How to use out-of-distribution detection method to argue AI/ML-based components into a safety lifecycle and related safety argument

Murat Erdogan- Senior Manager Confirmation Measures

11/27/2023

Motivation

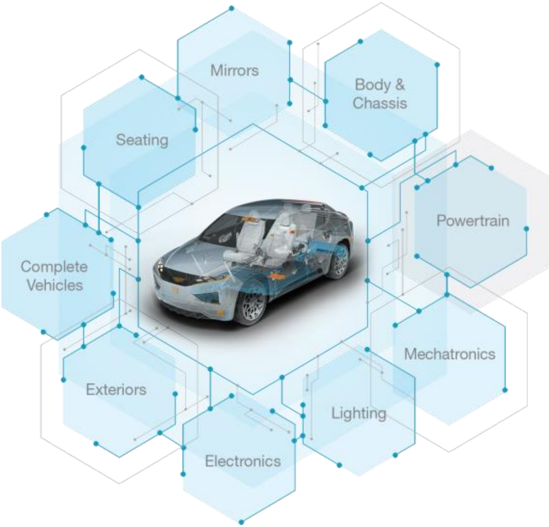
How can we increase safety for all road users?

MAGNA

Trends driving the evolution of mobility

Challenges

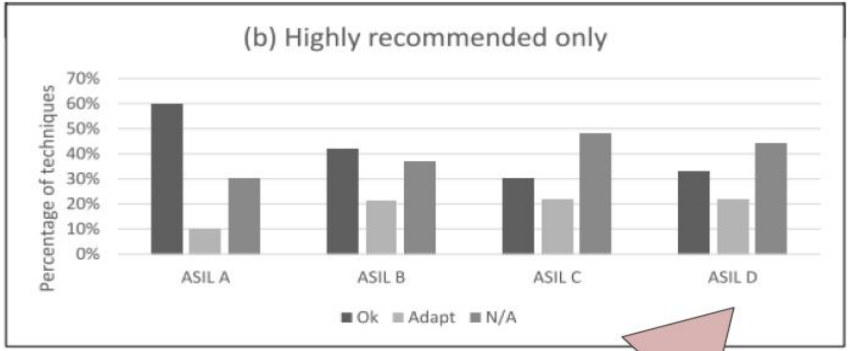
Challenge



REGULATIONS → **STANDARDS** → **COMPLEXITY**

Challenge

ISO 26262 + ML = ?



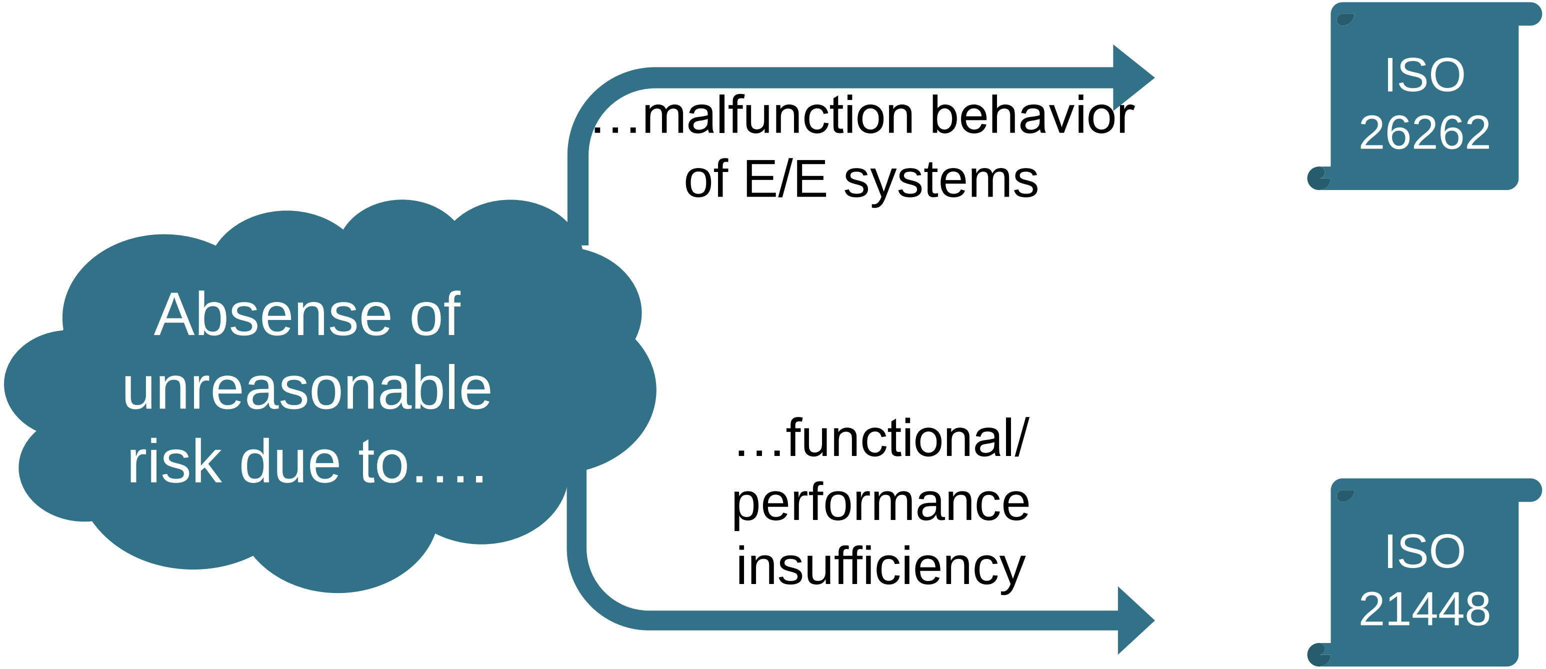
34 techniques:
Majority not
applicable for ML

