33rd Annual INCOSE international symposium

hybrid event

Honolulu HI USA

Marc Zeller | Siemens AG

Safety Assurance of Autonomous Systems using Machine Learning: An Industrial Case Study and Lessons Learnt

Why is safety assurance for AI-based Systems hard?

- Uncomplete specification ("specification by example")
- Highly non-linear behavior
- Very complex structure / state space explosion
- Hard to decompose into human-understandable functional blocks
- Safety assessment methodology needs to adapt to technology, domain, and use case

Motivation

- Artificial Intelligence (AI) / Machine Learning (ML) helps to implement novel functionalities (e.g. autonomous vehicles/trains, AGVs in factories, etc.)
- Safe systems incorporating AI/ML are required for many industrial use cases
- Safe AI is based on quality measures, quantitative performance and a thorough understanding of the system and methodologies for verification

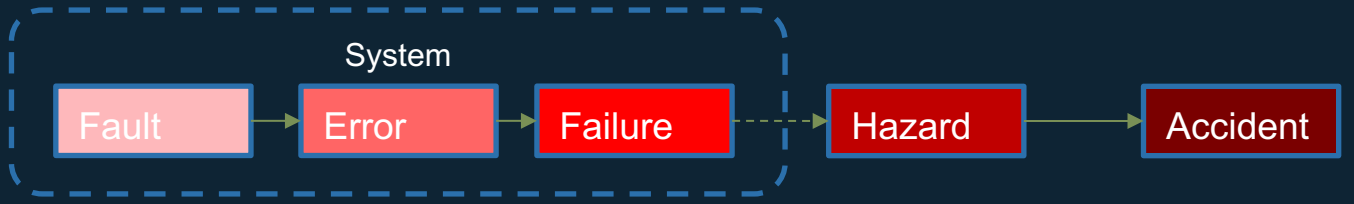
Challenges

- Safety Of The Intended Functionality (SOTIF) must be considered (ISO 21448) in context of AI/ML-based systems
- Show that all identified system hazards have been mitigated sufficiently
- Safety analyses techniques to create cause-effect-relationships for safety and SOTIF aspects

Background: Functional Safety & SOTIF

Functional Safety

- Absence of unacceptable risks (IEC 61508)
- Risk = combination of hazard probability and severity of the resulting accident
- Focus: Random hardware faults & systematic software faults

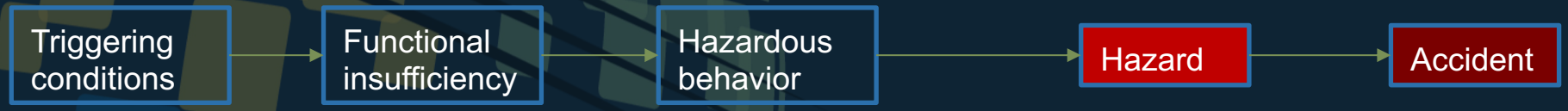


Fault-Error-Failure Chain according to

Avizienis, A., Laprie, J. C., et al. "Basic concepts and taxonomy of dependable and secure computing", 2004

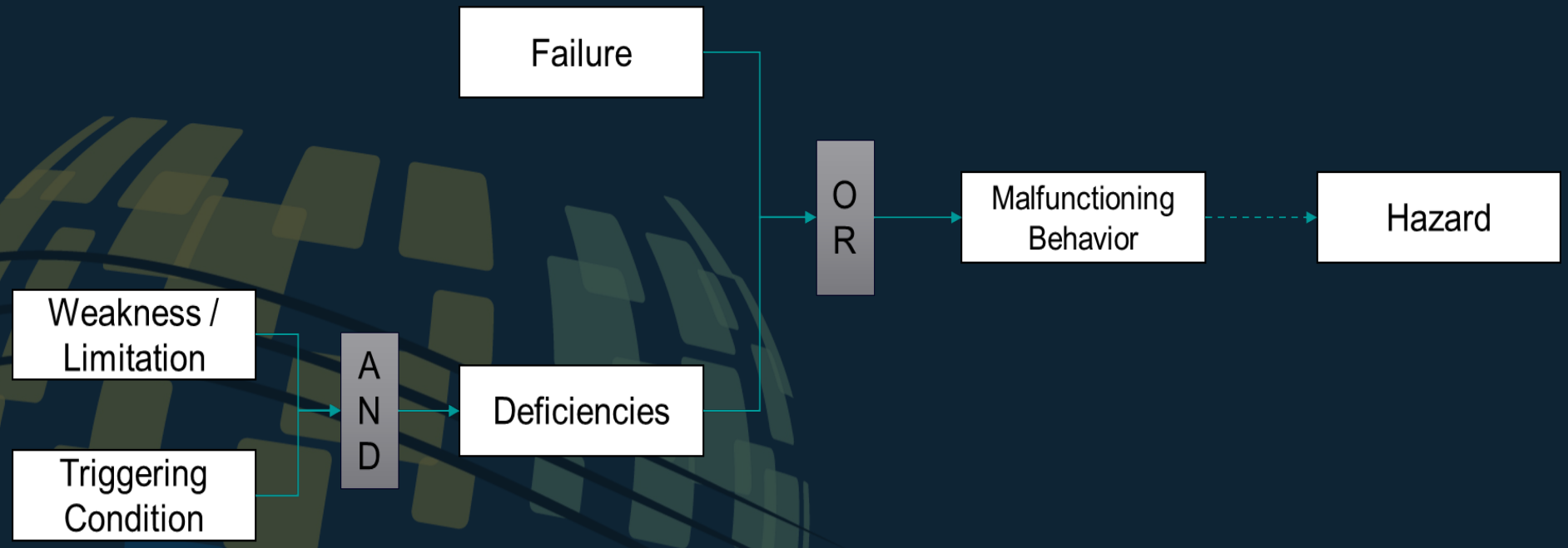
Safety Of The Intended Functionality (SOTIF)

- Absence of unreasonable risk due to hazards resulting from functional insufficiencies of the intended functionality or its implementation (ISO 21448)
- SOTIF activities include the identification of functional insufficiencies and the evaluation of their effects



