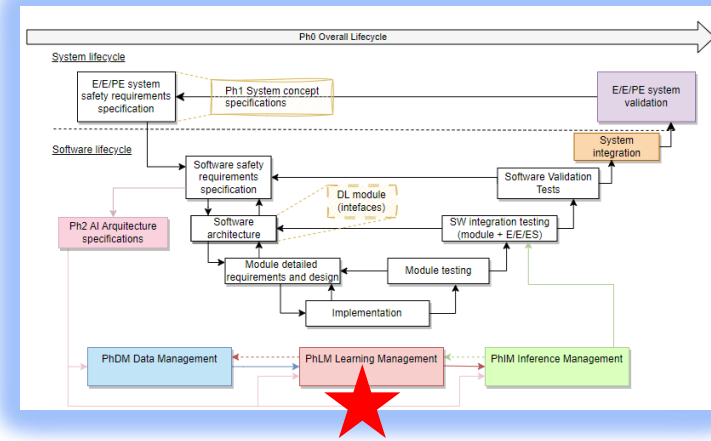
- Data explainers

- Data profiling, statistical inference
- **Data prototypes/descriptors**
 - Prototypical instances/patches
 - **Generative AI based descriptors**



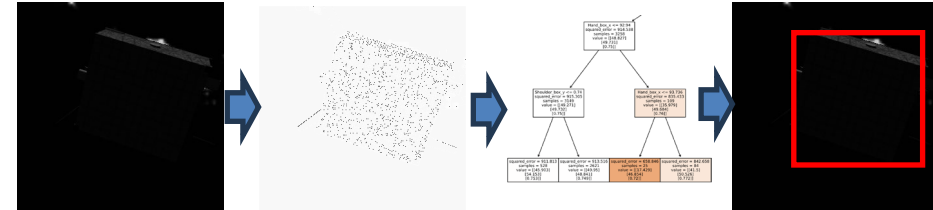
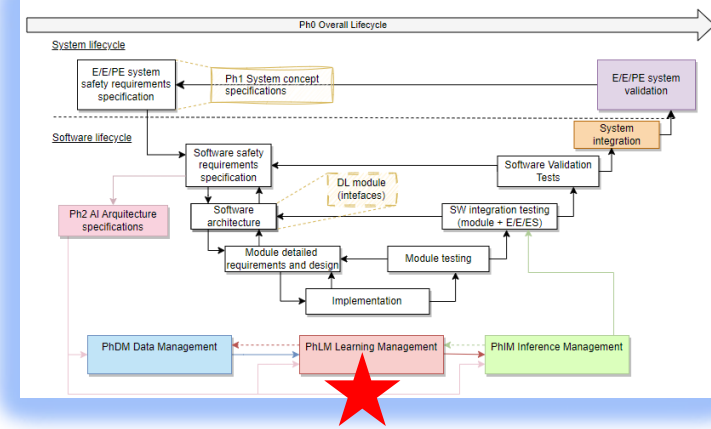
PhLM Learning Management

- Purpose
 - Safety driven metrics
 - Explainable by design, explanations as evidence
 - Reduce epistemic (model) uncertainty
 - Baseline model “normal” behaviours
- Model explainers
 - **Design: Disentanglement, Decomposition, Gradient/activation extractors**
 - Global explanation: Surrogate models
 - Local explanations to diagnose corner cases

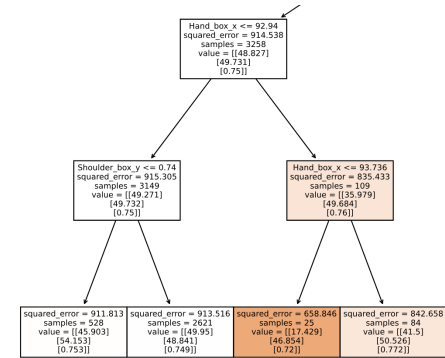


PhLM Learning Management

- Purpose
 - Safety driven metrics
 - Explainable by design, explanations as evidence
 - Reduce epistemic (model) uncertainty
 - Baseline model “normal” behaviours
- Model explainers
 - Design: Disentanglement, Decomposition, Gradient/activation extractors
 - **Global explanation: Surrogate models**
 - Local explanations to diagnose corner cases

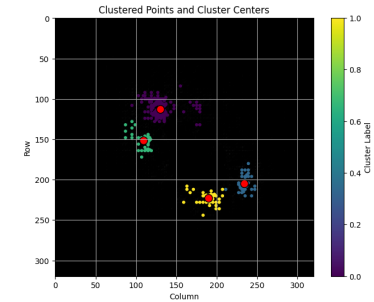
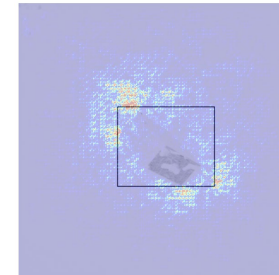
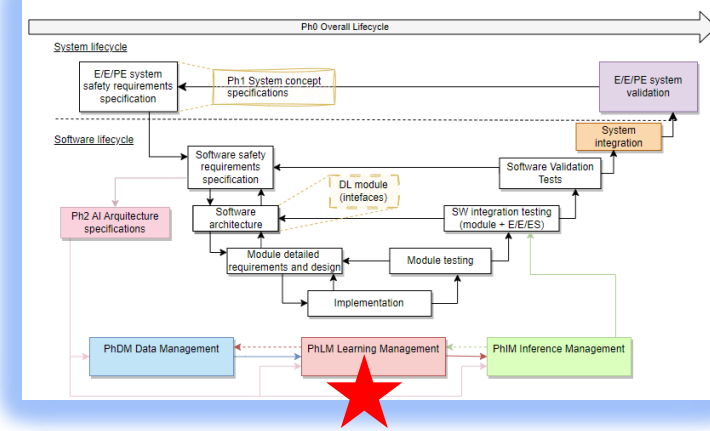


parts, edges, LBP...



