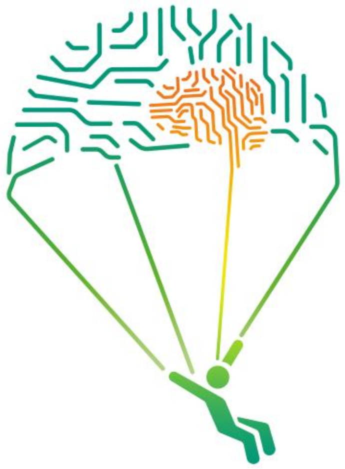
Safety-critical systems with Machine Learning component Challenges and Solutions

Daniel Schneider

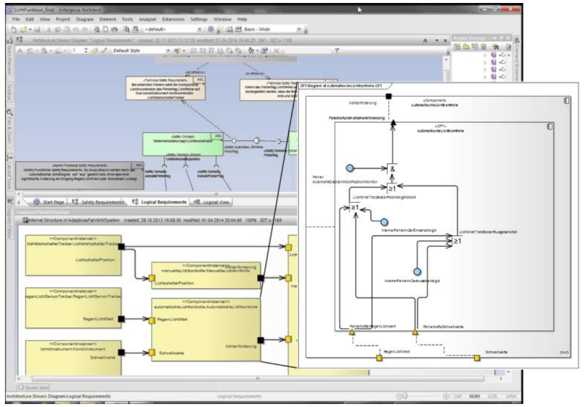
IFIP WG 10.4 Workshop on January 20th 2022

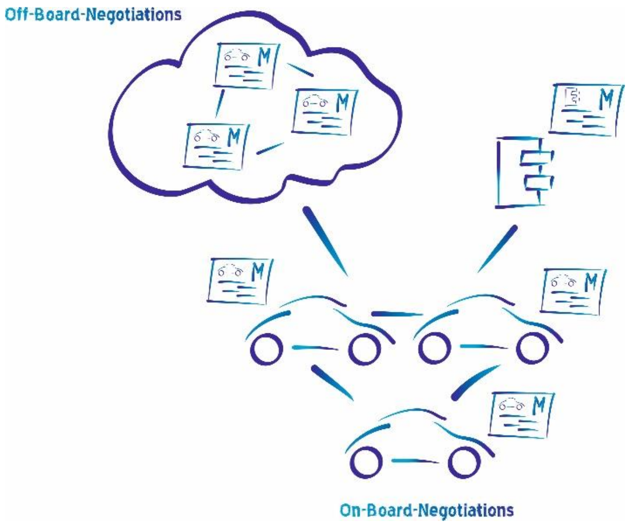


Fraunhofer
IESE

FRAUNHOFER IESE - SAFETY ENGINEERING DEPARTMENT

- Engineering of safety-related Solutions
  - Consulting, Tooling & Doing
- Model-based Safety Engineering
  - Hazard- and Riskanalyses
  - Safetyanalyses (FMEA, FTA, CFT etc.)
  - Safety Concepts and Safety Cases
  - Tools and methods (in particular www.safeTbox.de; https://youtu.be/VE_BiN-S7jw)
- Research Topics
  - Safety of collaborative autonomous systems
  - Dynamic Risk Management
  - Dependable AI
  - Security for Safety