Functional Safety and SOTIF: A combined view

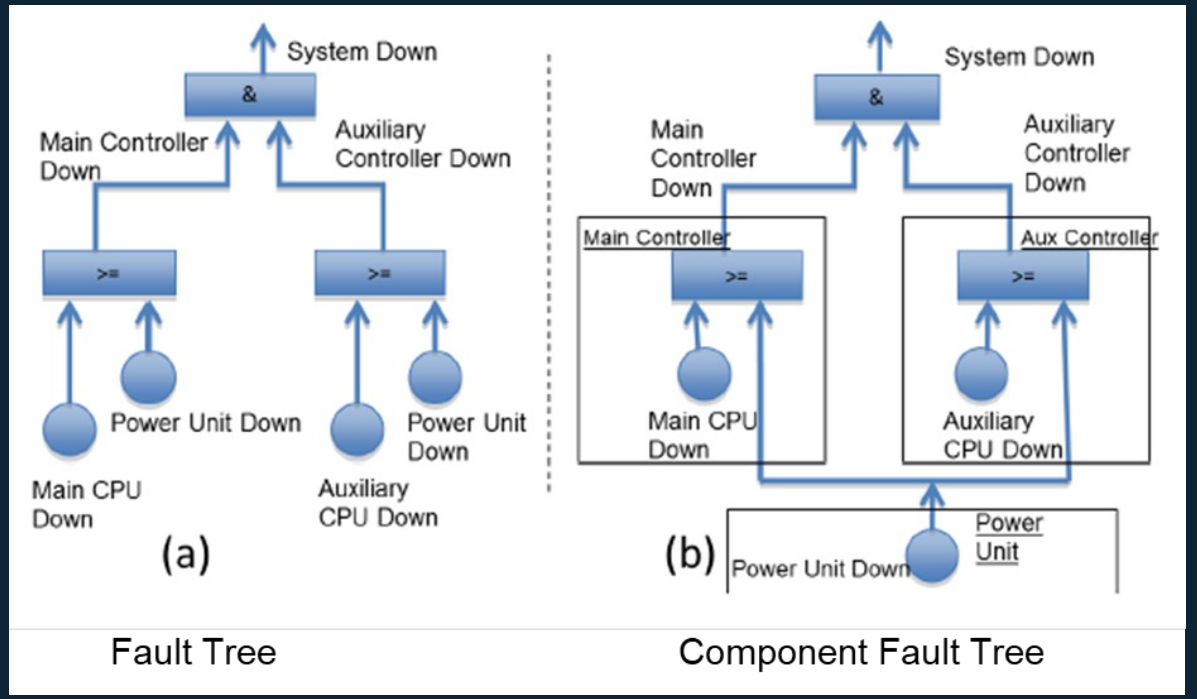
- *Deficiency* = combination of weakness/limitations (insufficiencies) and triggering conditions
- Deficiencies can lead to malfunction or malfunctioning behavior



Component Fault Trees (CFTs)*

- Extend classic fault trees with a component concept
- Focus on failure modes of an encapsulated system component
- Failures visible at the inport/outport of a component are modeled using Input/Output Failure Modes
- Modular, hierarchical composition of system fault trees

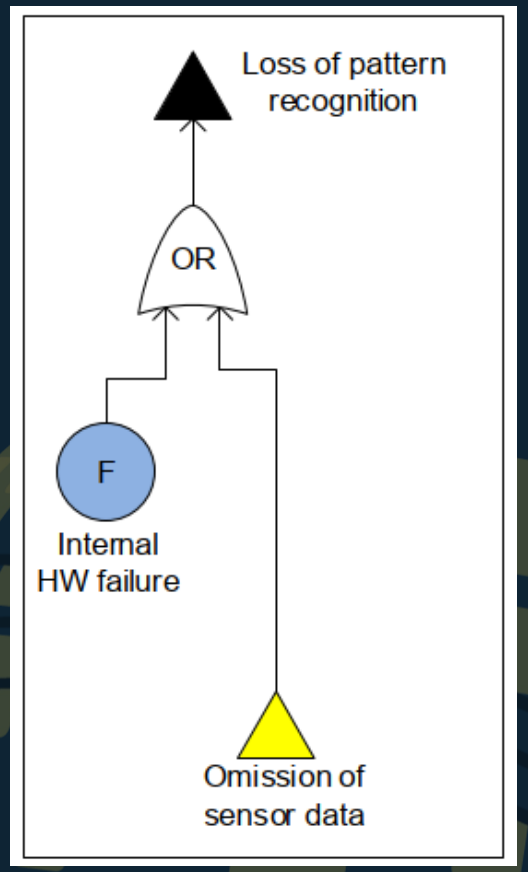
- Same information as Fault Tree (a), only different modeling concept (b)
- Divide-and-conquer strategy
- Systematic reuse of component CFTs
- Quantitative & qualitative Fault Tree

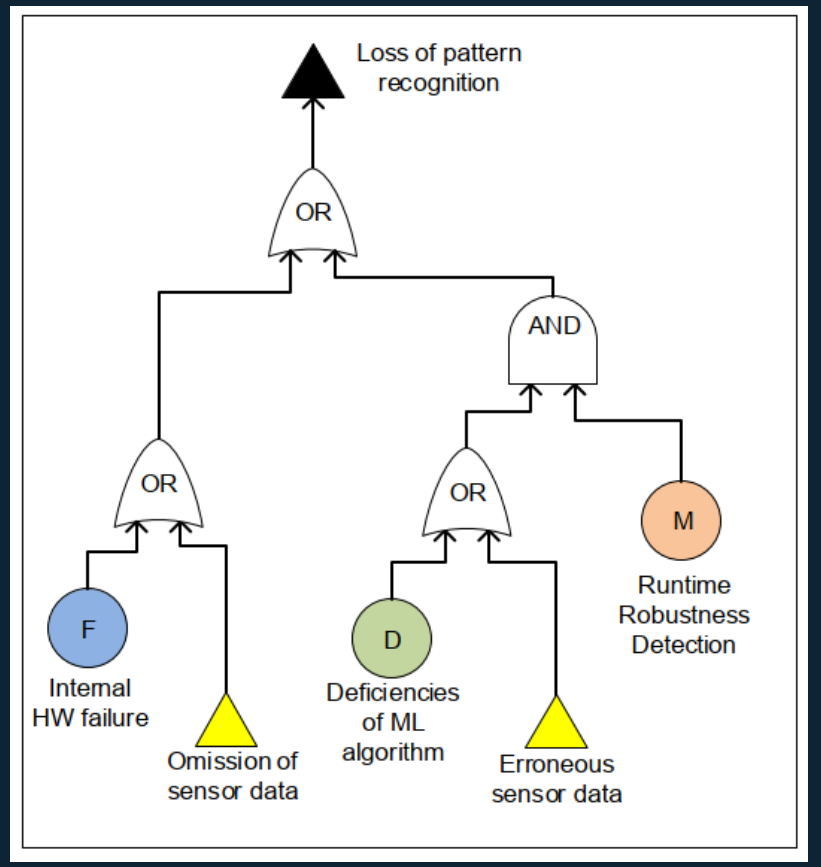


*) Kaiser, B., Liggesmeyer, P., Mäckel, O. (2003). A new component concept for fault trees, SCS '03: Proceedings of the 8th Australian workshop on Safety critical systems and software

Kaiser, B., Schneider, D., Adler, R., Domis, D., Möhrle, F., Berres, A., Zeller, M., Höfig, K., Rothfelder, M. (2018). Advances in Component Fault Trees, Proceedings of the 28th European Safety and Reliability Conference (ESREL)