PhLM Learning Management

- Purpose
 - Safety driven metrics
 - Explainable by design, explanations as evidence
 - Reduce epistemic (model) uncertainty
 - Baseline model “normal” behaviours
- Model explainers
 - Design: Disentanglement, Decomposition, Gradient/activation extractors
 - Global explanation: Surrogate models
 - **Local explanations to diagnose edge/corner cases**



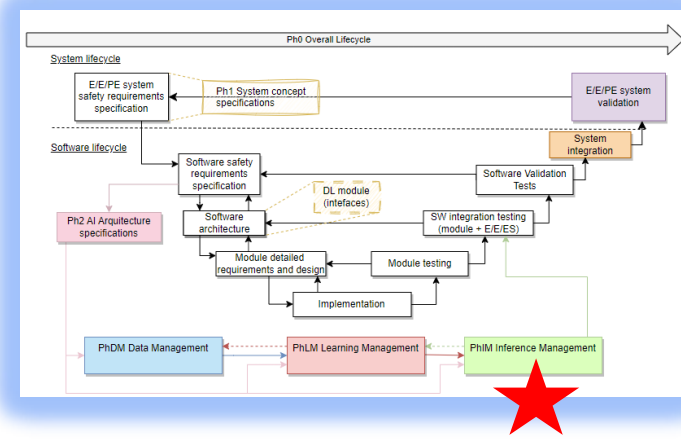
PhIM Inference Management

- Purpose

- Model conversion
- Model Optimization
- Inference Verification

- Model explainers

- **For Model Conversion:** Use XAI techniques (Local, global; model specific/agnostic) to compare explainability of converted model to original model
- **For Model Optimization:** Use XAI techniques as for model conversion but apply for layers/neurons (to evaluate effects of pruning) – model specific
- **For Inference Verification:** Use Surrogate models (to ensure final model is **acceptably** the same as the original model)



Agenda

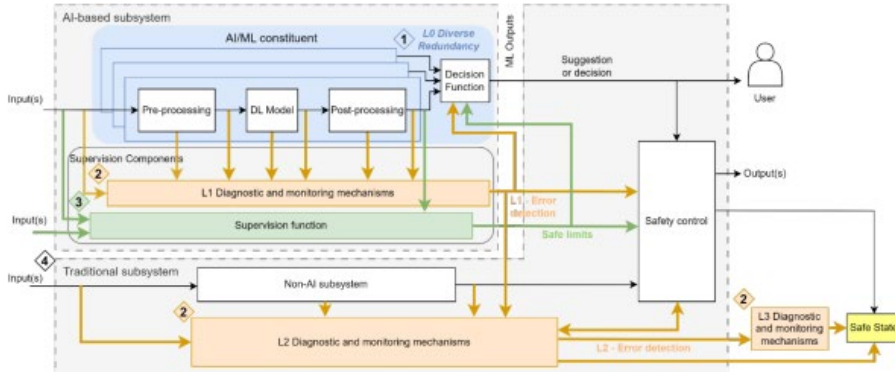
1) Context:

- Functional Safety Management for Artificial Intelligence (AI-FSM)
- XAI for supporting AI-FSM:
 - Data Explainers
 - Model Explainers

2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- **Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)**
- Libraries

Operation & Monitoring

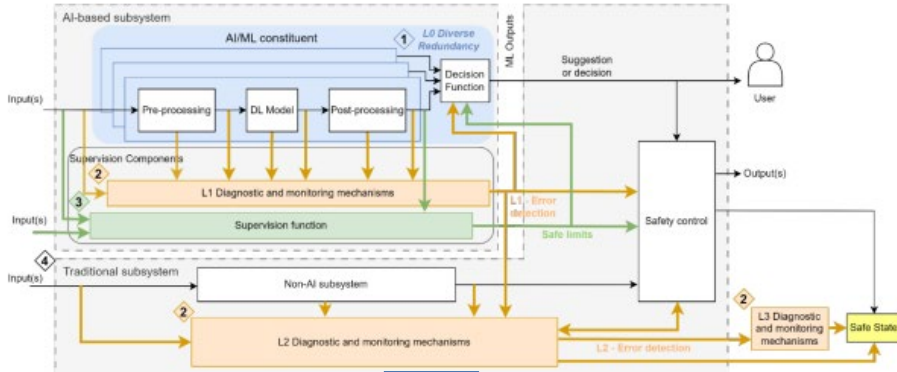


Safety pattern 2

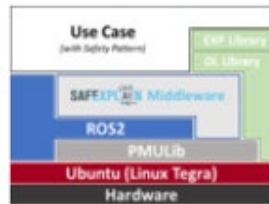
- SAFEXPLAIN *safety patterns* are used in the Operation and Monitoring stage (OM stage)
- Provide reference safety architecture with required functionalities of the components
- XAI tools can be used for building components within AI-based subsystem

Operation & Monitoring

Reference safety architecture

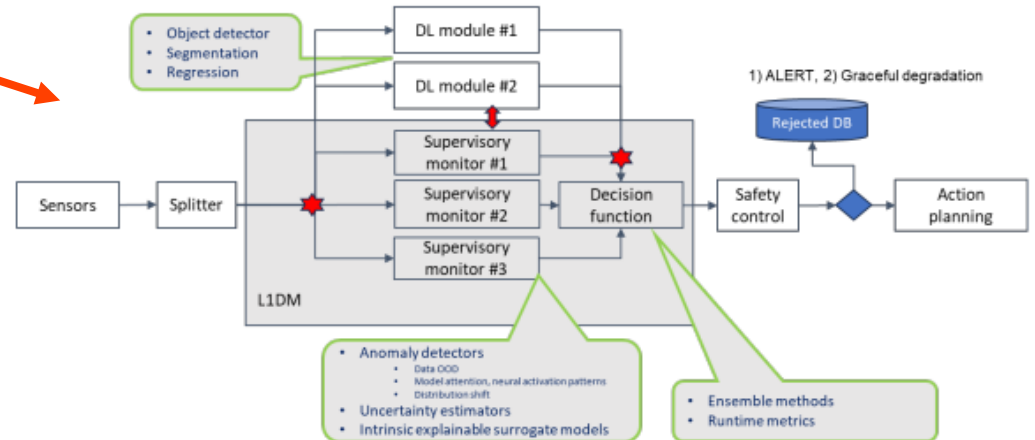
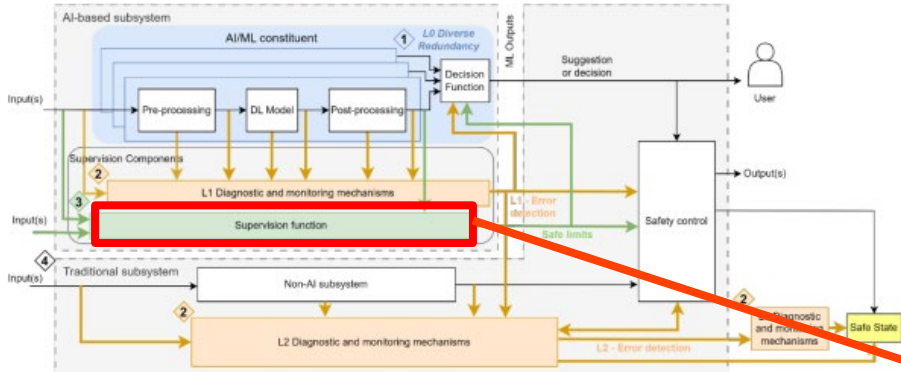