Ref.: [1]

Why not applicable?

- Non-deterministic nature of ML components
- V&V methods in ISO 26262 cannot be used to create KPIs for ML components

Challenge


Absense of unreasonable risk due to….
…malfunction behavior of E/E systems
ISO 26262
…functional/ performance insufficiency
ISO 21448

Background

SPRINGER LINK

Find a journal   Publish with us   Search

International Working Conference on Requirements Engineering: Foundation for Software Quality

↳ REFSQ 2023: **Requirements Engineering: Foundation for Software Quality** pp 233–242 | Cite as

Home > Requirements Engineering: Foundation for Software Quality > Conference paper

Out-of-Distribution Detection as Support for Autonomous Driving Safety Lifecycle

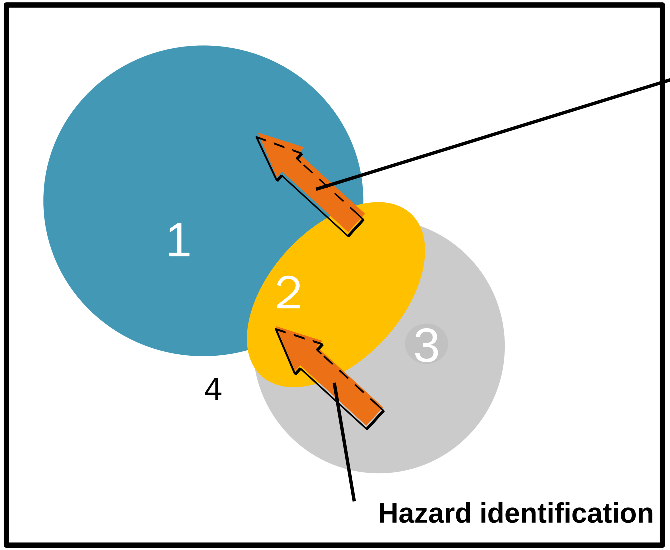
Jens Henriksson, Stig Ursing, Murat Erdogan, Fredrik Warg, Anders Thorsén, Johan Jaxing, Ola Örsmark & Mathias Örtenberg Toftås

Conference paper | First Online: 04 April 2023

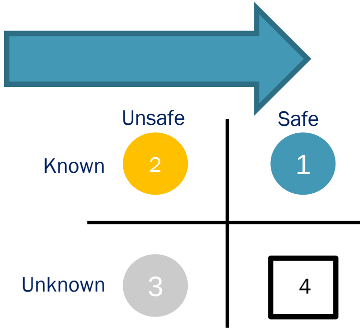
500 Accesses | **1** Citations

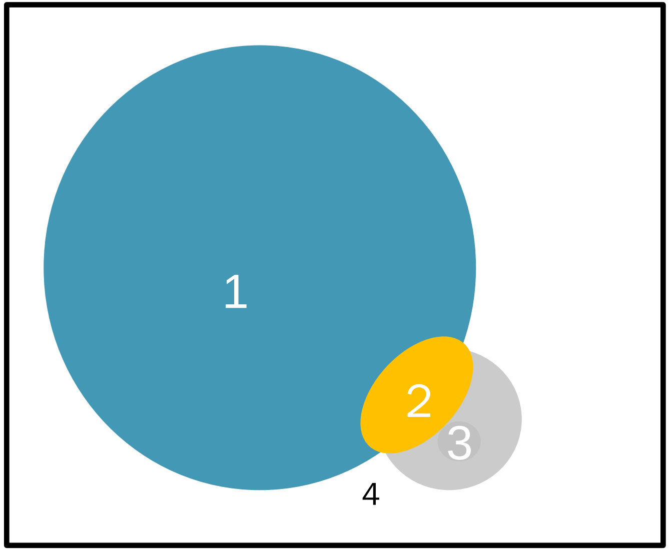
Part of the Lecture Notes in Computer Science book series (LNCS,volume 13975)

Hypothesis for OoD-Detection



Hazard mitigation
- Improving function
- Restricting performance
- Introducing redundancy
- etc…