From CFT to Component Fault and Deficiency Tree (CFDT)

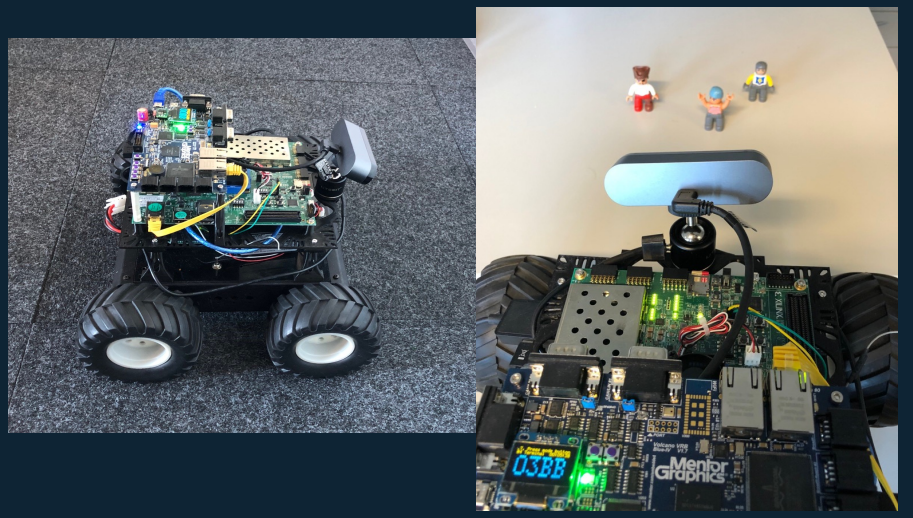




CFT = (IFM,OFM,B,G,SubCFT,C)

CFDT = (IFM,OFM,B,D,M,G,SubCFDT,C')