- The safety pattern is realized in selected SAFEXPLAIN platform with SW stack



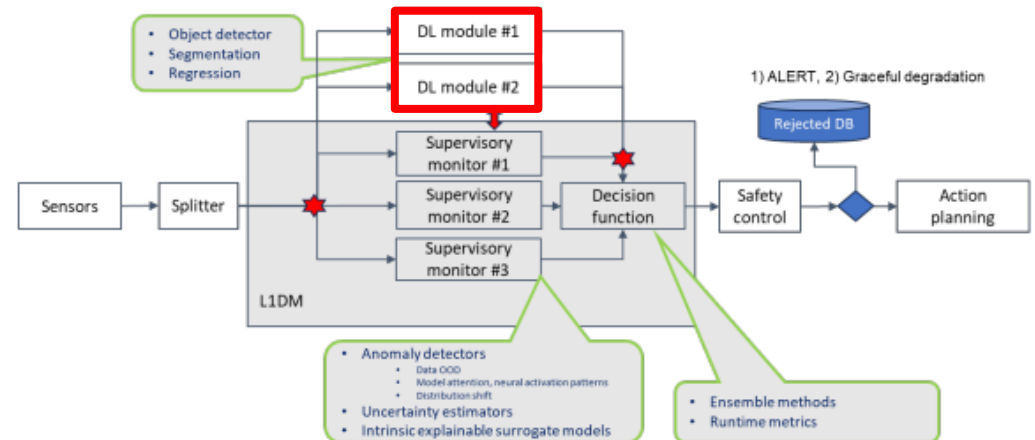
**Evaluation on NVIDIA JETSON
AGX ORIN & SAFEXPLAIN SW
Stack**

Operation & Monitoring



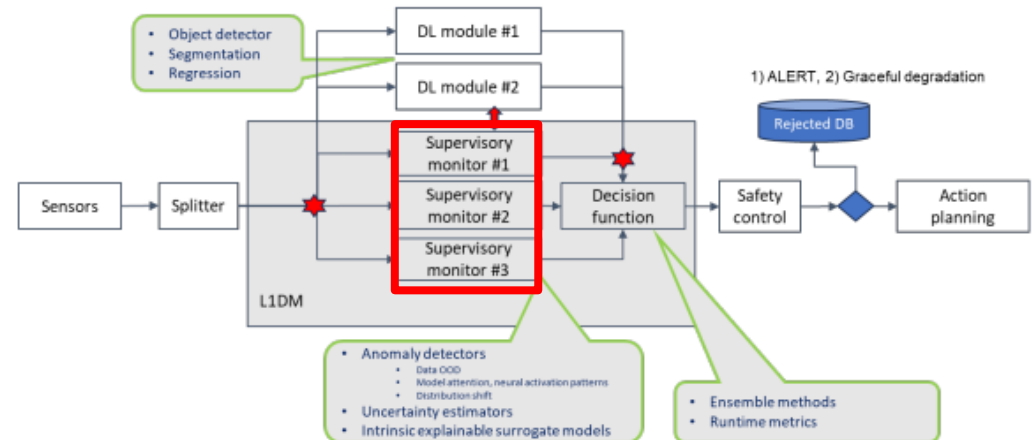
Operation & Monitoring

- OM architecture purpose:
 - Control unknowns
 - Control known uncertainties
- OM constitution:
 - Supervises one or more DL modules:
 - Redundancy of DL modules
 - All contribute to the decision function



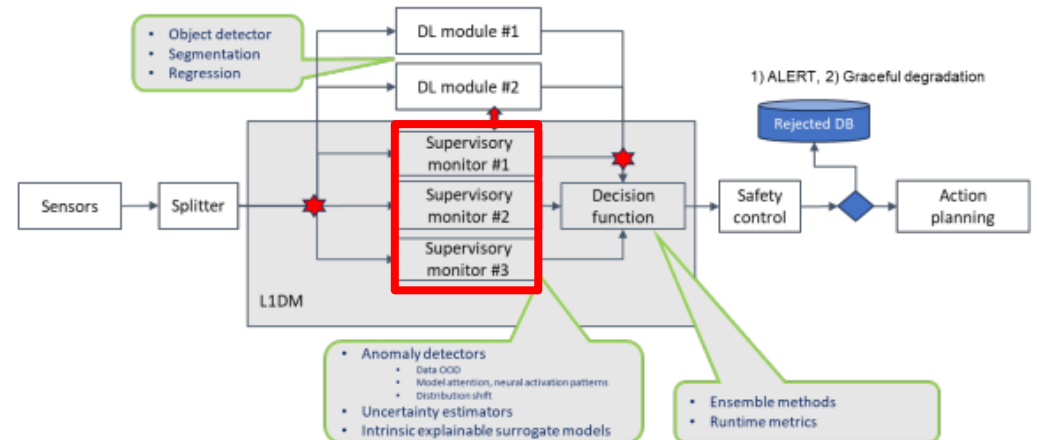
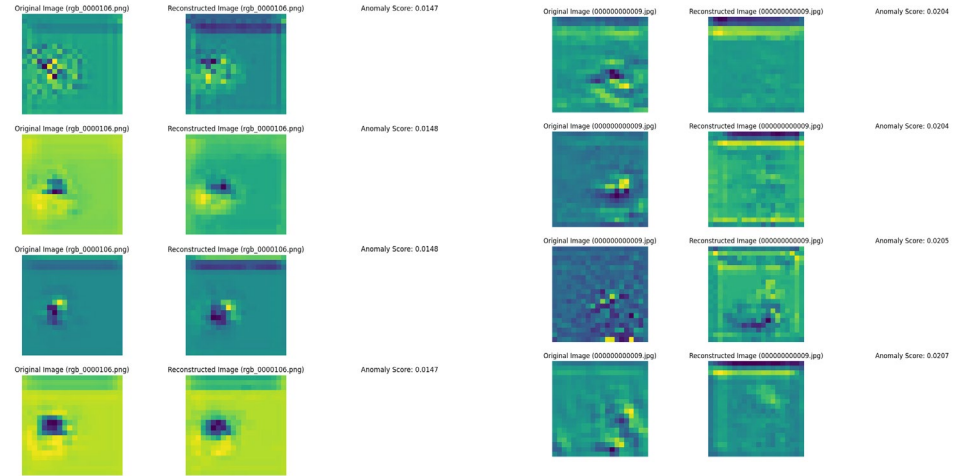
Operation & Monitoring

- OM architecture purpose:
 - Control unknowns
 - Control known uncertainties
- OM constitution:
 - Supervisory monitoring components
 - **#1 Anomaly checks:**
 - Anomaly detections (input/model/output)
 - **#2 Consistency of Interpretability check:**
 - Surrogate model
 - **#3 Uncertainty estimation check:**
 - Uncertainty aware DL module