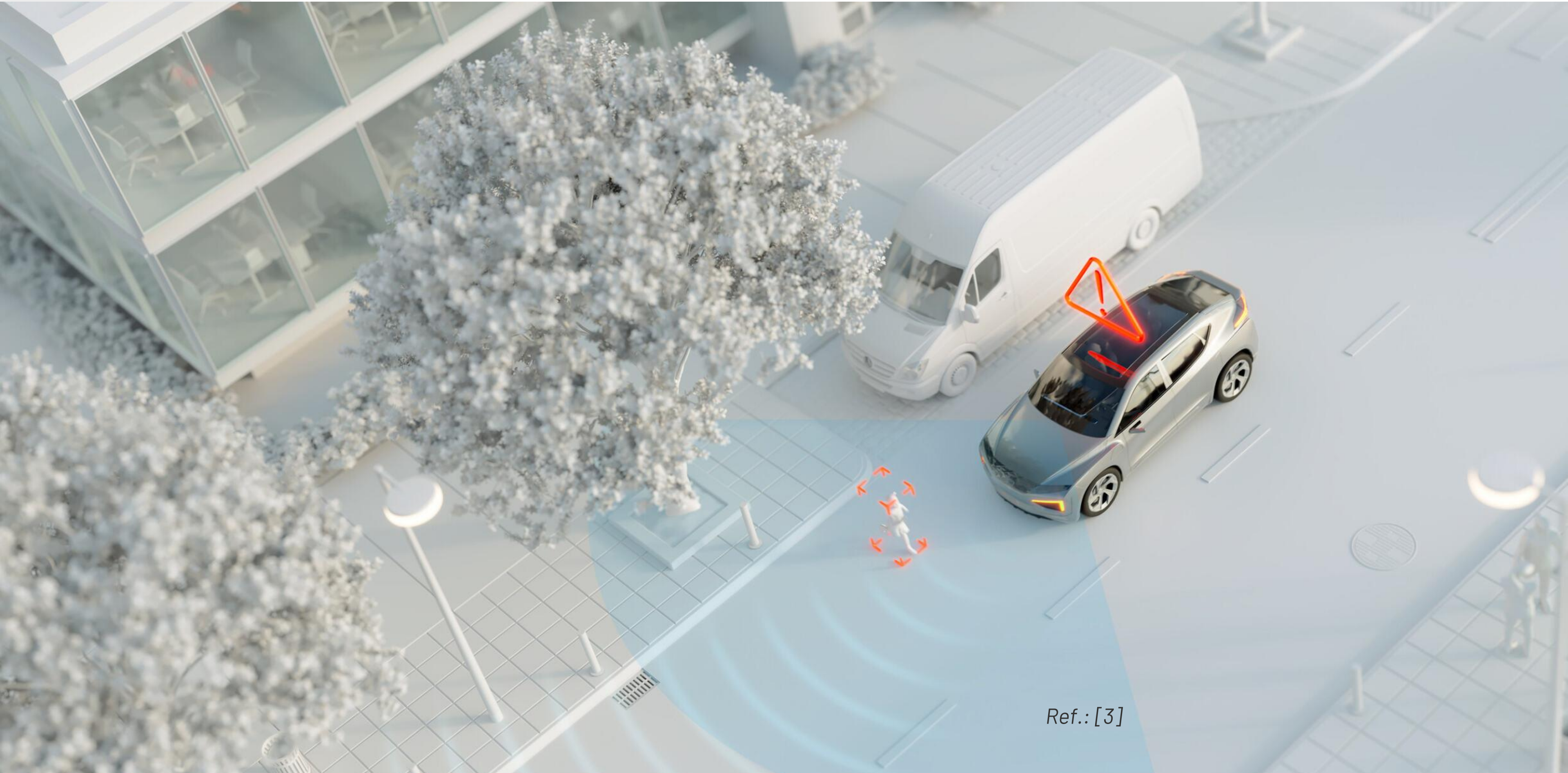


Out-of-distribution detection is one mitigation strategy that shall support hazard mitigation by providing a distance measure of how far off the model operates compared to linear data (known safe states)