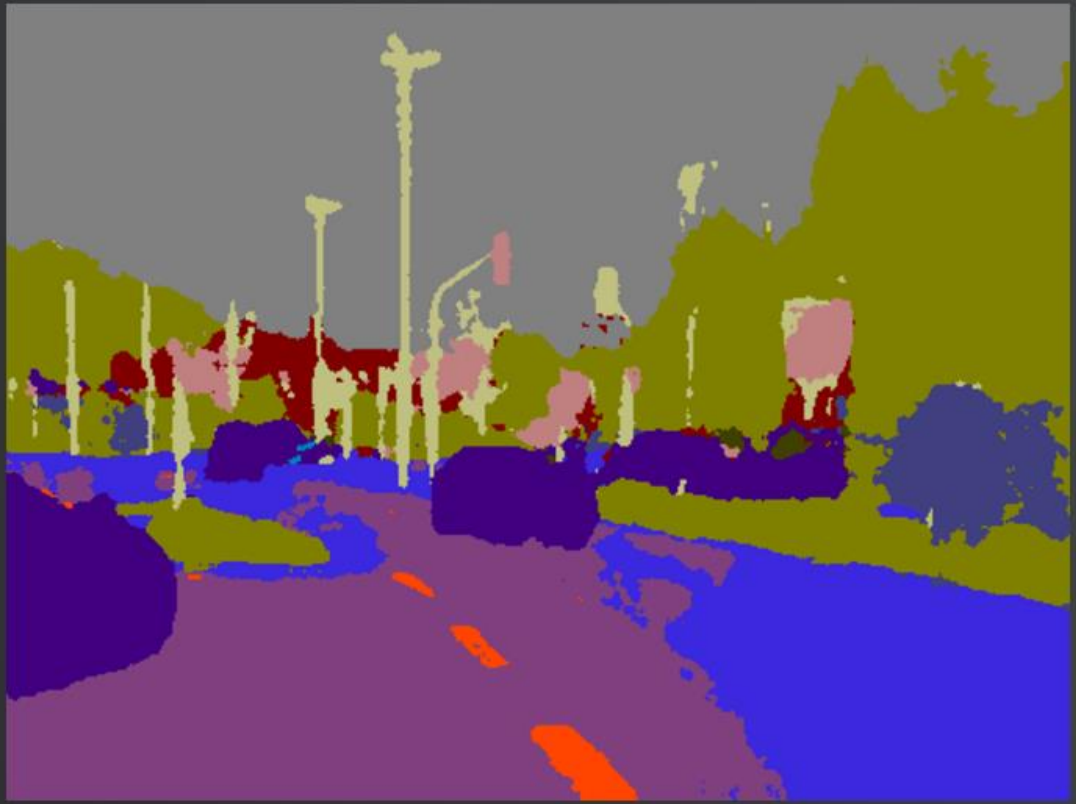
Fraunhofer IESE

CHALLENGE

	Inter-connectedness	Autonomy	Use of AI	Uncertainties and Unknowns
	Heterogenous collectives of CPS - global scale	Autopoietic Autonomy	prevalent dynamic learning (critical)	Unknown
	Heterogenous collectives of CPS		sporadic dynamic learning (uncritical)	
	Homogeneous collectives of CPS	Cooperative Autonomy		Partially known
	Groups of CPS	Mission Autonomy	Prevalent, pre-learned (critical)	
	Complex CPS	Multiple Functions	Sporadic, pre-learned (uncritical)	Known
	Closed standalone ES	Single functions		

FROM DIGITAL TO „INTELLIGENT“

- For specific tasks, Neural Networks show better performance than classic software (and even than humans)
- Example: Semantic segmentation of images



Fraunhofer IESE

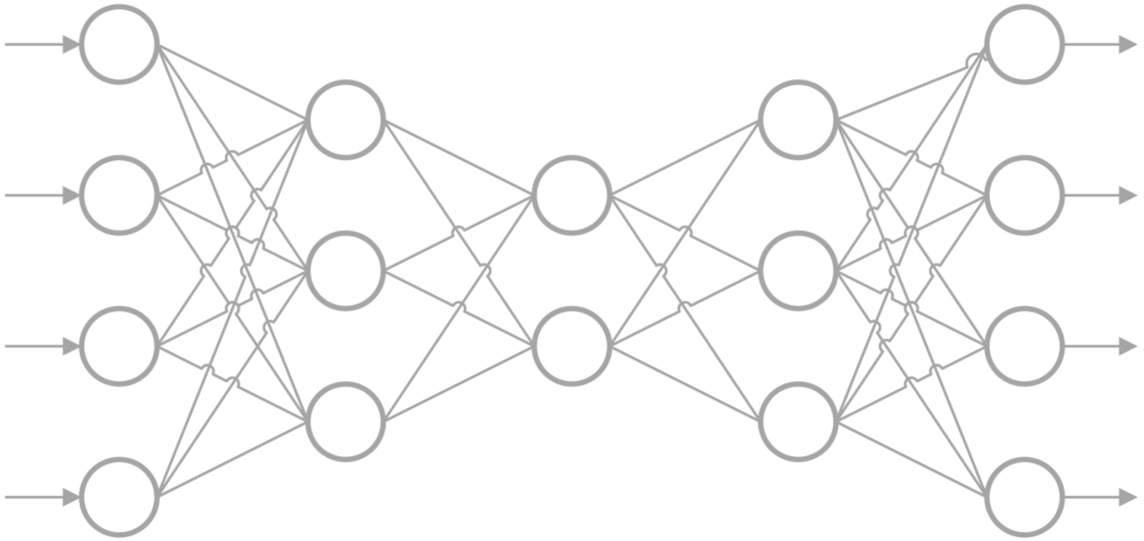
NEURAL NETWORK ENGINEERING (?)

```python
    def make_observation(self):
        raw_obs = self.measurements

        # Calculate distance of closest vehicle
        dist_min = 999999999
        Playerposition = np.array([
            raw_obs.player_measurements.transform.location.x,
            raw_obs.player_measurements.transform.location.y
        ])

        for agent in raw_obs.non_player_agents:
            if agent.HasField('vehicle'):
                x = np.array([
                    agent.vehicle.transform.location.x,
                    agent.vehicle.transform.location.y
                ])
                dist = int(np.linalg.norm(x - Playerposition))
                dist_min = np.min(np.array([dist_min, dist]))