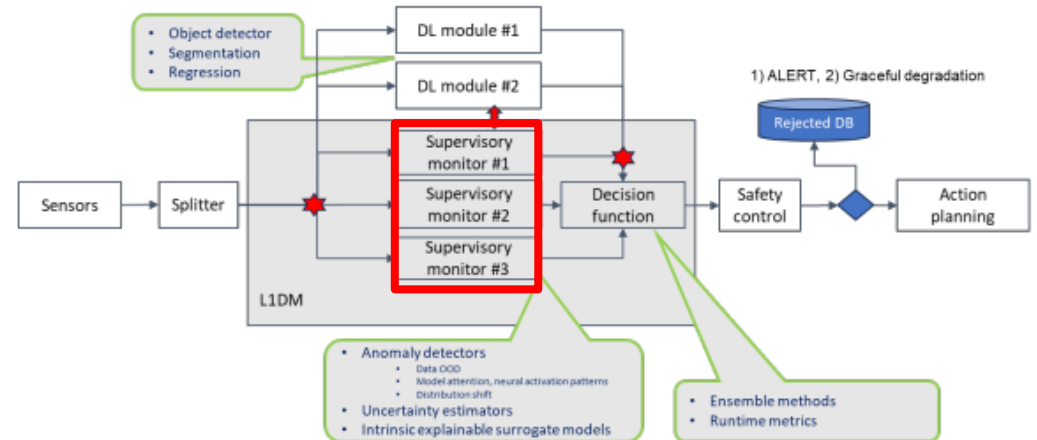
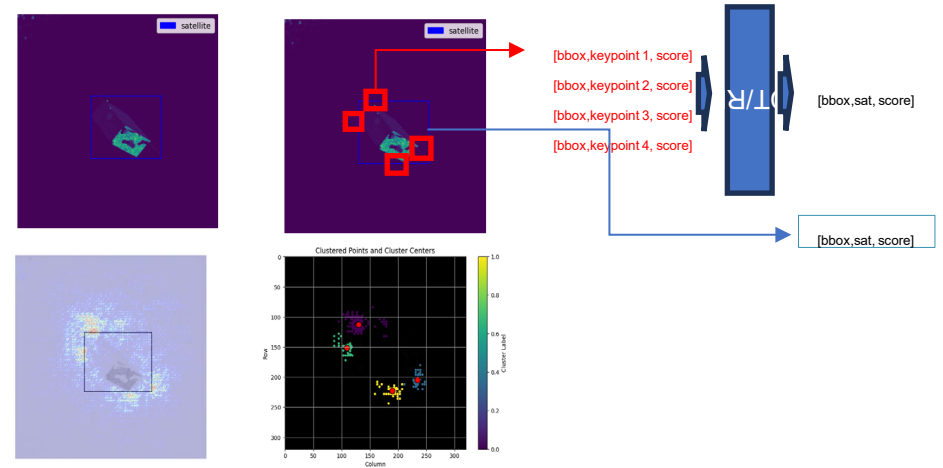
Operation & Monitoring

- OM architecture purpose:
 - Control unknowns
 - Control known uncertainties
- OM constitution:
 - Supervisory monitoring components
 - **#1 Anomaly checks:**
 - Anomaly detections (input/model/output)
 - #2 Consistency of Interpretability check:
 - Surrogate model
 - #3 Uncertainty estimation check:
 - Uncertainty aware DL module



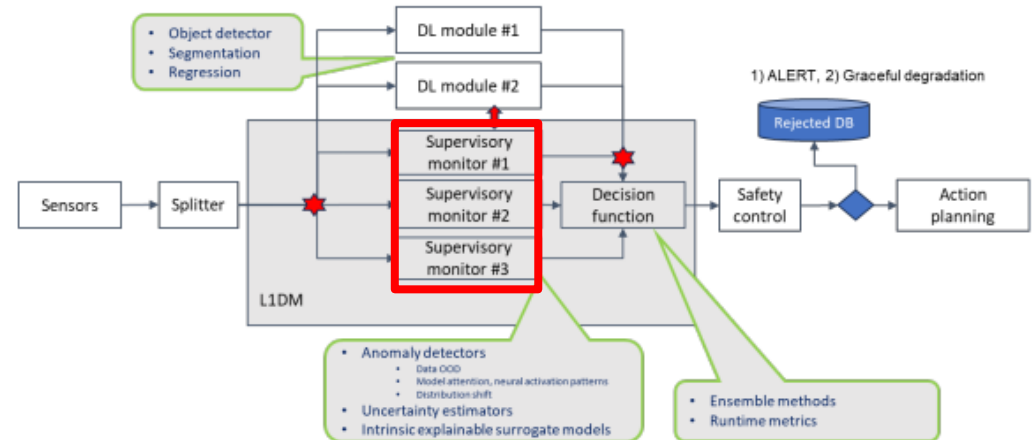
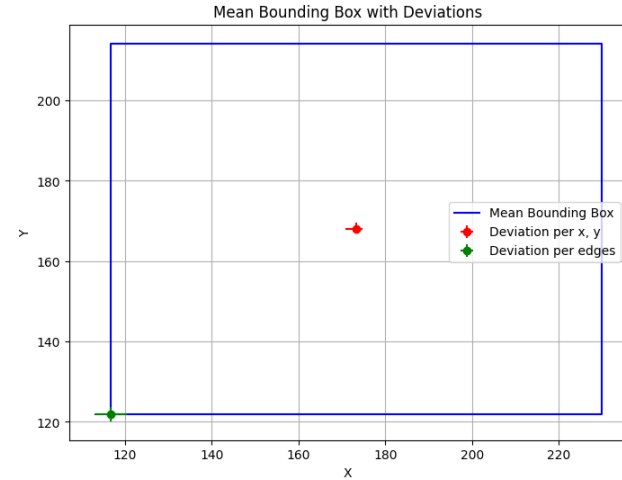
Operation & Monitoring

- OM architecture purpose:
 - Control unknowns
 - Control known uncertainties
- OM constitution:
 - Supervisory monitoring components
 - #1 Anomaly checks:
 - Anomaly detections (input/model/output)
 - #2 Consistency of Interpretability check:
 - Surrogate model
 - #3 Uncertainty estimation check:
 - Uncertainty aware DL module



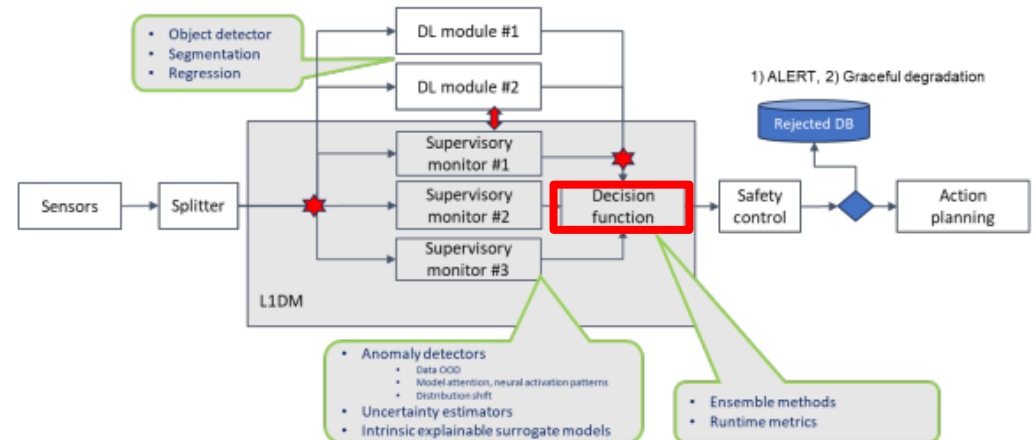
Operation & Monitoring

- OM architecture purpose:
 - Control unknowns
 - Control known uncertainties
- OM constitution:
 - Supervisory monitoring components
 - #1 Anomaly checks:
 - Anomaly detections (input/model/output)
 - #2 Consistency of Interpretability check:
 - Surrogate model
 - #3 Uncertainty estimation check:
 - Uncertainty aware DL module