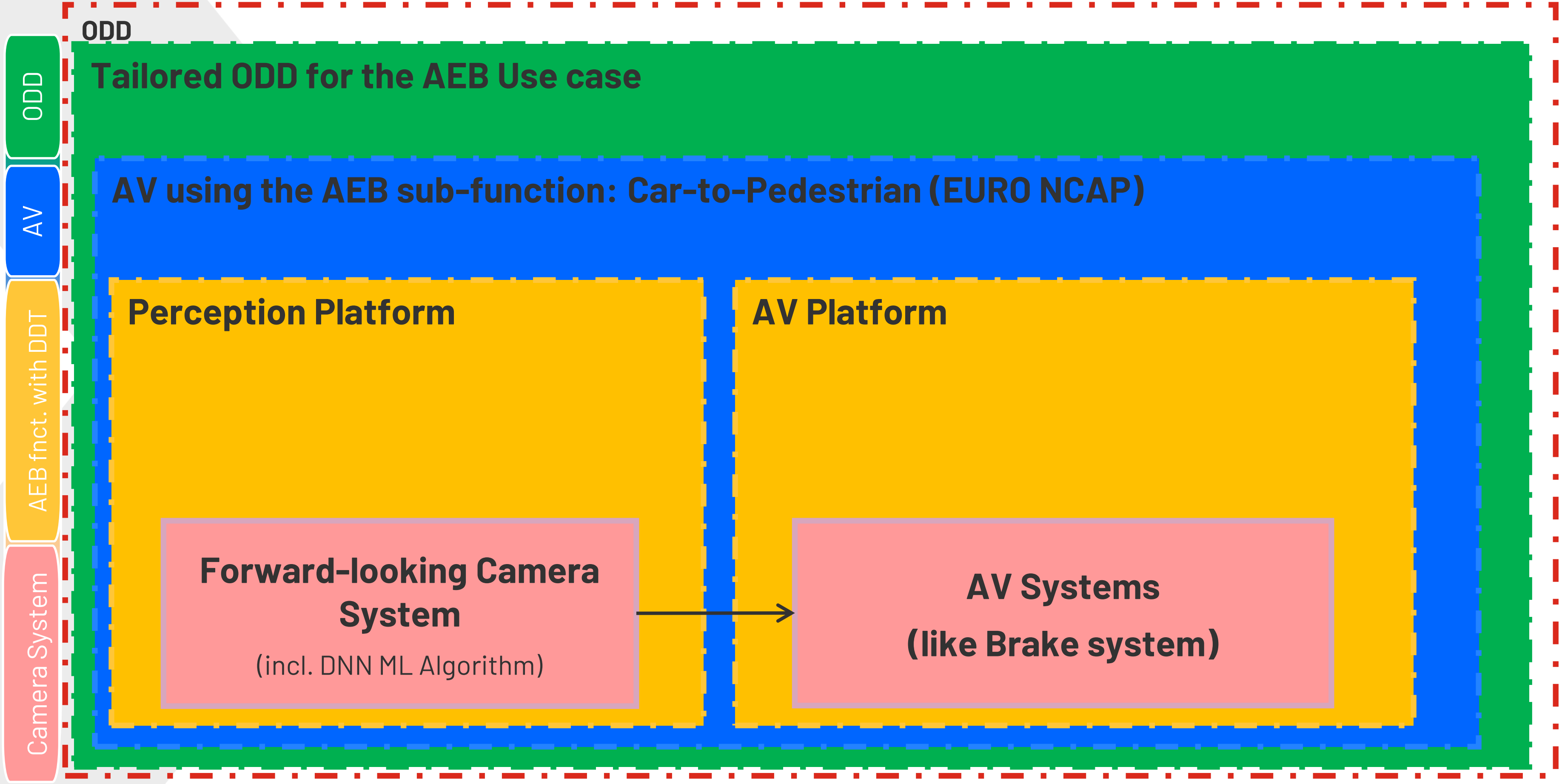
Ref.: [2]