PANORover Case Study

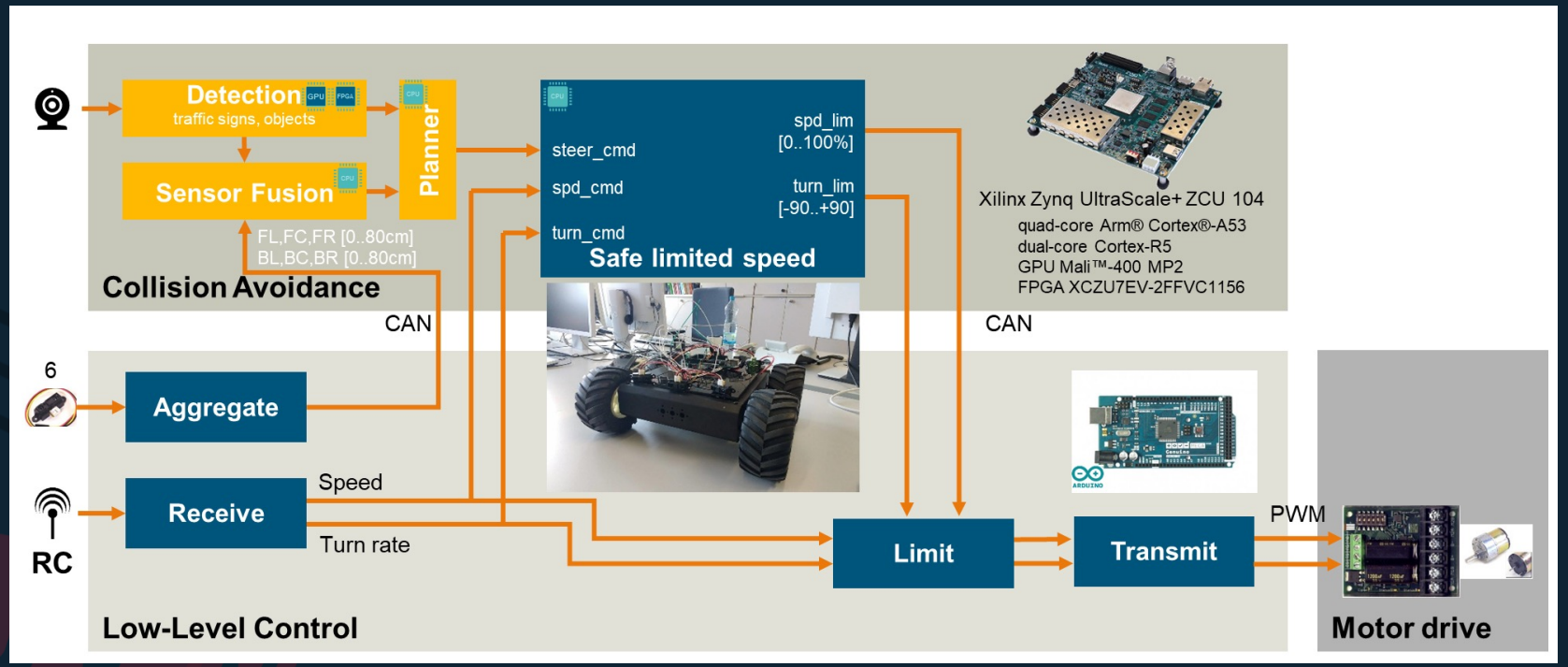


Use Case

- Rover with automated braking and collision avoidance function

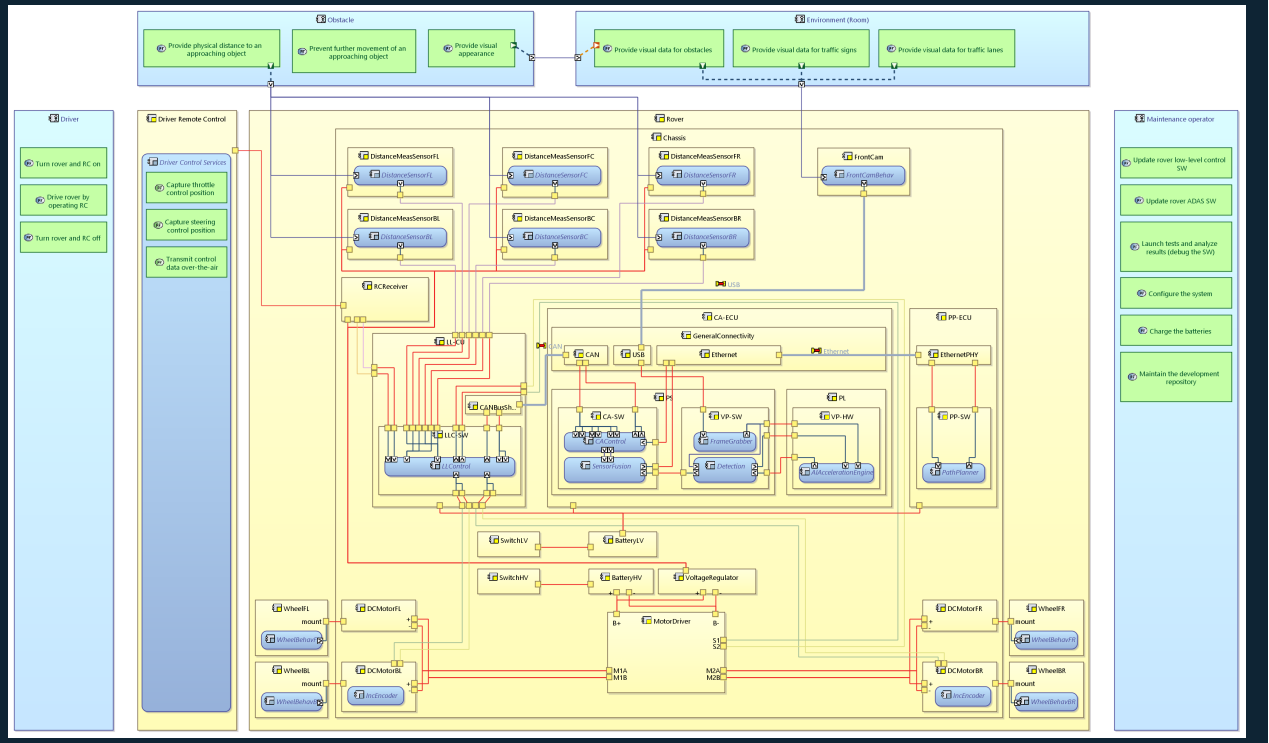
Function

- Avoid collisions by measuring distance via ultrasonic sensors
- Deep Neural Network (DNN) YOLO to detect pedestrians via camera
- Sensor fusion

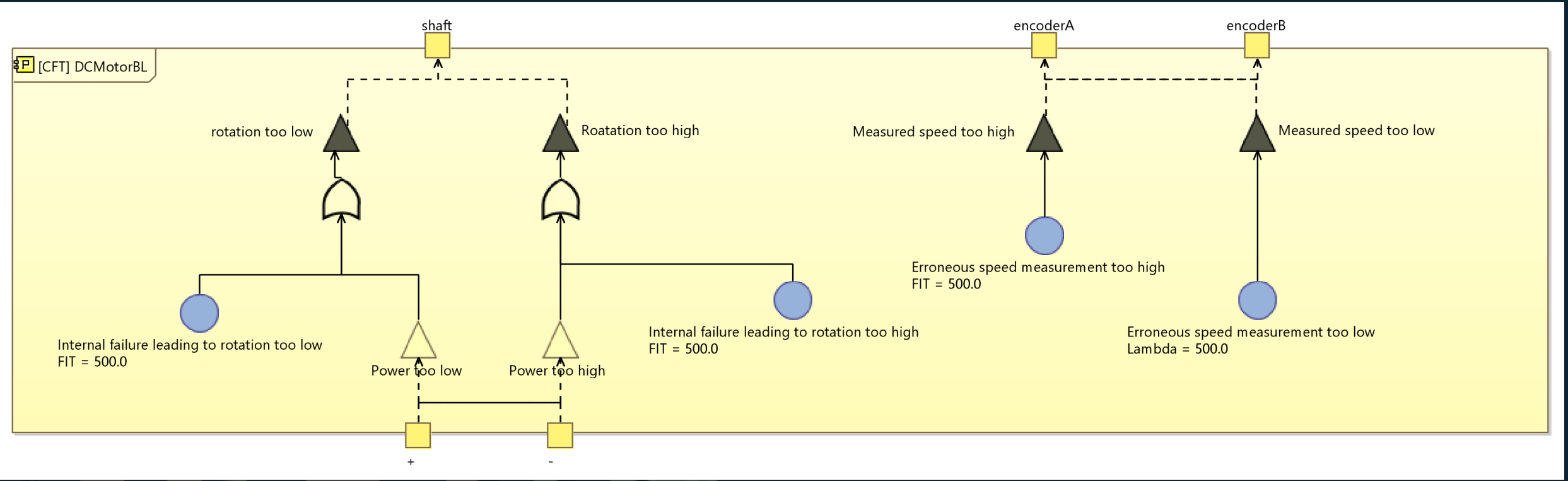


Safety Assurance using CFDTs

- Hazard Identification
  - Collision
  - Unintended Stop
- Physical system architecture defined in Capella using the ARCADIA methodology
- Qualitative Safety Analysis performed using CFDTs
  - To show that all hazards are mitigated sufficiently
  - Asses the combinations of deficiencies/failures (cut sets) leading to hazards
- Capella model of PANORover is enriched with a new view to create CFDT models
  - CFDT element for each component in physical architecture
  - From actuator to sensor