```

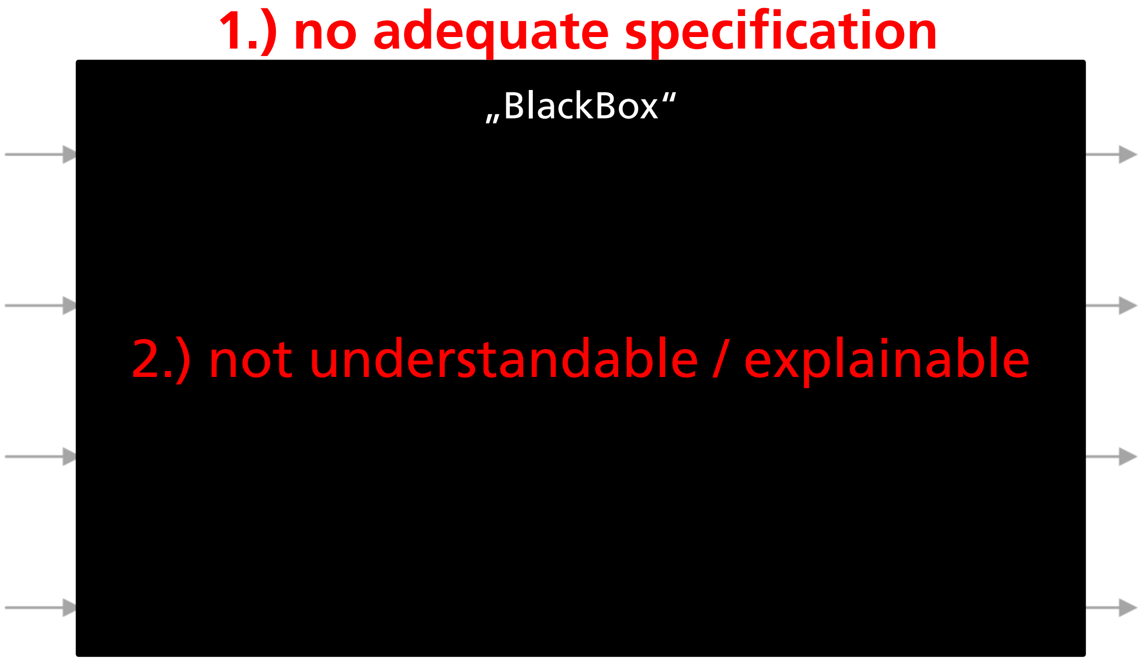


- A (very) different way of engineering software
- Neural networks are very different from source code
- Established methods, techniques and tools are not directly applicable
  - There are specific techniques and tools, but
  - there is a need for a more systematic engineering of neural networks

Fraunhofer IESE

SAFETY CHALLENGES

- Typically, a sound requirements specification is missing
  - There will be training data and maybe a partial requirements specification
  - This is not very surprising, because ML is particularly attractive to address problems where it is hard to come up with a sound specification (e.g. camera-based object classification)
  - This complicates V&V and the generation of sound evidence for a safety argument
- In addition, proper analysis and verification is difficult due to a lack of explainability
  - BlackBox: Established WhiteBox Techniques (such as Inspections, Walkthroughs) not applicable
  - Apparently insignificant changes at the inputs can lead to very significatn changes at the output
  - Physical Hacks a problem