AEB Use-Case

AEB Use-Case: Pedestrian Detection

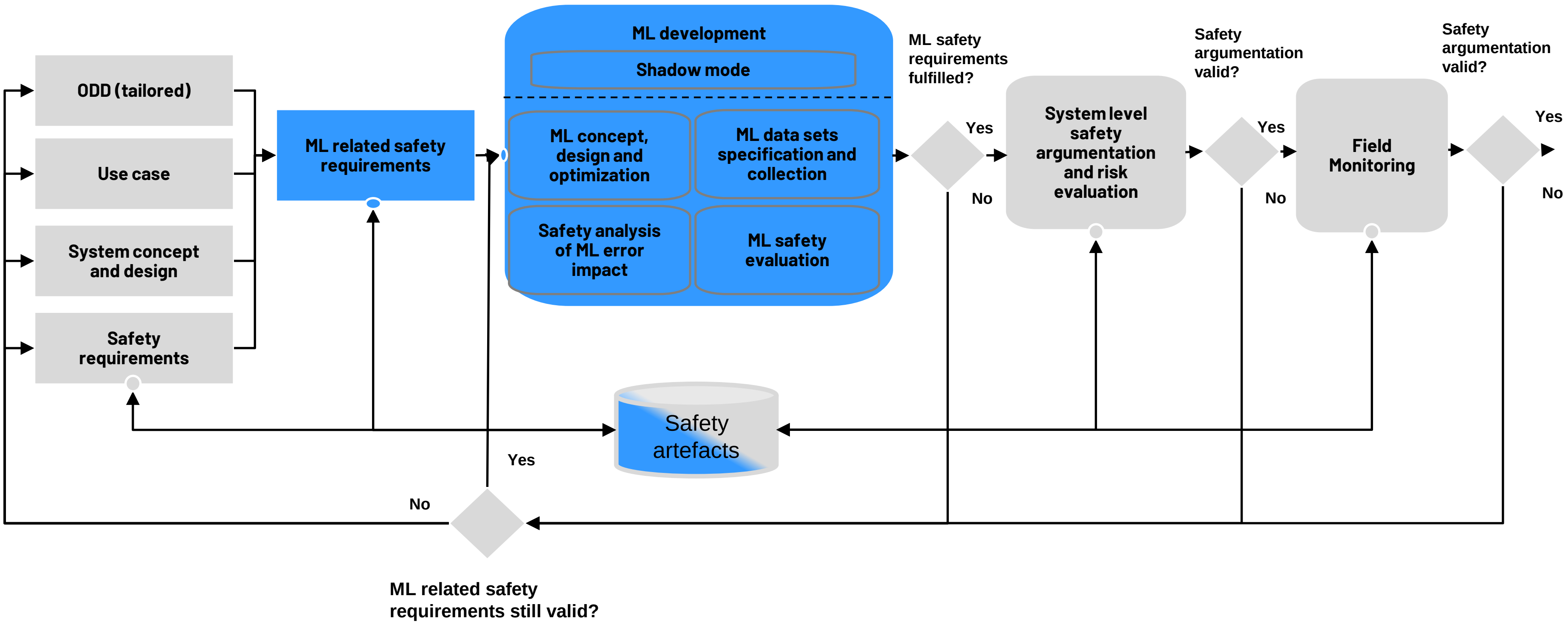


Holistic System Abstraction Levels



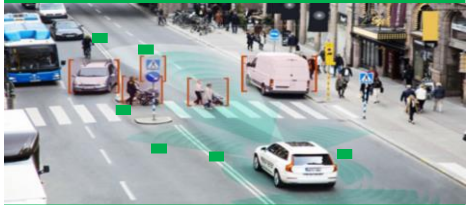
Ref.: [3]

Development Lifecycle



Holistic System Abstraction Levels

ODD



AV

PLTFRM 1
PLTFRM 2
PLTFRM 3

PLTFRM 1

FLC | ADAS | Brake

Systems supporting ADS fnct. with DDT

Forward-looking camera

ISO 26262

3.74- harm
physical injury or damage to the health of persons

3.75- hazard
potential source of harm caused by malfunctioning behaviour of the item

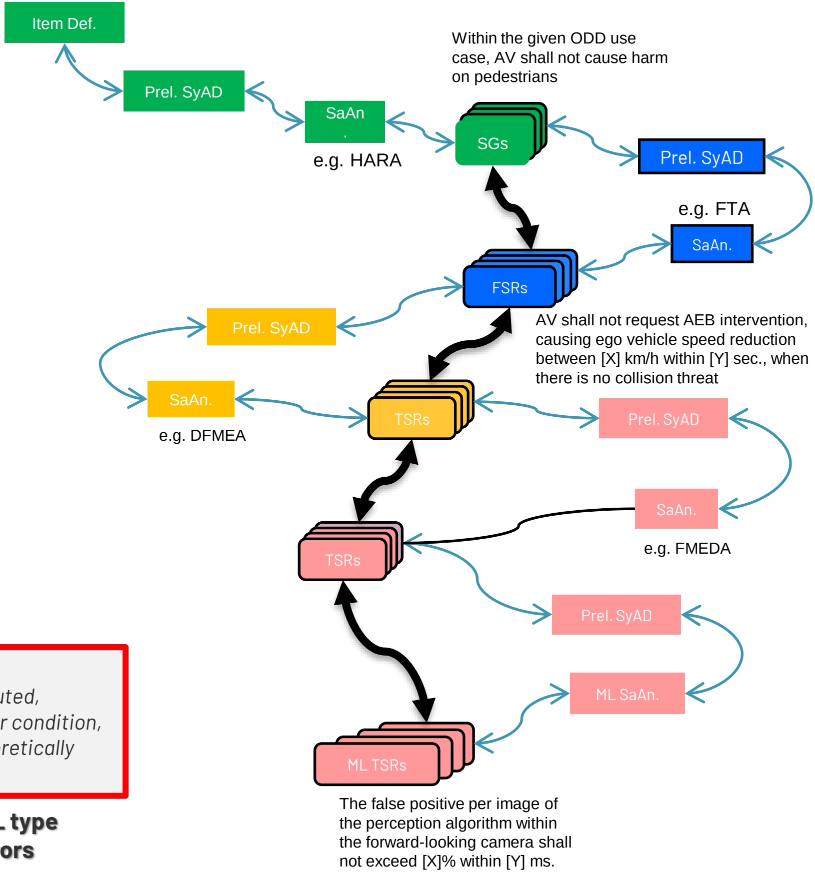
3.88- malfunctioning behaviour
failure or unintended behaviour of an item with respect to its design intent

3.50- failure
termination of an intended behaviour of an element or an item due to a fault Manifestation.

