AI-FSM: Towards Functional Safety Management for Artificial Intelligence-based Critical Systems

Javier Fernández[†], Irune Agirre[†], Jon Perez-Cerrolaza[†],
Lorea Belategi[†], Ana Adell[†], Carlo Donzella[⋆], Jaume Abella[‡]

Presenter: Javier Fernández - IKERLAN

(‡) Barcelona Supercomputing Center
Centro Nacional de Supercomputación

(⋆) exida

(†) ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Funded by the European Union

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

SAFEXPLAIN

CONTENTS

01 CONTEXTUALIZATION

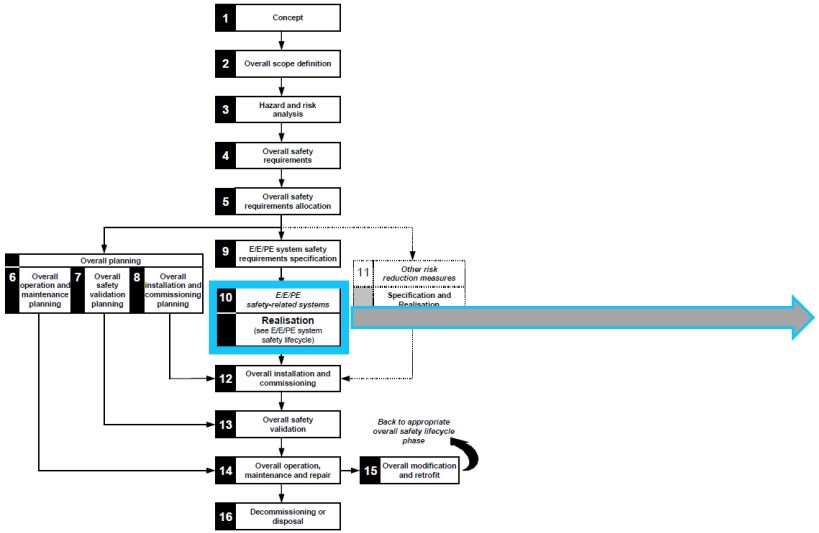
02 PROPOSED LIFECYCLE

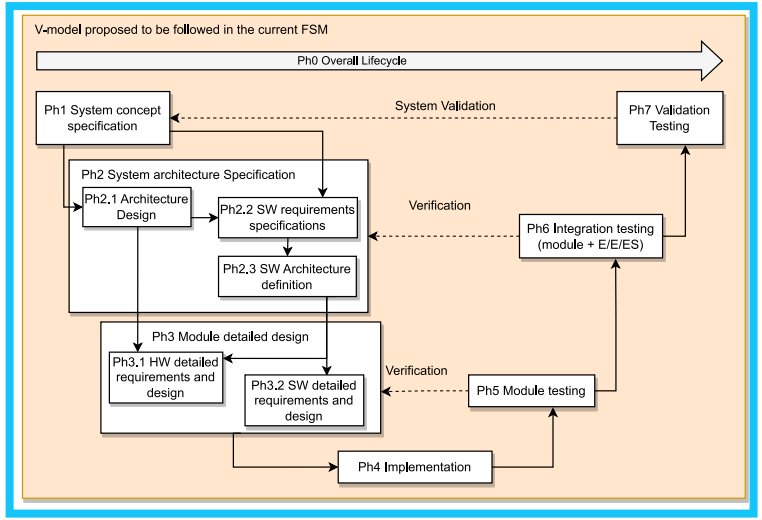
03 CONCLUSION & FUTURE RESEARCH

04 DISCUSSION

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**.