STARTING POINTS FOR SAFETY ASSURANCE OF ML COMPONENTS

STARTING POINTS FOR ASSURING ML COMPONENTS
INTEGRATED SAFETY AND ML ENGINEERING

- Only use ML components when there is no acceptable conventional solution
  - Accordingly, keep the ML part of the system as small as possible
- Integrate/align the activities and work products of Safety and ML Engineering
- At least a partial and as-good-as-possible requirements spec shall be created. Benefits:
  - Traceability wrt. safety engineering; e.g. clear association with safety requirements broken down from a hazard and risk analysis
  - Inform training data engineering, tailoring and QA of training data
  - Argue completeness or coverage regarding important quality aspects
  - V&V of the trained ANN against the spec
  - The specification can be the basis for a safety supervisor or similar runtime measures

STARTING POINTS FOR ASSURING ML COMPONENTS
INTEGRATED SAFETY AND ML ENGINEERING

- Methods and techniques for analyzing and hardening (zB: XAI)
- In case of object classification and conv nets, you can improve the performance of the NN by using techniques such as heatmapping or GradCAM, but you cannot assure it will always work
- E.g. you cannot know if every classification will be correct
- In general, guidance is required wrt. adequacy of techniques and generated evidence

Fraunhofer
IESE

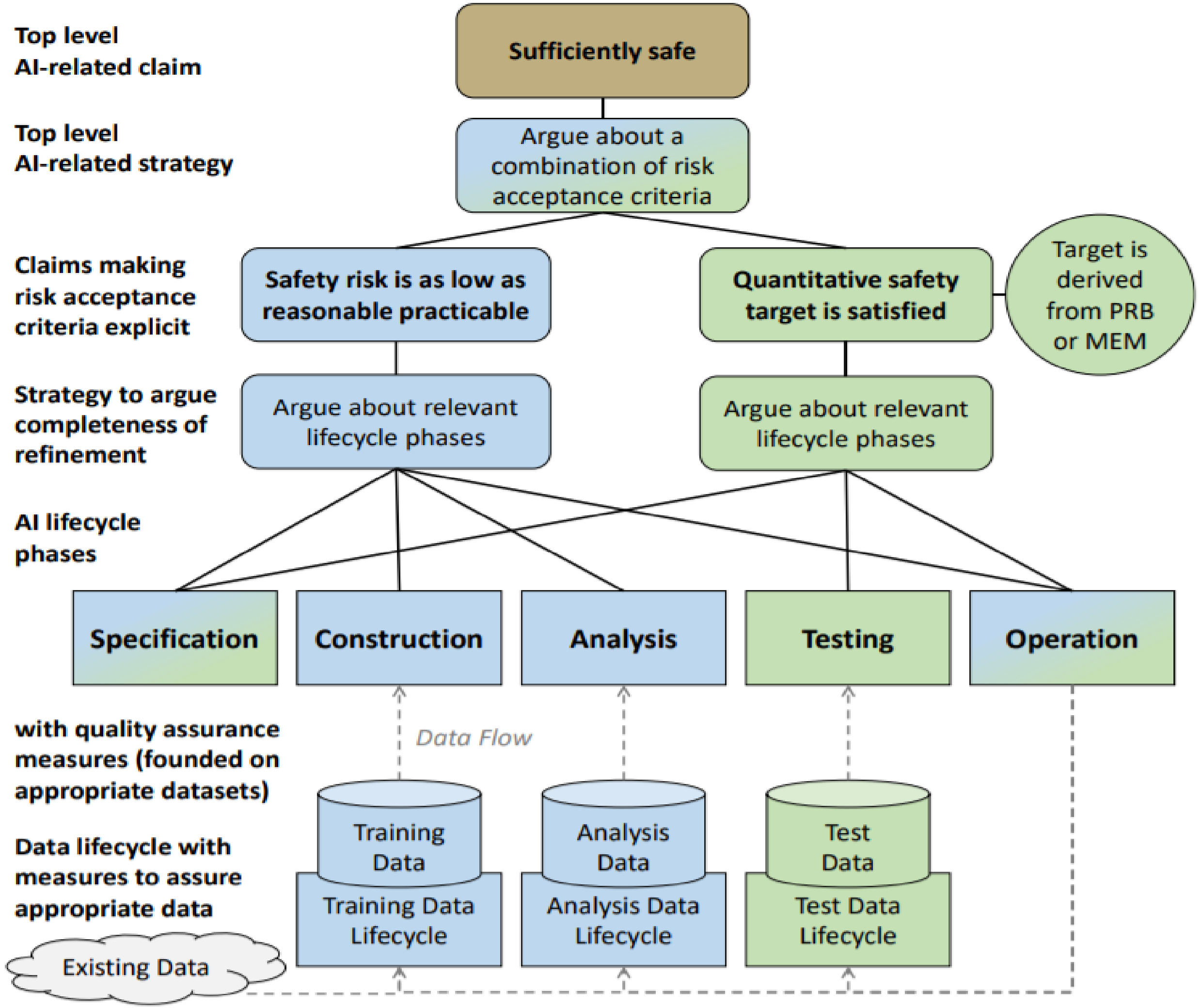
STARTING POINTS FOR ASSURING ML COMPONENTS
INTEGRATED SAFETY AND ML ENGINEERING

- Ongoing research wrt ML: Assuring robustness of the learned model, enable predictability and integrate explainability into the ML components
- (Redundancy-)Measures on an architectural level; e.g.:
  - Safety Supervisor / Simplex architecture
  - Homogenous and diverse redundancy (e.g. parallel utilization of ML components with different training data, architecture etc.)
  - Layered supervisor concept (layers of protection architecture)
- Validation as central element of assurance (i.e. for generating safety evidence)
  - Challenge lies in the selection of test cases and in arguing coverage and completeness
  - Currently a lot of research
    - E.g. PEGASUS and V&V Methoden projects in Germany

STARTING POINTS FOR ASSURING ML COMPONENTS

INTEGRATED SAFETY AND ML ENGINEERING

- Overall there shall be a seamless integration between ML Engineering and Software-, Systems- and Safety-Engineering
- We recommend setting up an explicit and adequately specified argumentation structure (e.g. in form of an assurance case) for the key properties of the system
- Argumentation patterns can be re-used