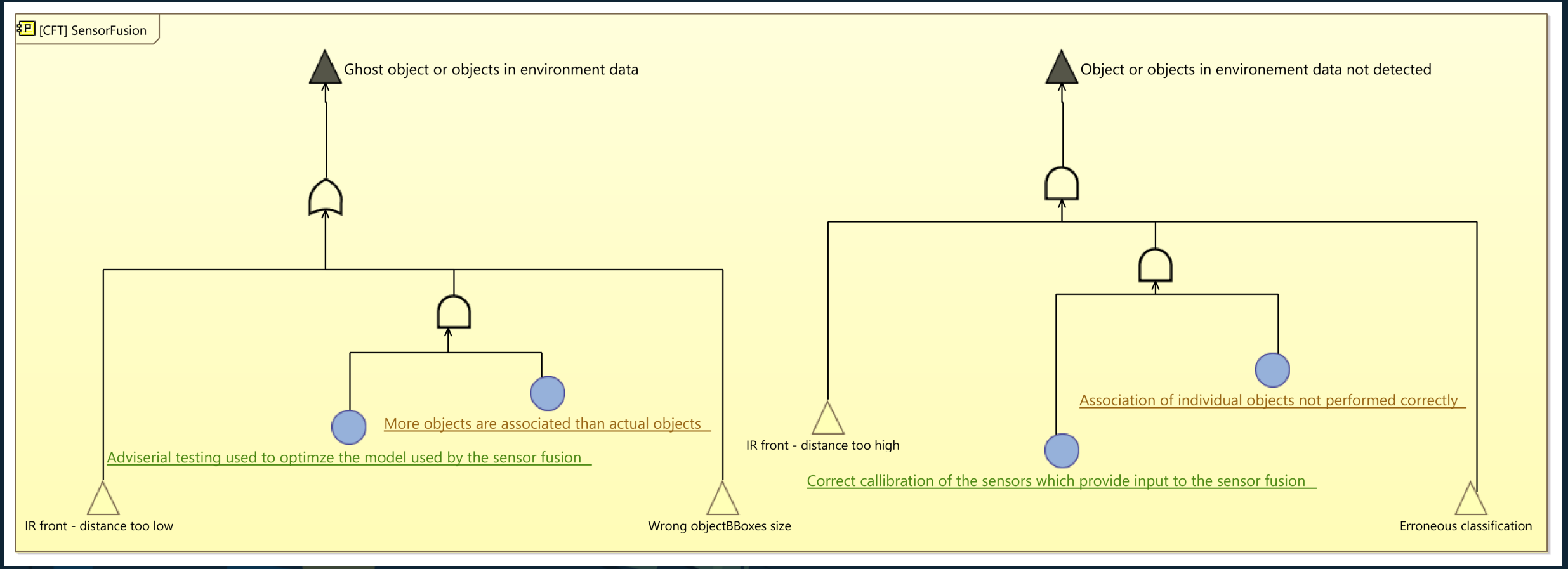


Safety Assurance using CFDTs



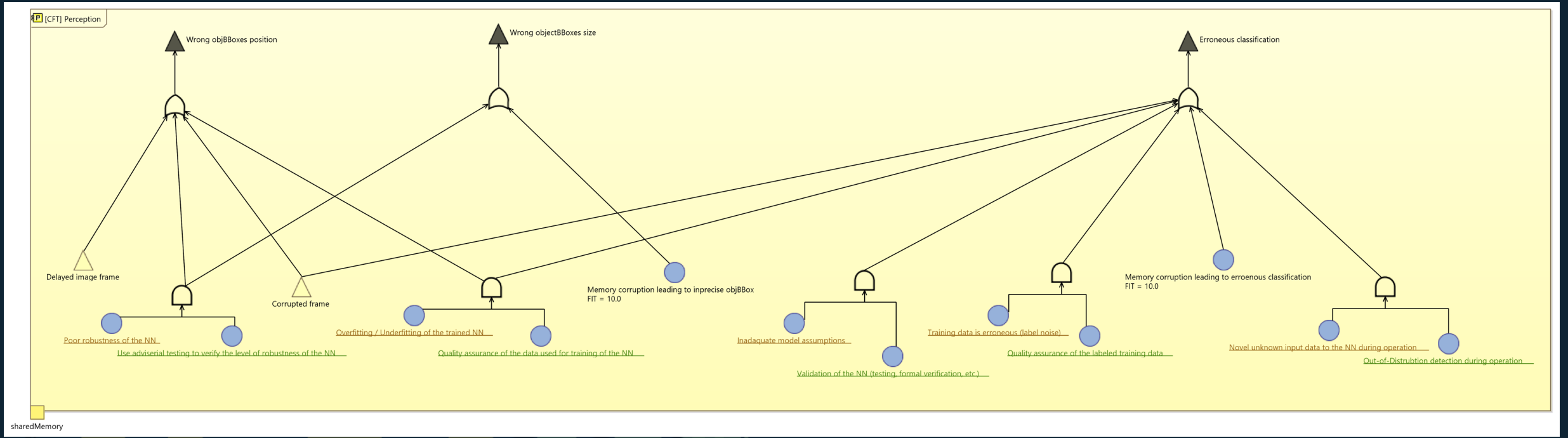
Example of a CFDT element of an actuator in the PANRover case study

Safety Assurance using CFDTs



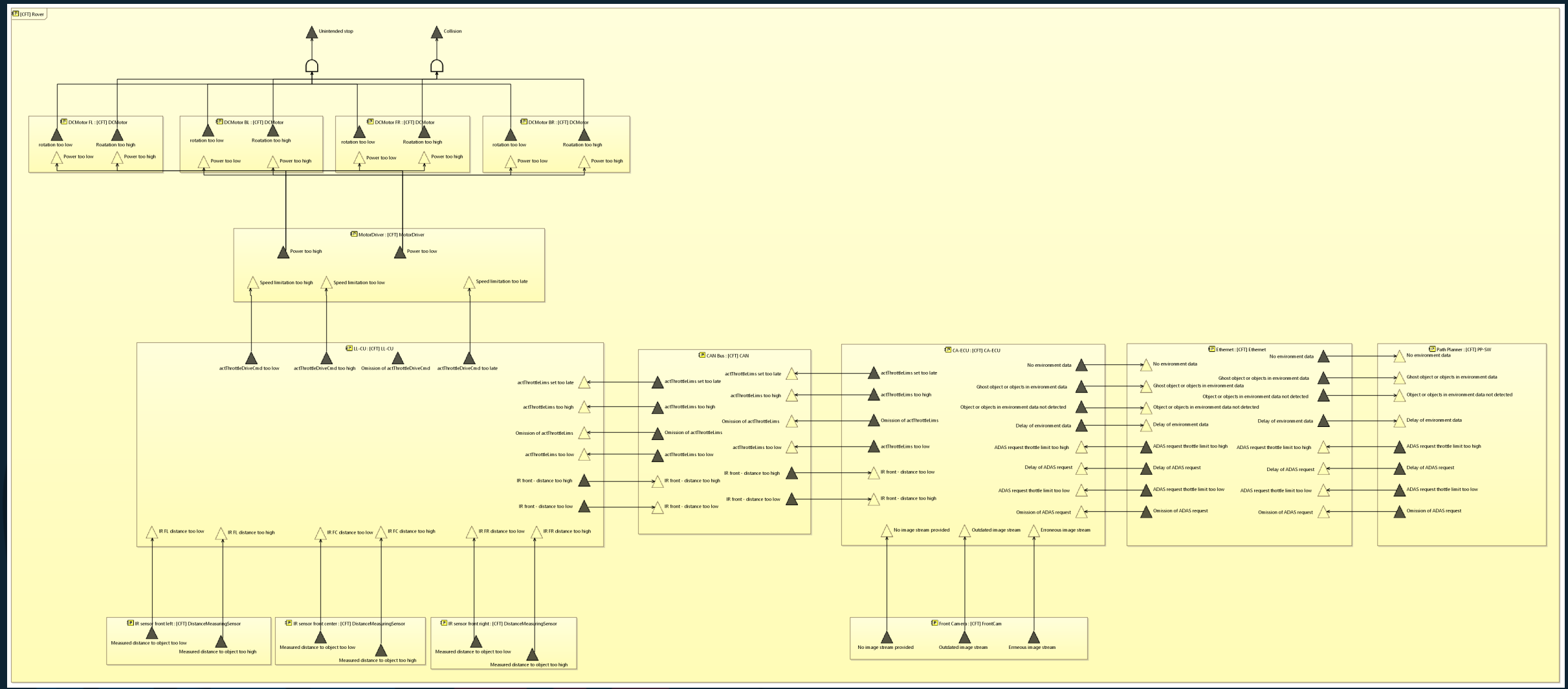
CFDT of the sensor fusion component in the PANRover case study