3.50- error
discrepancy between a computed, observed or measured value or condition, and the true, specified or theoretically correct value or condition

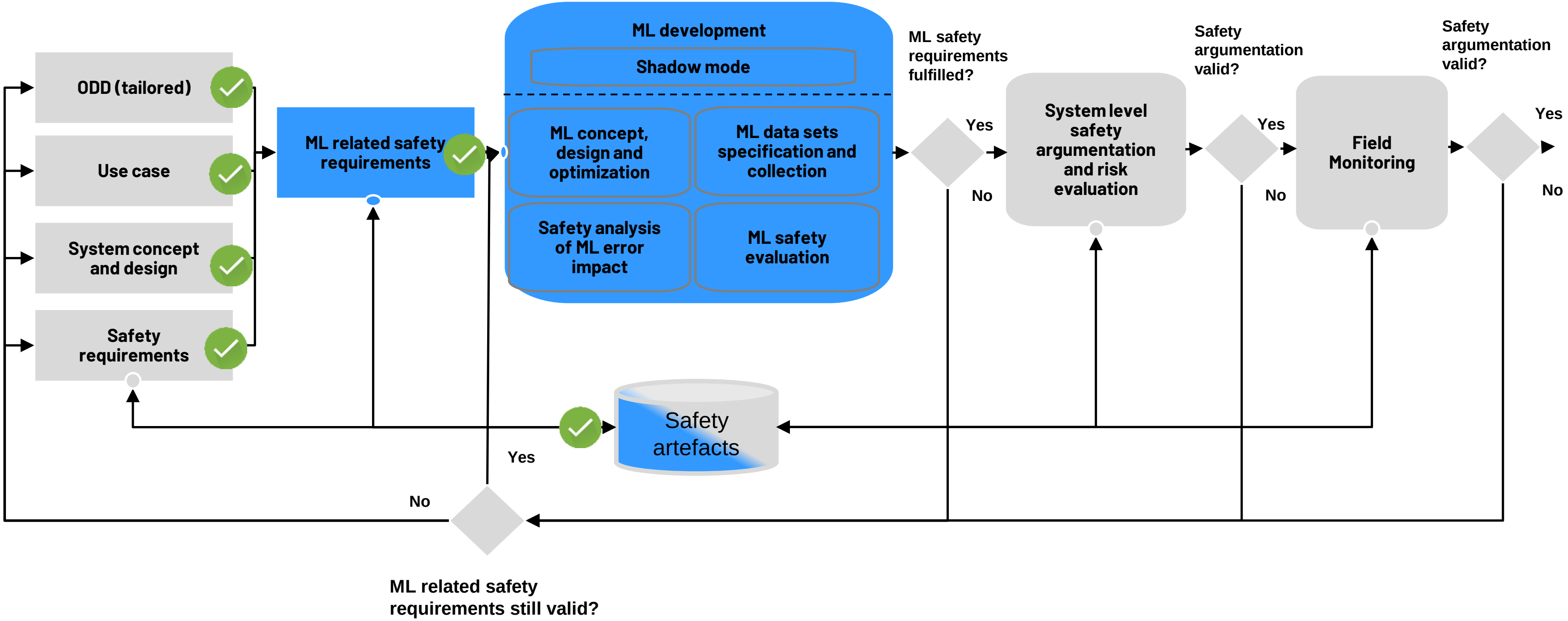
3.50- fault
abnormal condition that can cause an element or an item to fail

AI/ML type of errors

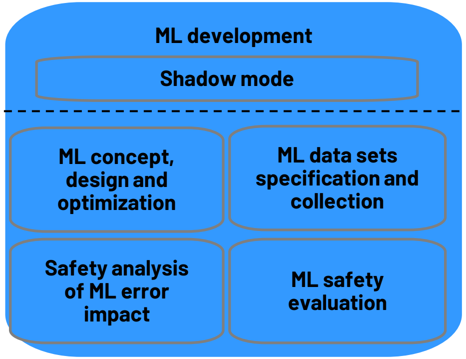


Ref.: [5]

Development Lifecycle (cont.)

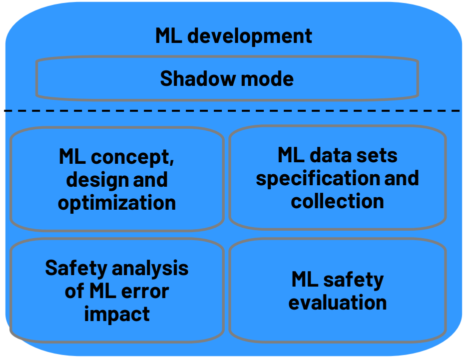


Development Lifecycle (cont.)



Development Lifecycle (cont.)

MAGNA



ISO 26262

3.74- harm
physical injury or damage to the health of persons

3.75- hazard
potential source of harm caused by malfunctioning behaviour of the item

3.88- malfunctioning behaviour
failure or unintended behaviour of an item with respect to its design intent

3.50- failure
termination of an intended behaviour of an element or an item due to a fault Manifestation.