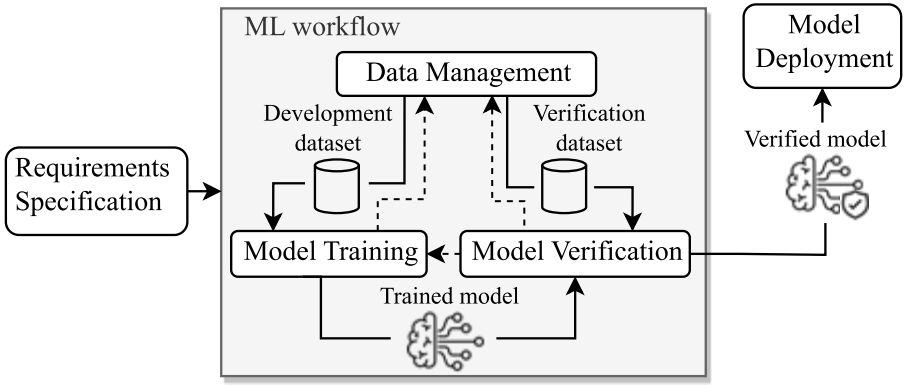




Context

AI lifecycle phases

- Five main stages:
  - Requirements Specification
  - Data Management
    - Development dataset
      - Training + Validation* dataset
    - Verification dataset
  - Model training
    - Trained model
  - Model verification
    - Verified model
  - Model Deployment
    - Inference model



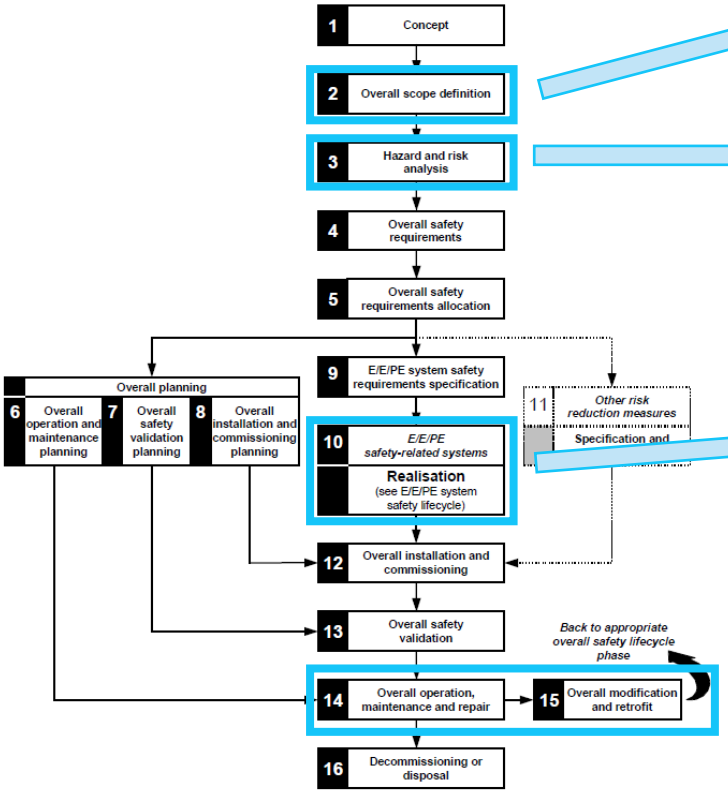
Context

The general DL-based systems development process clashes frontally with traditional safety development processes:

- **DL SW** is designed monolithically following empirical **training processes** with example training data, **rather than implementing specific safety requirements.**
- **DL SW,** as opposed to any other kind of SW in safety critical systems, **cannot be considered as correct by design due to the predictive nature** that comes along with mispredictions and confidence values.
- **DL SW design** is no longer independent of data, **and its parameters are** set empirically **based on training datasets.**
- **DL SW imposes high-performance demands on the underlying HW and its inherent complexity** (both HW and SW) entails challenges to comply with safety standards.

Context

Main phases affected by including DL



- Definition of the ODD and operational scenarios
- HARA shall identify potential hazards caused by the DL-based system. The ODD and operational scenarios are used as input for this stage.
- New phases not contemplated by the traditional V-model:
  - Data management
  - Learning management
  - Inference management
- In traditional software, after a product release an update involves a re-assessment process taking a lot of time. This can be challenging in DL models since their product lifecycle is more likely to be updated.