Safety Assurance using CFDTs

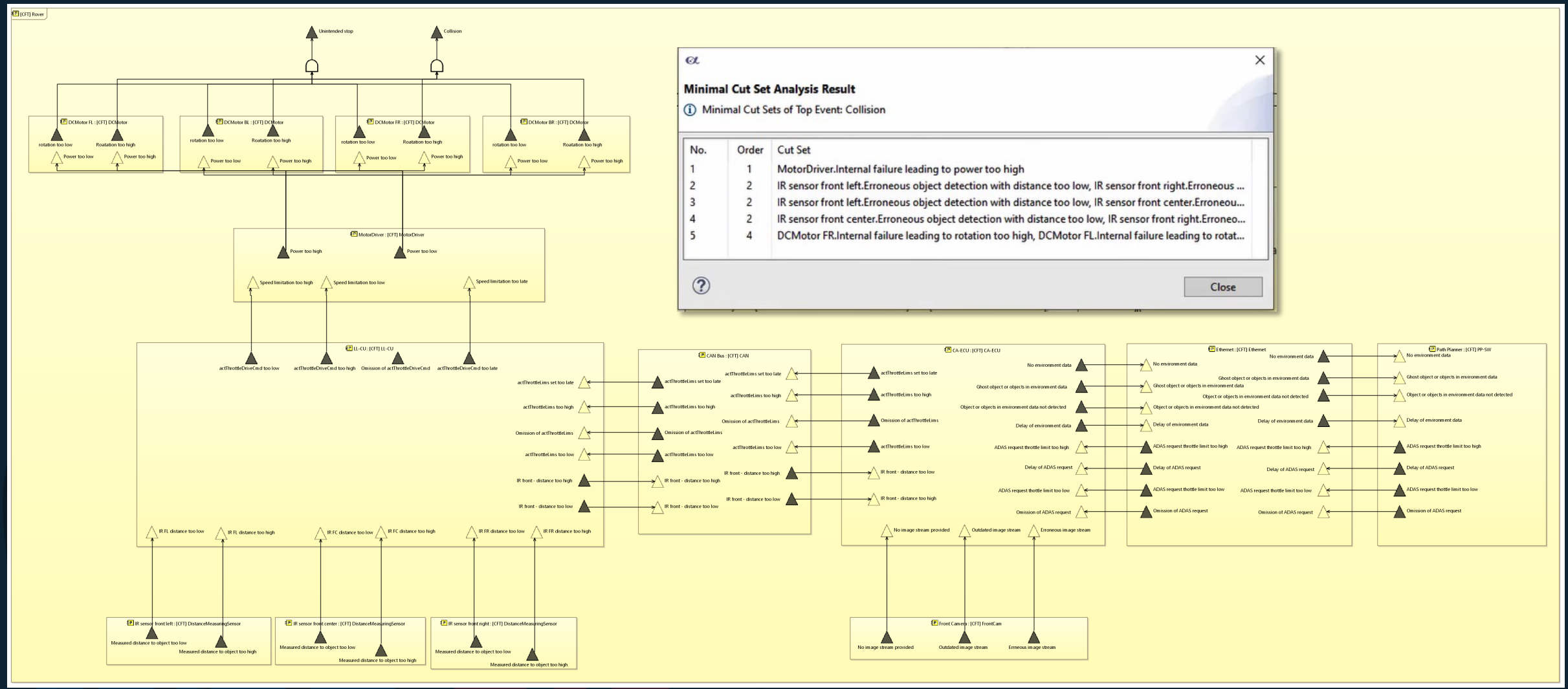


CFDT of the perception component in the PANRover case study

Safety Assurance using CFDTs



Safety Assurance using CFDTs



Lessons Learnt from the Case Study

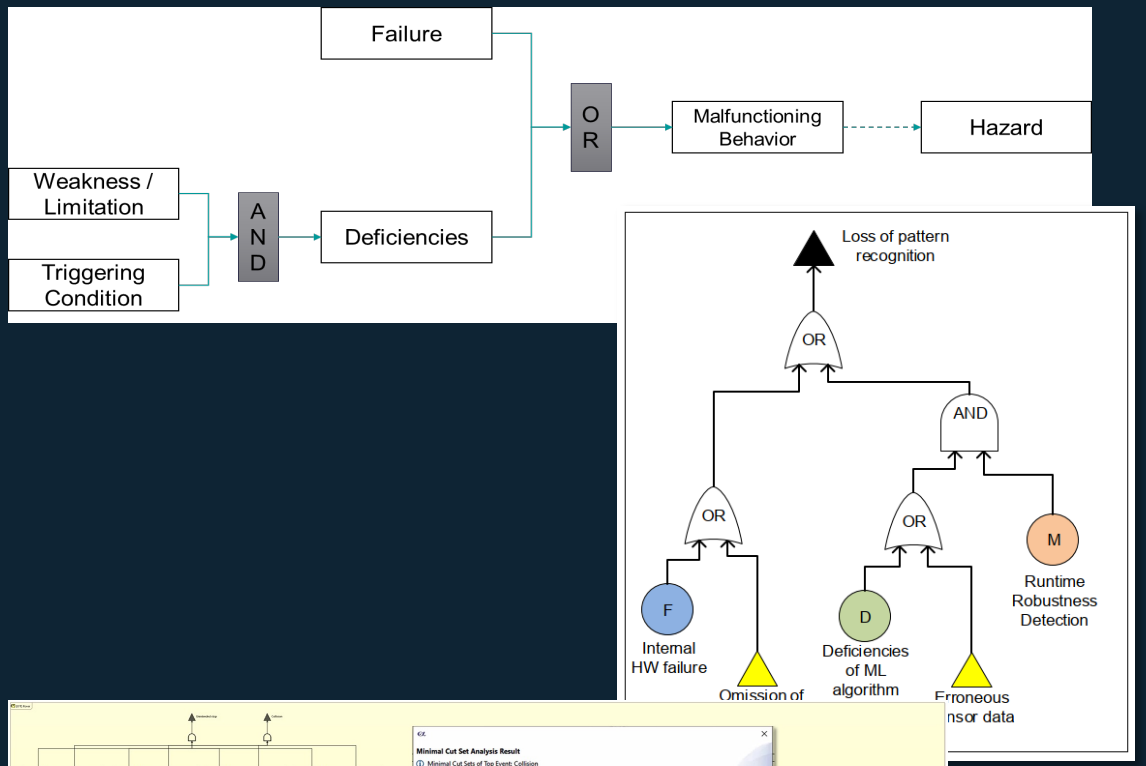
- CFDT methodology allows to conduct analysis to show that all risks have been mitigated
- Combined model for functional safety and SOTIF
- Integration into the MBSE model
  - Eases communication between system engineers and safety experts
  - Ensures consistency of the safety analysis model
- CFDTs provide abstraction for SOTIF aspects by aggregating functional insufficiencies/weaknesses and triggering condition
- Modular approach of the CFT/CFDT methodology
  - Allows to mix different levels of abstraction for different components
  - Enables reuse of parts of the safety analysis model in different projects
- Divide-and-conquer approach of the CFT/CFDT methodology
  - Helps experts to systematically identify the causes of hazards
  - Supports effort to minimize the number of unknown, potentially unsafe failures
- CFDT elements can be created by different experts and integrated seamlessly into the overall CFDT model
- Safety analysis model can be developed in an iterative or agile way
- Using qualitative fault tree analysis (minimal cut set analysis) shows where safety mitigation approaches are missing and provides the basis for further reviews