Operation & Monitoring

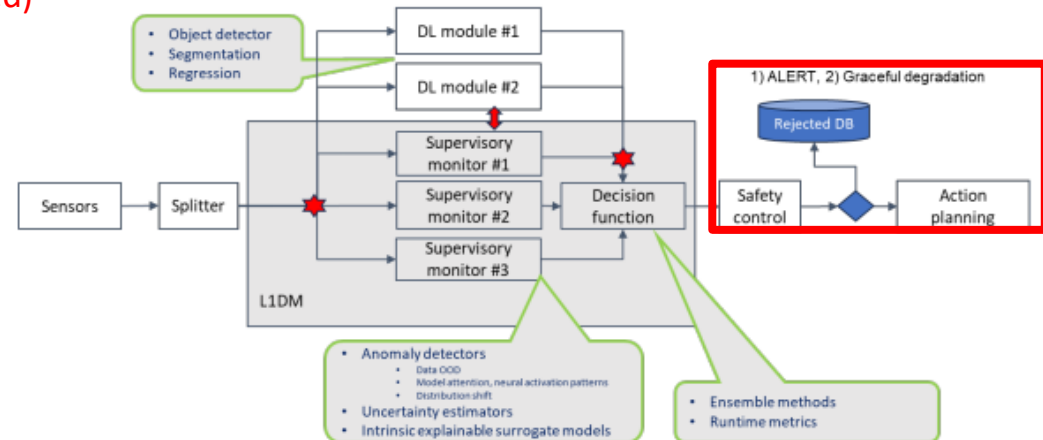
- OM architecture purpose:
 - Control unknowns
 - Control known uncertainties
- OM constitution:
 - Decision function inputs:
 - DL modules predictions
 - Supervisor scores
 - Decision function output
 - Ensembled prediction
 - Trustworthiness score



1) ALERT, 2) Graceful degradation

Operation & Monitoring

- OM architecture purpose:
 - Control unknowns
 - Control known uncertainties
- OM constitution:
 - System can enact safety control (if DL output rejected)
 - Can store rejected data for future versions



Agenda

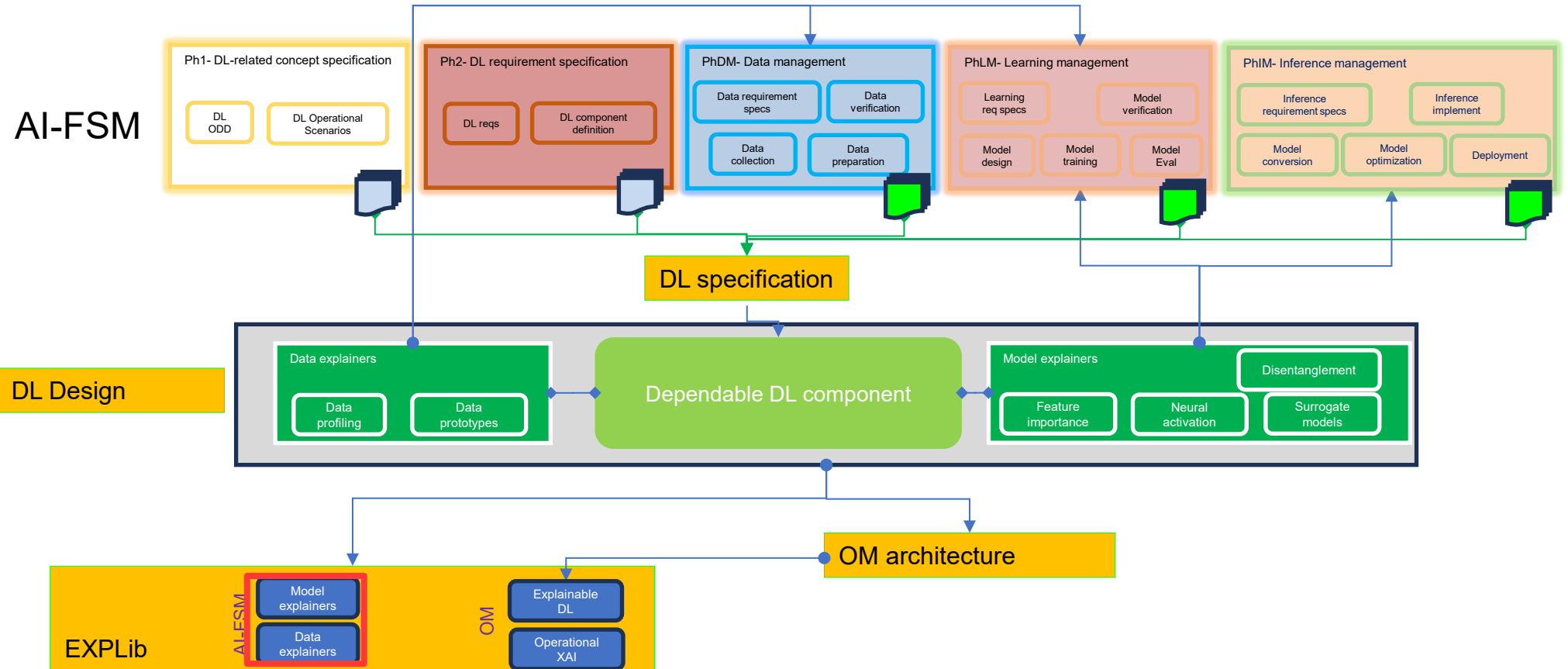
1) Context:

- Functional Safety Management for Artificial Intelligence (AI-FSM)
- XAI for supporting AI-FSM:
 - Data Explainers
 - Model Explainers