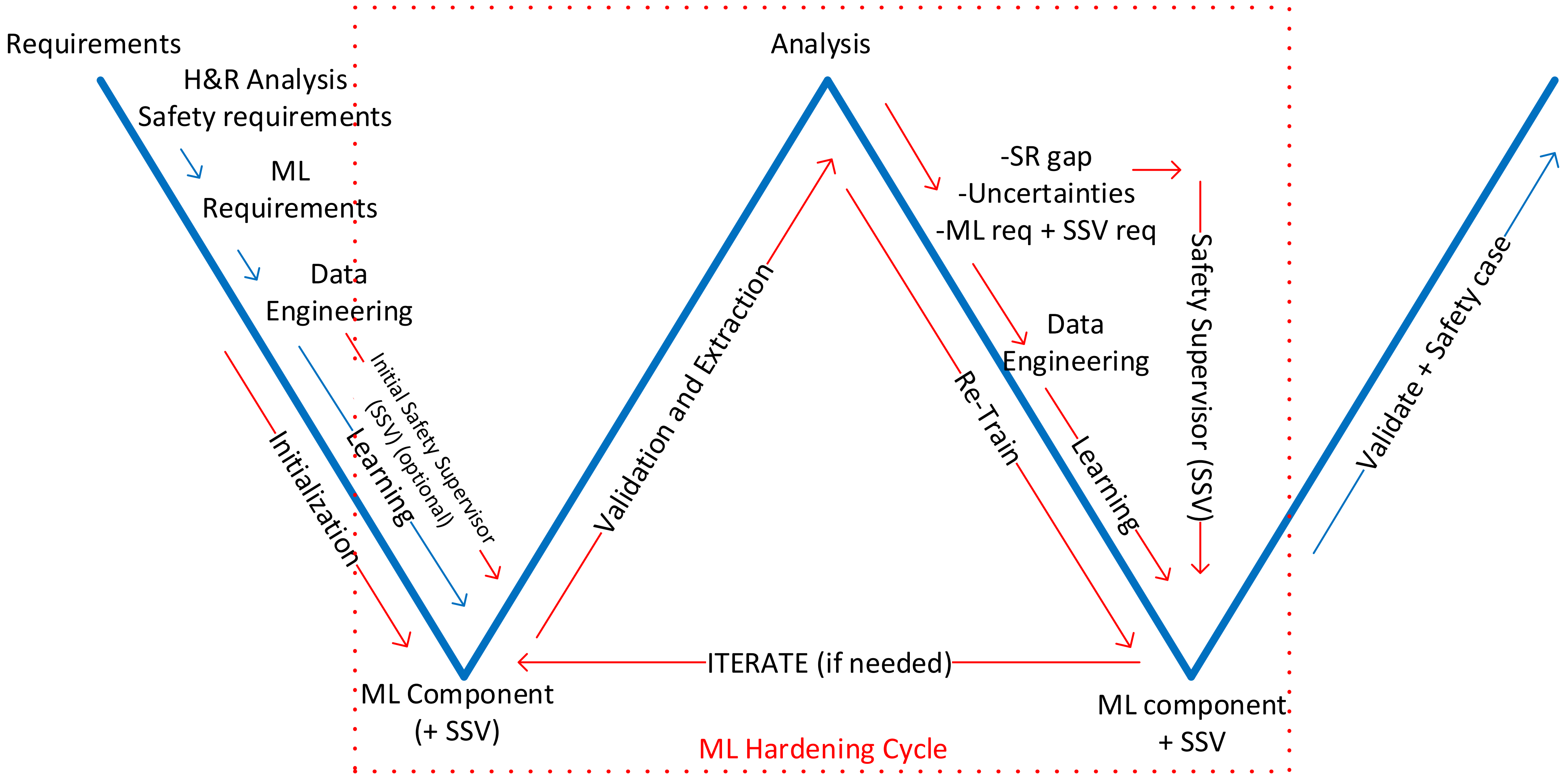


https://www.omg.org/spec/SACM/2.0/About-SACM/

Fraunhofer IESE

STARTING POINTS FOR