3.50- error
discrepancy between a computed, observed or measured value or condition, and the true, specified or theoretically correct value or condition

3.50- fault
abnormal condition that can cause an element or an item to fail

AI/ML type of errors

Examples:
- *False positive*
- *False negative*
- *incorrect classification*
- *Inaccurate estimation*
- *...*

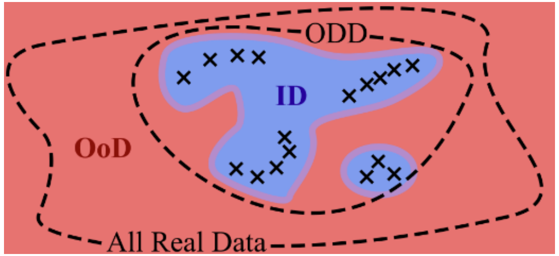
Insufficiencies
Examples:
- *Bias*
- *Lack of robustness*
- *Lack of generalization*
- *Variance*
- *...*

Causes
Examples:
- *Scalable oversight*
- *Distributional shift*
- *Under specification*
- *Overtraining*
- *...*

Ref.: [4]

Model Development

MAGNA



discrete data points in the input data space

area where the model has comprehensed to ID

area where the model has failed to OOD

ML development

Shadow mode

ML concept, design and optimization

ML data sets specification and collection

Safety analysis of ML error impact

ML safety evaluation

Process for Model training

Labeling format

OoD

Architecture type

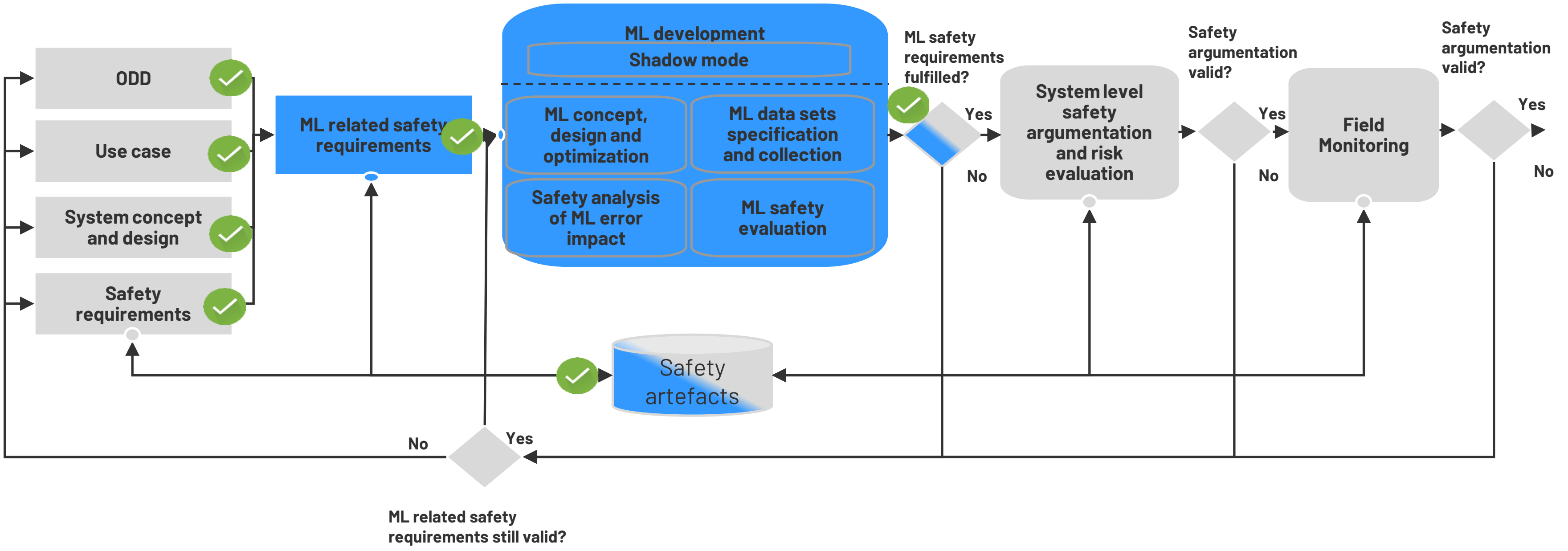
Optimization methods

Safety Analysis

Reference Paper

Distributional shift

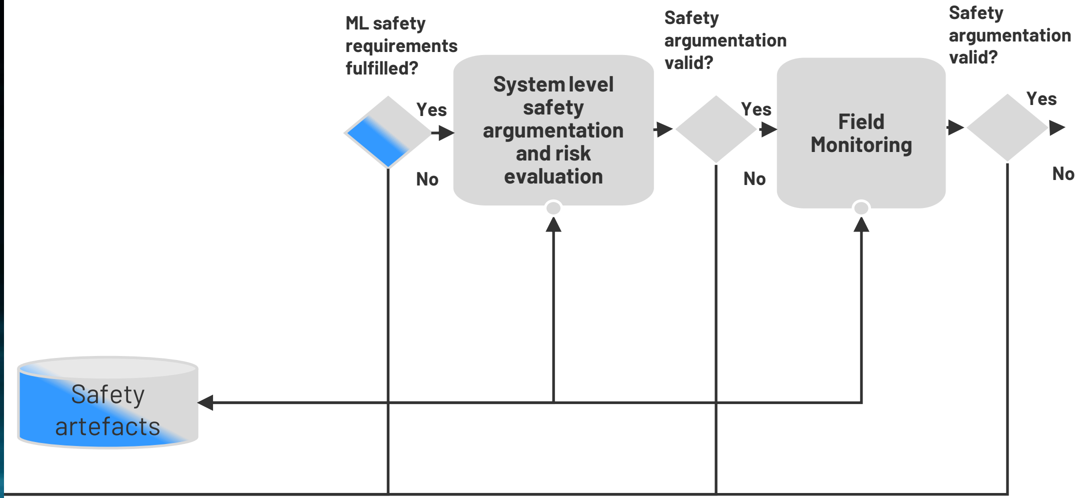
Development Lifecycle (cont.)



Remaining Lifecycle

Integration to next level systems and operational monitoring:

- Condition to start the integration?
  - ➢ Acceptance criteria e.g. sufficient coverage of the allocated safety requirements
- Shadow mode operation can be used to:
  - ➢ Deploy the feature in the background to gather data
  - ➢ Collect and classify data as ID and OoD
  - ➢ Identifying the challenging scenarios for the model
  - ➢ Supporting the model improvement



Conclusion

Conclusion

Outcome

- Need of new standards for AI/ML based components in automotive
- Introduction of system abstraction layers
- Introduction of OoD detection method as one of guiding principles

Outlook

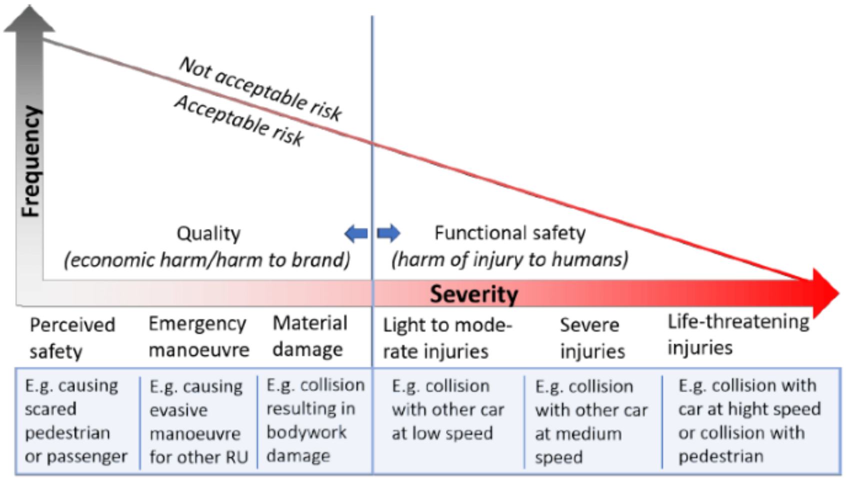
- Continuation on the development lifecycle
- Strategies for arguing ML-components into safety argument