CONTENTS

01 CONTEXTUALIZATION

02 PROPOSED AI-FSM LIFECYCLE

03 CONCLUSION & FUTURE RESEARCH

04 DISCUSSION

AI-FSM

Scope:

The current version of this AI-FSM is restricted to **DL constituents** with the following features:

- DL algorithms based on **supervised learning** for **visual perception classification tasks**.
- Applications based on **offline learning processes** in which the model remains fixed at approval time, while excluding online learning processes.

The AI-FSM is based on the recommendations drawn from:

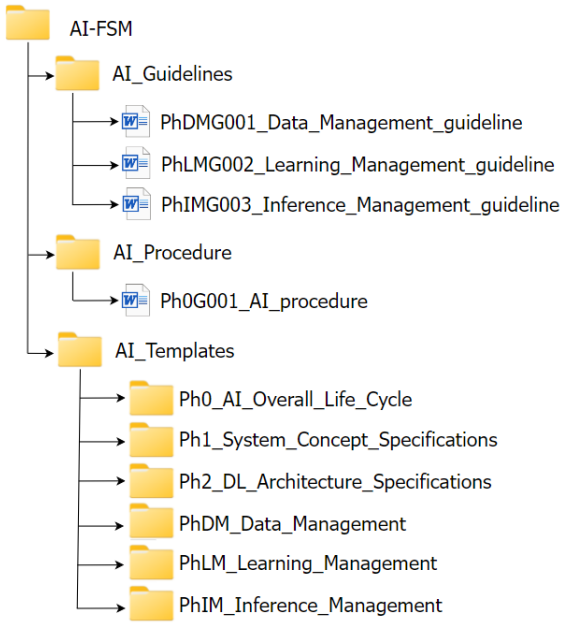
- ISO 26262 and IEC 61508
- IEC TR 5469
- EASA Concept Paper
- AMLAS
- A-SPICE for ML

AI-FSM

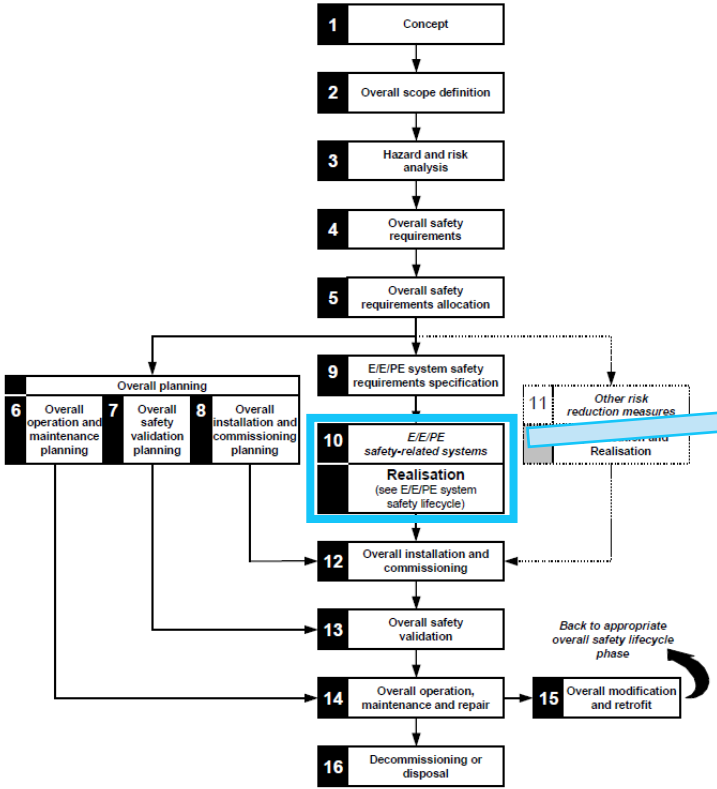
Types of documents:

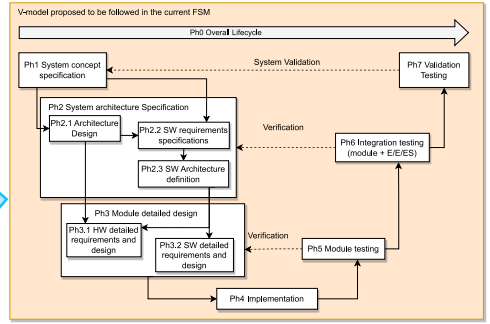
- Main procedure: It provides a set of steps required to generate the basic structure for a specific safety-related project. It serves as an internal guideline for fulfilling the procedure template.
- Procedure template: This document compiles how functional safety has been assessed within the organization.
- Guidelines: These documents offer additional guidance for specific processes.
- Templates: Standard documents used to collect the information consistently. They often include examples and tables to be completed.
- Internal Reviews (IRs): reviews based on the activities of the left side of the safety lifecycle. Objective: Check that the activities defined in each phase have been properly carried out:
  - **Quality Assurance**

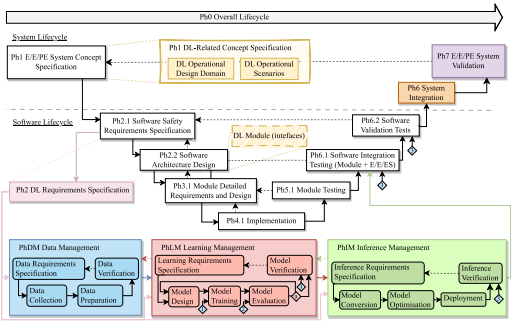
Structure of folders:



Proposed lifecycle



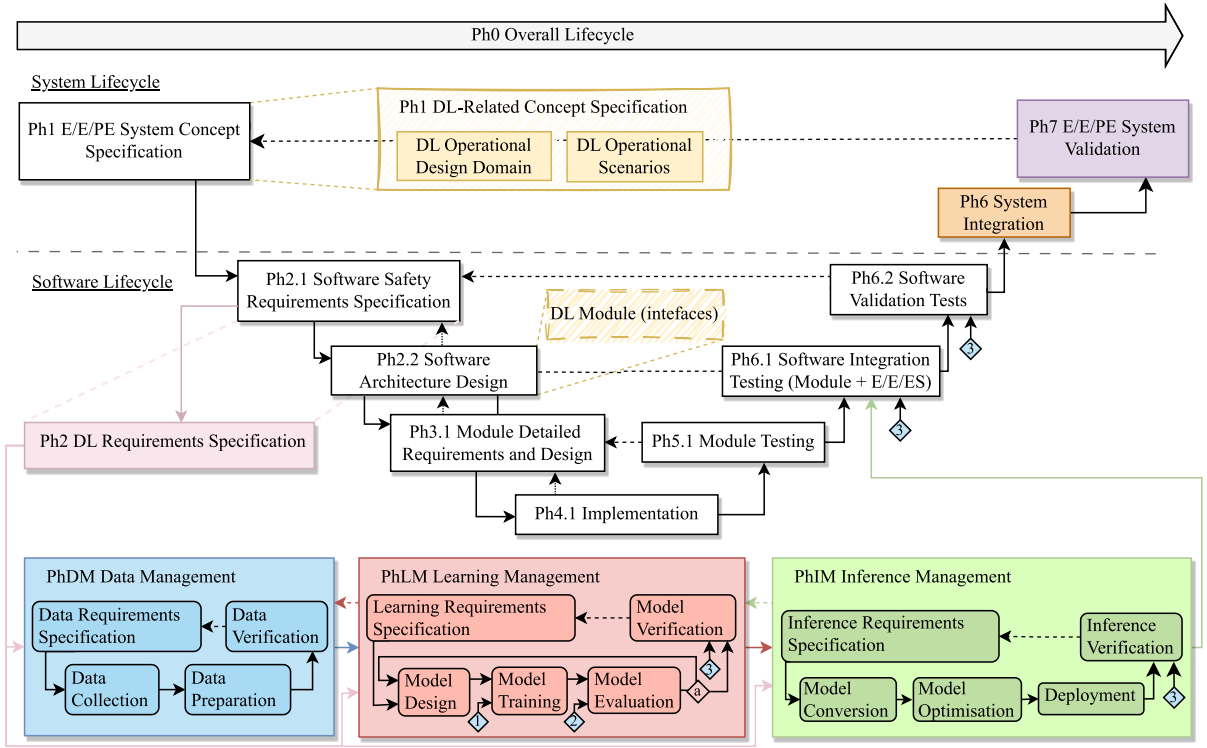




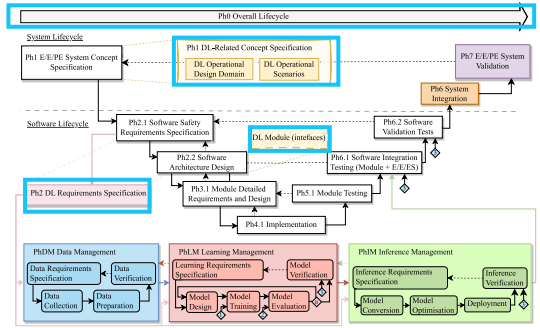
SAFEXPLAIN

Proposed lifecycle

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



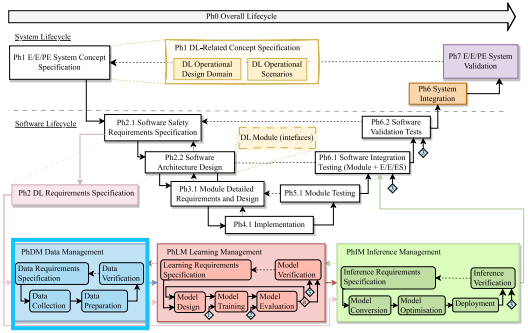
Proposed lifecycle: phases' objectives



- **<u>Ph0 AI Overall Lifecycle:</u>** It is a transversal phase that ***collects*** all the ***generic project information***
  - Documents generated
  - Organization chart
  - Tools selection
- **<u>Ph1 DL-Related Concept Specification:</u>** This phase encompasses the ***definition*** of the ***DL Operational Design Domain (ODD)*** and ***operational scenarios*** in which the DL will operate. In the case the safety-related system entails the use of DL, these definitions are required besides the traditional description of the use case and the definition of the operation reflected in the requirements.
- **<u>DL Modules (interfaces):</u>** This box highlights that Ph2.2 shall define all the interfaces of the DL modules.
- **<u>Ph2 DL Requirements Specification</u>**: This phase ***allocates*** the ***software requirements to DL*** constituents and ***refines them***:
  - Safety, operation, functional and non-functional requirements specification (among others)

SAFEXPLAIN

Proposed lifecycle: phases' objectives



- **<u>PhDM Data Management</u>**. It is responsible for collecting and preparing the datasets. Four steps:
  - <u>Data req. Specifications</u>. It allocates the DL req. to the data req. and refine them. It shall collect:
    - Data and datasets req.
    - Req. Associated with the collection and preparation steps.
    - Data filename policy.
  - <u>Data collection</u>. It involves collecting all the data to generate the datasets:
    - Data gathering. It involves gathering data from different sources.
    - Data generation. It relates to generating new data to complete the data gathering.
  - <u>Data preparation</u>. In this step, the previous data is cleaned, processed, or annotated to meet the reqs.
  - <u>Data verification</u>. This phase checks if the datasets meet the data req. specification.

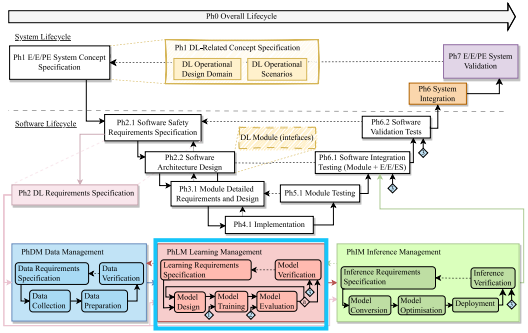
All actions and decisions taken shall be documented

- Inputs:
  - DL reqs specifications
  - ODD
  - Operational scenarios

- Ouputs generated:
  - Development dataset (training + validation*)
  - Verification dataset