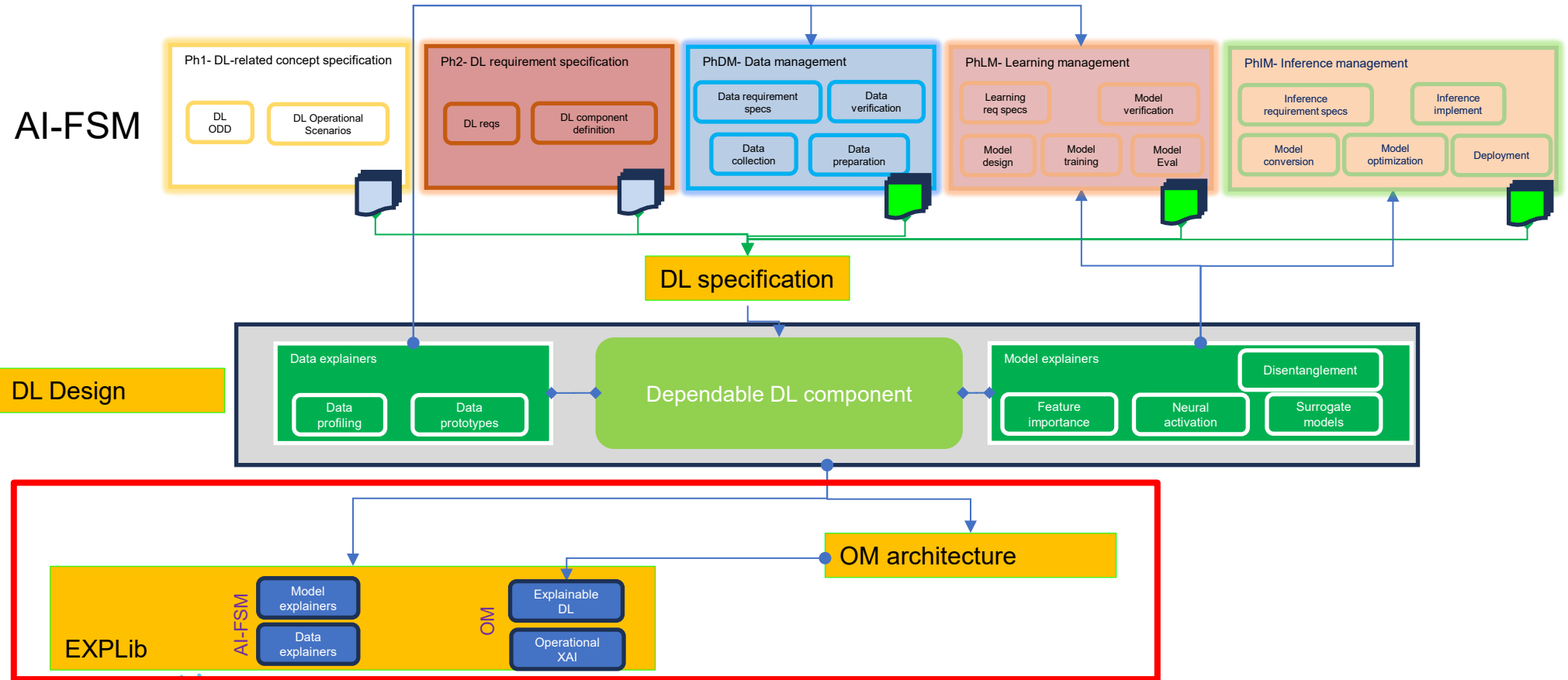
2) Walkthrough – Explainable AI for systems with Functional Safety Requirements:

- Making DL component dependable within the AI-FSM lifecycle
- Operation and Monitoring: Deploying DL component in compliance with Safety Pattern(s)
- **Libraries**

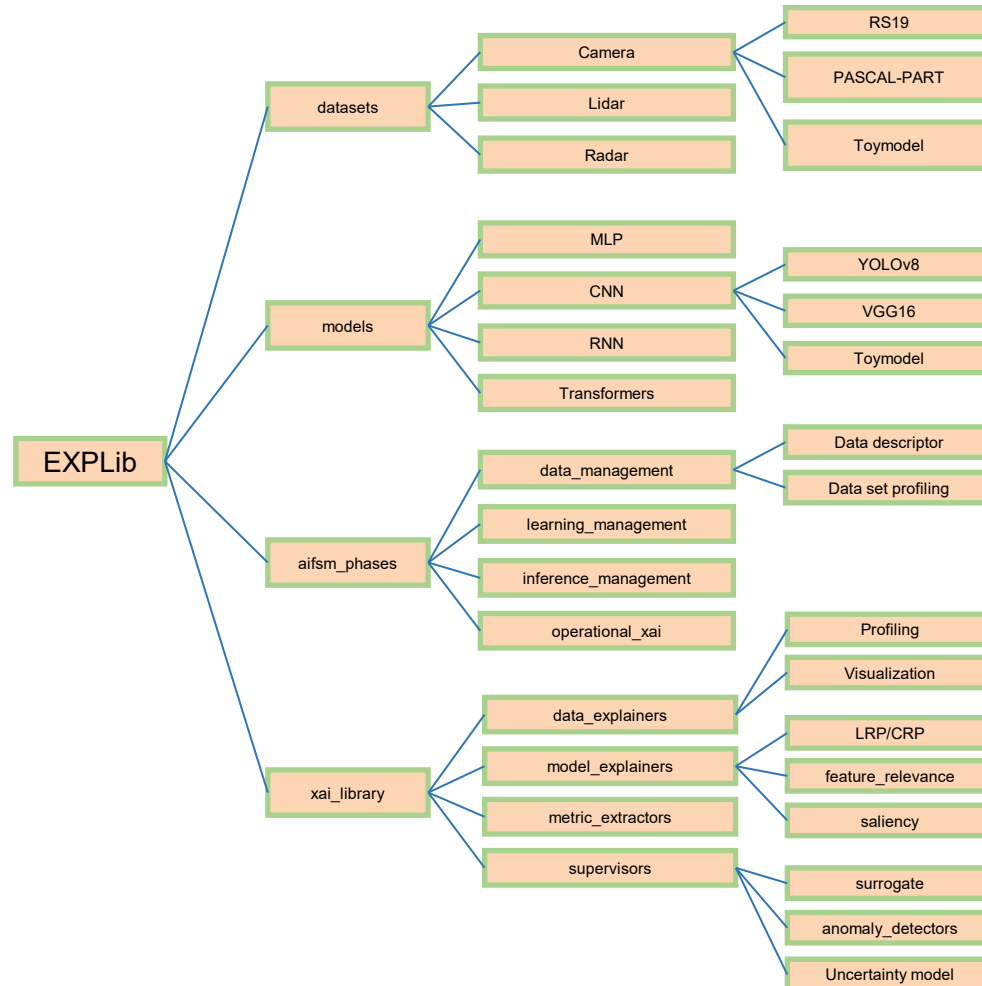
DL explainability in connection with AI-FSM