SALIENCE4CAV

About the Project

Connected automated vehicles (CAV) are expected to be able to provide more efficient, accessible and safer transport solutions, but the development of such complex safety-critical systems is a challenge. SALIENCE4CAV is a research project with the goal of developing methods for safety assurance for CAVs that can be part of an iterative development process supporting continuous deployment. This enables easier introduction of new automated functions to the market, where the function can initially support only a few key use cases, followed by gradual development of performance and number of use cases.

SALIENCE4CAV is a successor to the project ESPLANADE. The project started on January 1st 2021 and will run to June 30th 2023.



Source: The Quantitative Risk Norm - A Proposed Tailoring of HARA for ADS

The SALIENCE4CAV project (ref. 2020-02946) is supported by the Strategic vehicle research and innovation programme (FFI).

Reference

Ref. Nr.	Organisation, Conference	Title	Publisher	Publishing date
1	University of Waterloo ON, Canada	An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software	Rick Salay, Rodrigo Queiroz and Krzysztof Czarnecki	September 2017
2	Chalmers University of Technology Göteborg, Sweden	Outlier Detection as a Safety Measure for Safety Critical Deep Learning	Jens Henriksson	October 2023
3	10th Scandinavian Conference on System and Software Safety	An AEB use-case approach for robustness and safety using AI and ML for autonomy	Murat Erdogan	November 2022
4	Fraunhofer IKS	A causal model of safety assurance for machine learning	Prof. Simon Burton	2022, Munich
5	Springer Nature 2021	Ergo, SMIRK is Safe: A Safety Case for a Machine Learning Component in a Pedestrian Automatic Emergency Brake System	Markus Borg, Jens Henriksson, Kasper Socha, Olof Lennartsson, Elias Sonnsjö Lönegren,Thanh Bui, Piotr Tomaszewski, Sankar Raman Sathyamoorthy, Sebastian, Brink and Mahshid Helali Moghadam	September 2022


Forward.
For all.
MAGNA