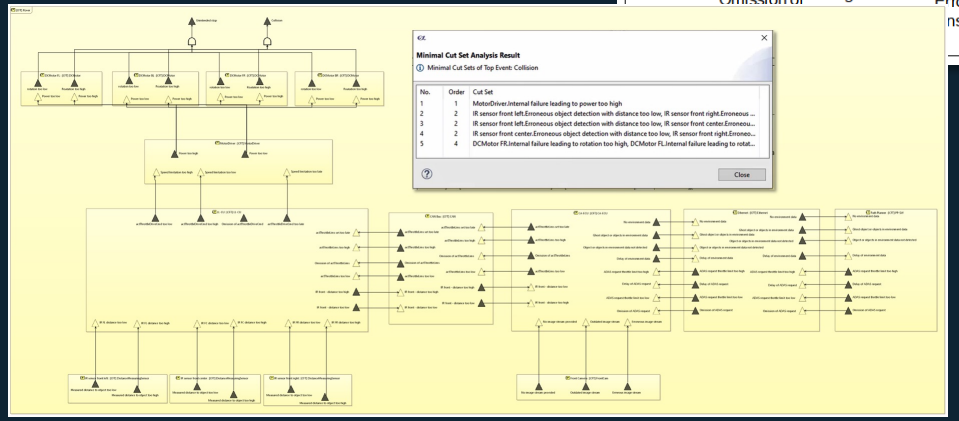
Open Points

- Is the abstraction of functional insufficiencies/weaknesses and triggering condition using so-called deficiencies always appropriate?
  - Model functional insufficiencies/weaknesses and triggering condition explicitly as separated elements in CFDT is more suitable in some cases
- The hazard identification and risk assessment process is very challenging for autonomous systems operating in open world context
  - Different operational scenarios may lead to different hazards and also different failures and deficiencies which can cause those hazards
  - In CFDT model multiple top events and their causes can be specified (however this increases the complexity)
- Quantitative analysis of CFDTs
  - No approach for quantification of deficiencies available yet
  - Uncertainties must be considered when annotating deficiencies with probabilities
  - Techniques such Bayesian Networks might be used
  - Quantification of deficiencies and extension of CFDT methodology must be investigated in detail in the future

Summary & Outlook

- Safe systems incorporating AI/ML are required for many industrial use cases
- Failures of the system & functional insufficiencies of the intended functionality must be analyzed w.r.t.
  - their effects
  - mitigation strategies
- CFDTs = methodology to describe cause-effect-relationships for failures & functional insufficiencies and system hazards
  - Modular, compositional
  - Seamless integration into MBSE model
  - Covering functional safety and SOTIF aspects
- Evaluation of PANORover use case showed feasibility of the methodology
- Next Steps: Extension of CFDT w.r.t. quantitative analysis of using e.g. Bayesian Networks





Thank You! Questions?

33rd Annual INCOSE international symposium

hybrid event

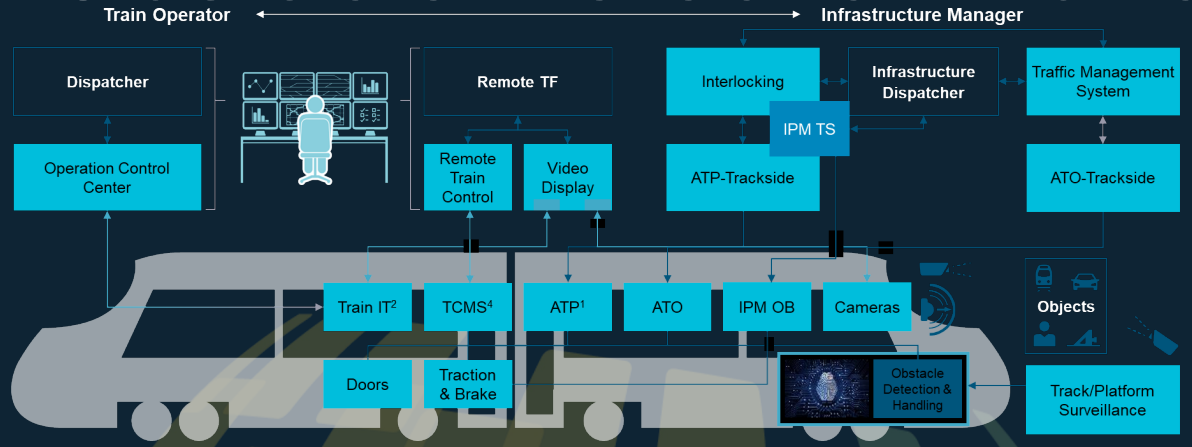
Honolulu HI USA

www.incose.org/symp2023

for Autonomous Trains Using CFDTs to Assess a ML-based Obstacle Detection Function

safe·trAIn

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag



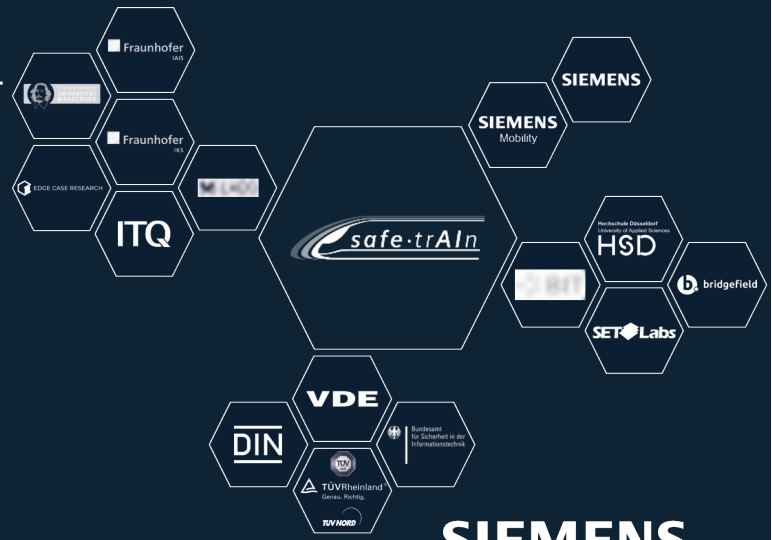
- The digitalization of train operations aims at **fully automated rail mobility** (GoA3/4)
- **Obstacle Detection** is the **most critical** new onboard system to reach GoA4 operation
- **Safe and robust AI-based methods** and **novel strategies for certification** are needed to implement ML-based Obstacle Detection in autonomous trains