SAFEXPLAIN

Proposed lifecycle: phases' objectives



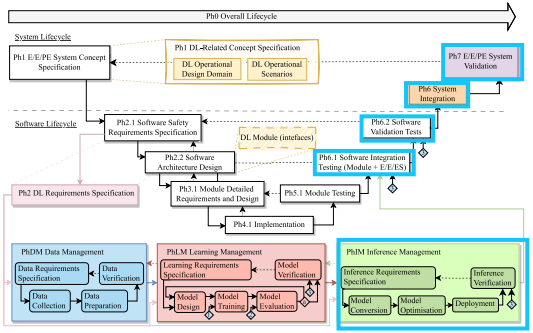
- **<u>PhLM Learning Management</u>**. It is responsible for generating a DL model that meets the DL req. specification. Five steps:
  - <u>Learning req. Specifications</u>. It allocates the DL req. to learning reqs. and refine them. It shall collect:
    - Qualitative and quantitative learning reqs.
    - Model post-training selection criteria.
    - Reqs. associated with the model design and training.
  - <u>Model design</u>. It focuses on the specification of a set of DL models that best suit the application.
  - <u>Model training.</u> In this step, the specified models are generated employing the training dataset.
  - <u>Model evaluation.</u> Once the model(s) are trained, they are evaluated employing the validation dataset.
  - <u>Model verification</u>. This step not only evaluates the generalization capabilities and identifies potential issues using the verification dataset but also checks if the reqs. are met.

All actions and decisions taken shall be documented

- Inputs:
  - Development dataset (training + validation*) from PhDM
  - Verification dataset from PhDM
  - DL req. specification
- Ouputs:
  - Trained model
  - Evaluated model
  - Verified learning model

SAFEXPLAIN

Proposed lifecycle: phases' objectives



- **PhIM Inference Management**. Its purpose is to adapt the verified model for its deployment on the target HW while ensuring that it still meets the DL reqs. after converting and even optimising it. Five steps:
  - Inference req. specification. It allocates the DL and learning reqs. to inference reqs. and refine them. It shall collect:
    - Inference reqs.
    - Req. associated with the model conversion, optimization and deployment

All actions and decisions taken shall be documented

  - Model conversion. The model is transformed into a format suitable for deployment that must ensure compatibility with the specific target inference platform.
  - Model optimisation. the model may undergo optimization to enhance its performance, reduce its size, or adapt it for resource-constrained environments.
  - Deployment. This steps entails the implementation of the model in the target platform.
  - Inference verification. This phase not only evaluates the generalization capabilities and identifies potential issues using the verification dataset but also checks if the reqs. are met.
  - Input:
    - Verified learning model from PhLM
    - Verification dataset from PhDM
    - Learning and DL req. specification
  - Ouput:
    - Verified inference model

CONTENTS

01 CONTEXTUALIZATION

02 PROPOSED LIFECYCLE

03 CONCLUSION & FUTURE RESEARCH

04 DISCUSSION

Conclusions and Future Work

In recent years various emerging initiatives and standards have been developed to align the use of AI in safety-critical systems, but it remains an open research challenge. In this context, this paper addresses the updates required in the realization phase of traditional functional safety standards to support the safe development of AI in safety-critical systems.

In the future, the AI-FSM may be updated to align with upcoming iterations of emerging standards such as:

- **ISO/CD PAS 8800**: Road Vehicles - Safety and artificial intelligence
- **IEC TS 6254**: Information technology - Artificial intelligence - Objectives and approaches for explainability and interpretability of ML models and AI systems

CONTENTS

01 CONTEXTUALIZATION

02 PROPOSED LIFECYCLE

03 CONCLUSION & FUTURE RESEARCH

04 DISCUSSION

DISCUSSION

QUESTIONS ?

THANK YOU!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

exida

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

SAFEXPLAIN

NAME: JAVIER FERNÁNDEZ

EMAIL: JAVIER.FERNANDEZ@IKERLAN.ES

Acknowledgements:

- The research leading to these results has received funding from the European Union's Horizon Europe Programme under the SAFEXPLAIN Project (www.safexplain.eu), grant agreement num. 101069595.
- Jaume Abella has also been partially supported by the Spanish Ministry of Science and Innovation under grant PID2019-107255GB-C21/AEI/10.13039/501100011033.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

exida

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

SAFEXPLAIN