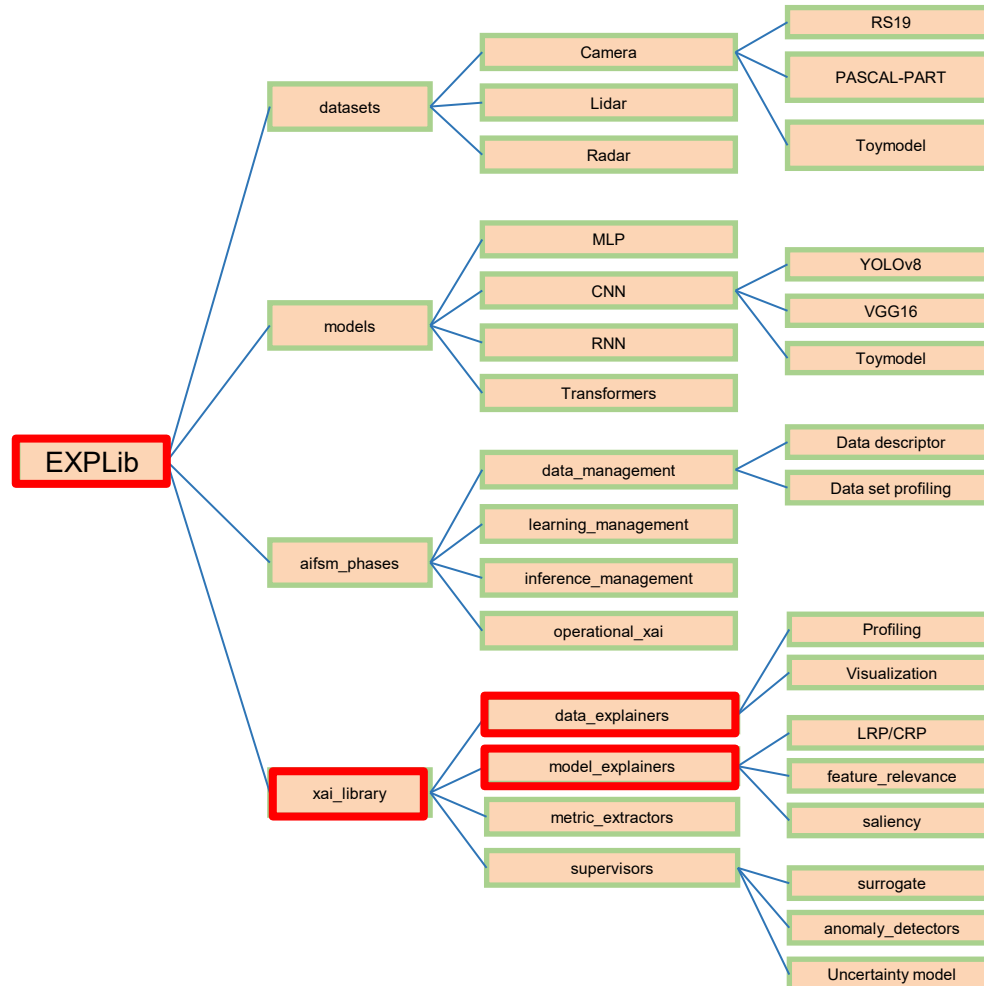
DL explainability in connection with AI-FSM