Project Objectives :

- Safe AI-based functions for a driverless regional train
- Method to enable the assurance of AI-based functions in terms of safety
- Guidelines and concepts to enable certification of AI-based functions in railway domain
- Input for national and European standardization activities

Key Data :

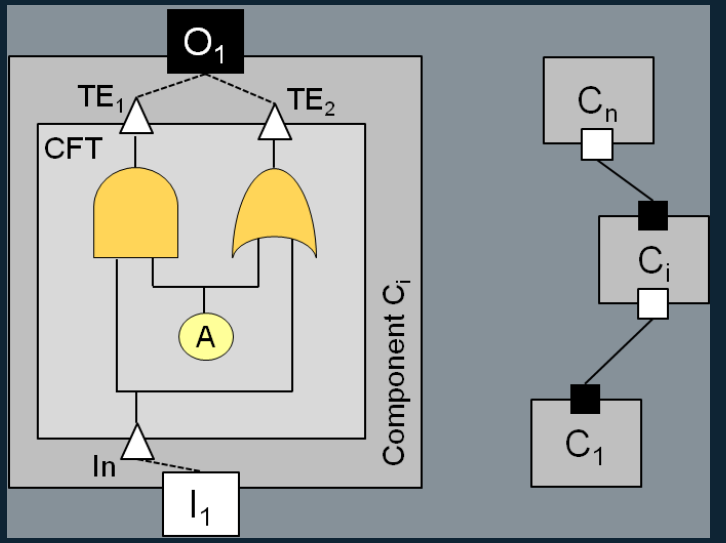
- Duration 01/2022-12/2024
- Overall budget: €23 m
- 16 partners

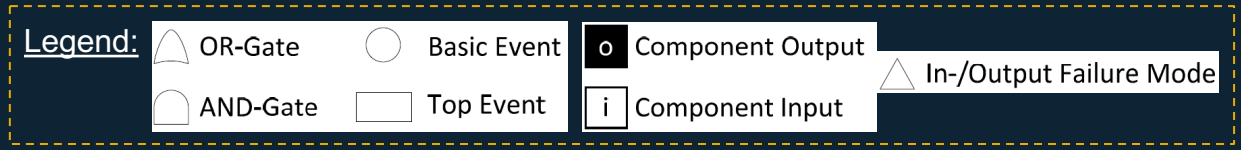


SIEMENS

Component Fault Trees (CFTs)*

- Extend classic fault trees with a component concept
- Focus on failure modes of an encapsulated system component
- Failures visible at the inport/outport of a component are modeled using Input/Output Failure Modes
- Modular, hierarchical composition of system fault trees





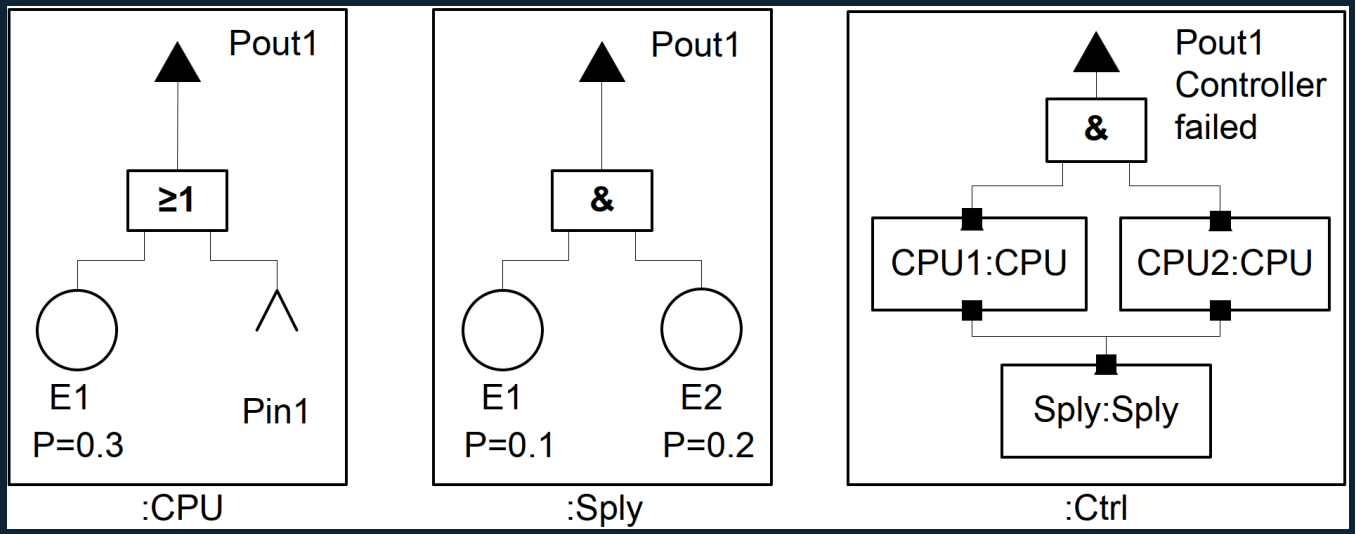
*) Kaiser, B., Liggesmeyer, P., Mäckel, O. (2003). A new component concept for fault trees, SCS '03: Proceedings of the 8th Australian workshop on Safety critical systems and software

Kaiser, B., Schneider, D., Adler, R., Domis, D., Möhrle, F., Berres, A., Zeller, M., Höfig, K., Rothfelder, M. (2018). Advances in Component Fault Trees, Proceedings of the 28th European Safety and Reliability Conference (ESREL)

Component Fault Trees (CFTs)*

- Extend classic fault trees with a component concept
- Focus on failure modes of an encapsulated system component
- Failures visible at the inport/outport of a component are modeled using Input/Output Failure Modes
- Modular, hierarchical composition of system fault trees

- Same information as Fault Tree (a), only different modeling concept (b)
- Divide-and-conquer strategy
- Systematic reuse of component CFTs
- Quantitative & qualitative Fault Tree



*) Kaiser, B., Liggesmeyer, P., Mäckel, O. (2003). A new component concept for fault trees, SCS '03: Proceedings of the 8th Australian workshop on Safety critical systems and software

Kaiser, B., Schneider, D., Adler, R., Domis, D., Möhrle, F., Berres, A., Zeller, M., Höfig, K., Rothfelder, M. (2018). Advances in Component Fault Trees, Proceedings of the 28th European Safety and Reliability Conference (ESREL)