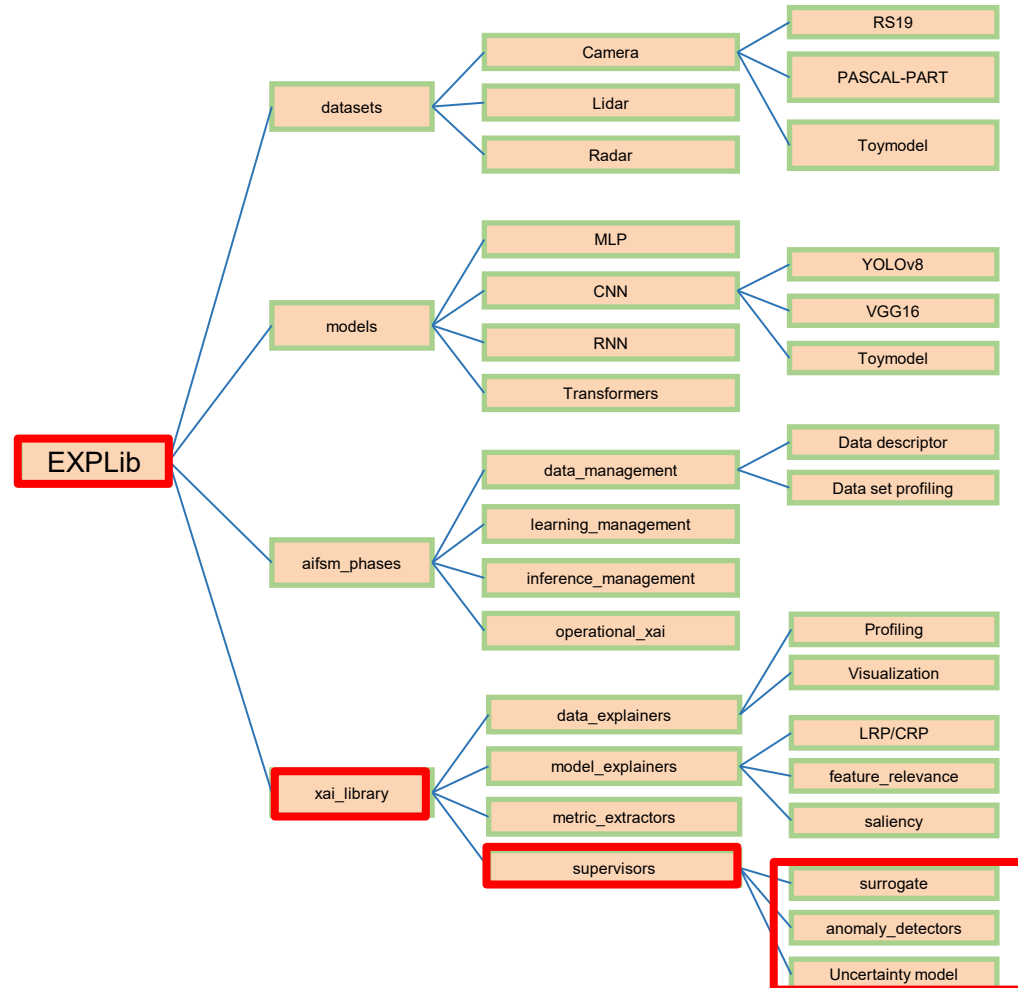
Implementation consideration - EXPLib



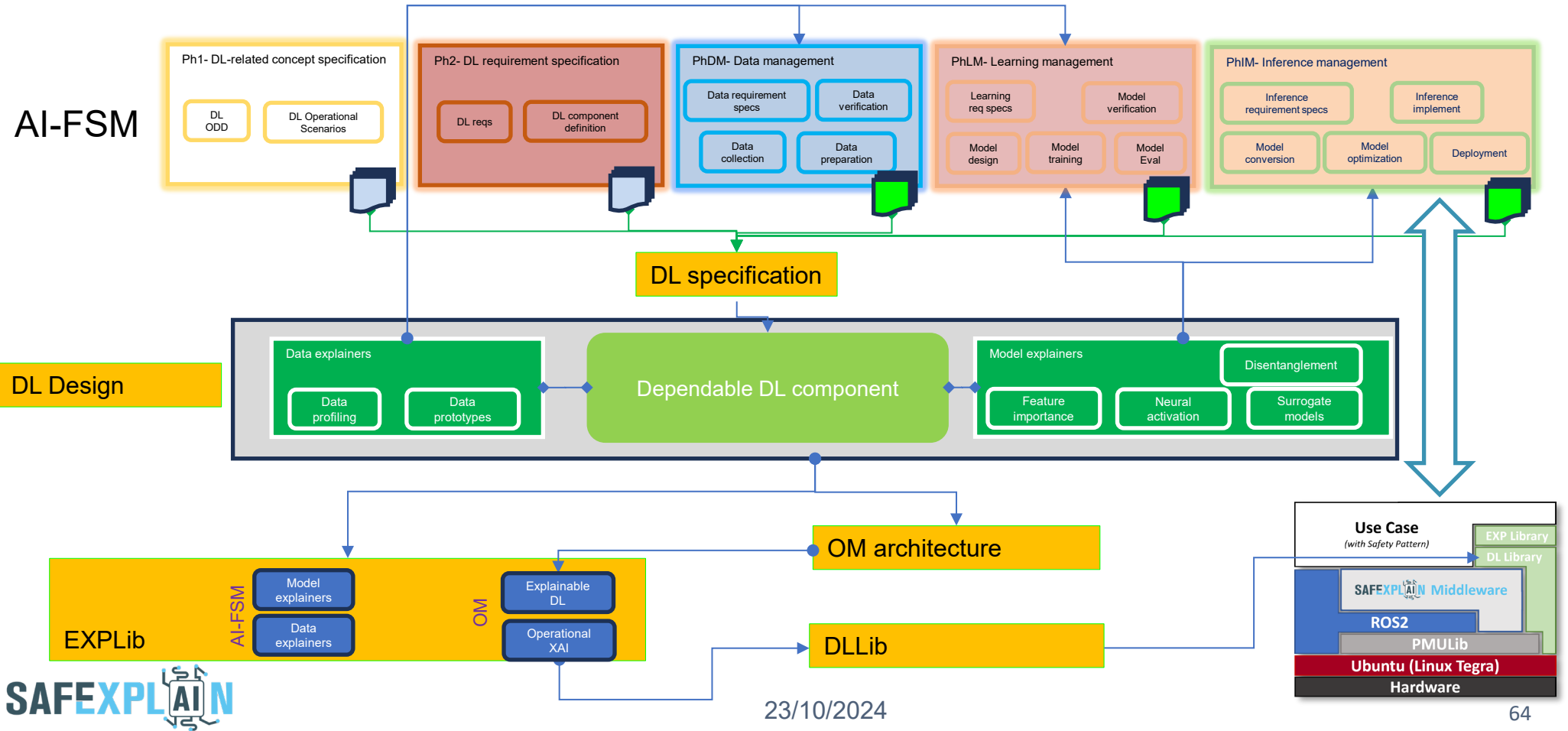
Implementation consideration - EXPLib



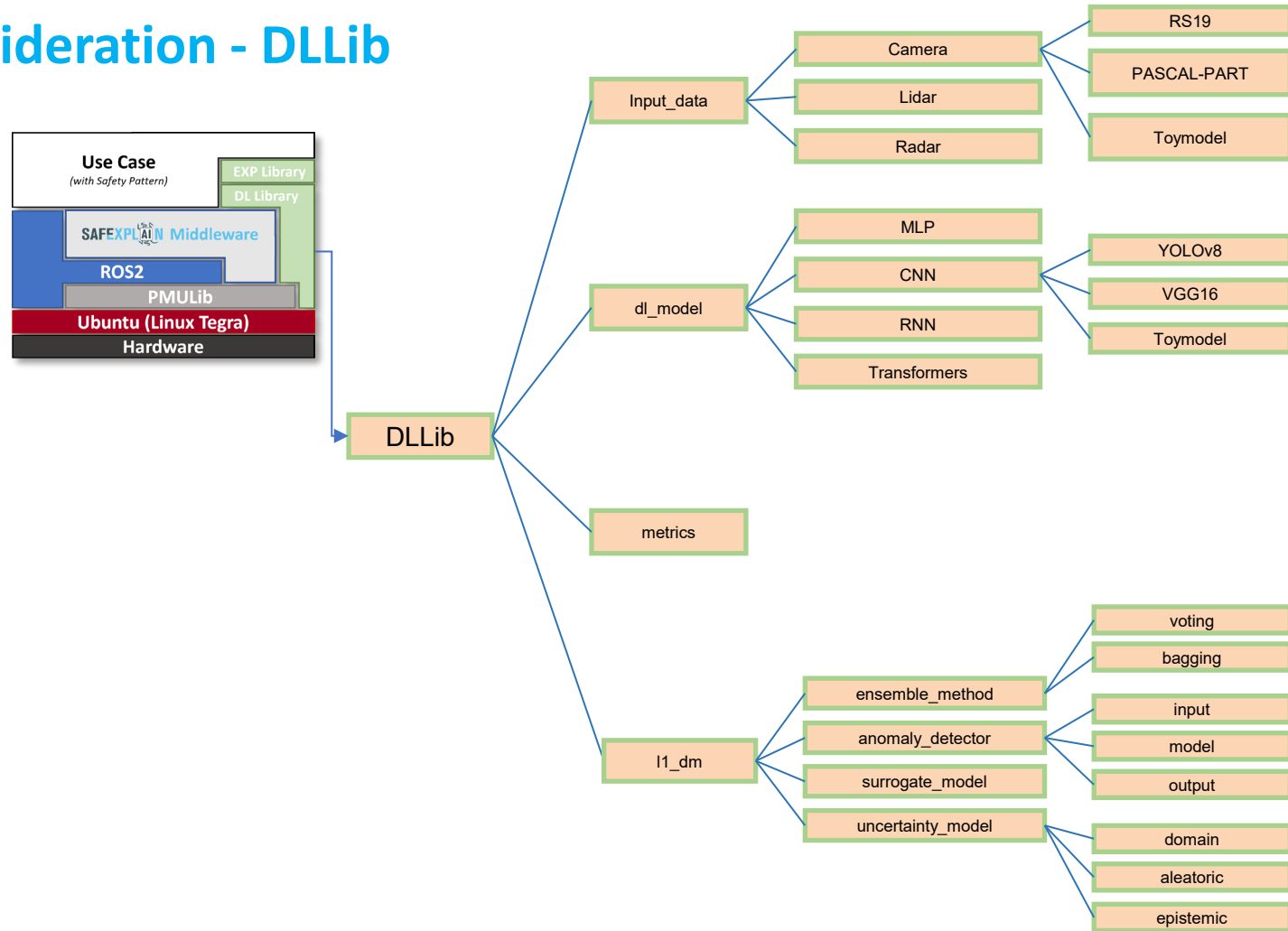
Implementation consideration - EXPLib



DL explainability in connection with AI-FSM



Implementation consideration - DLib





Questions?



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.