ASSURING ML COMPONENTS

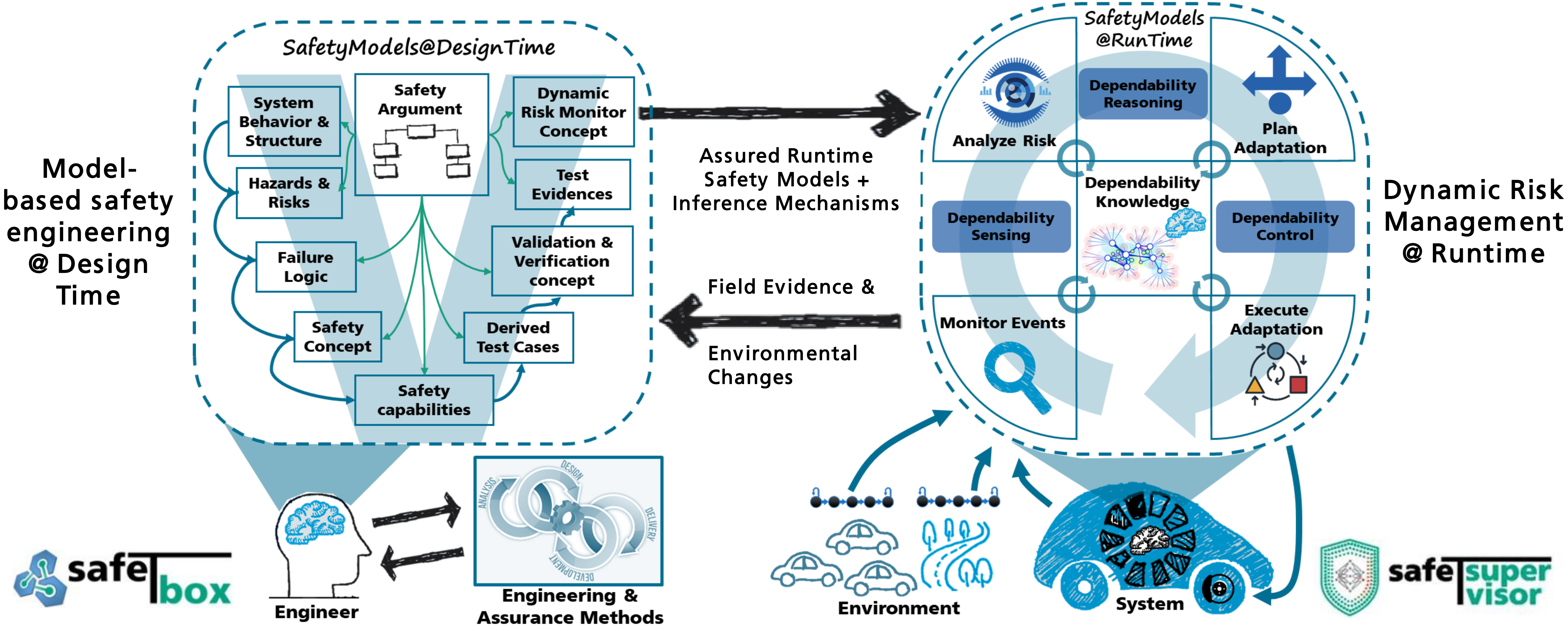
INTEGRATED SAFETY AND ML ENGINEERING – SUMMARY



Fraunhofer IESE

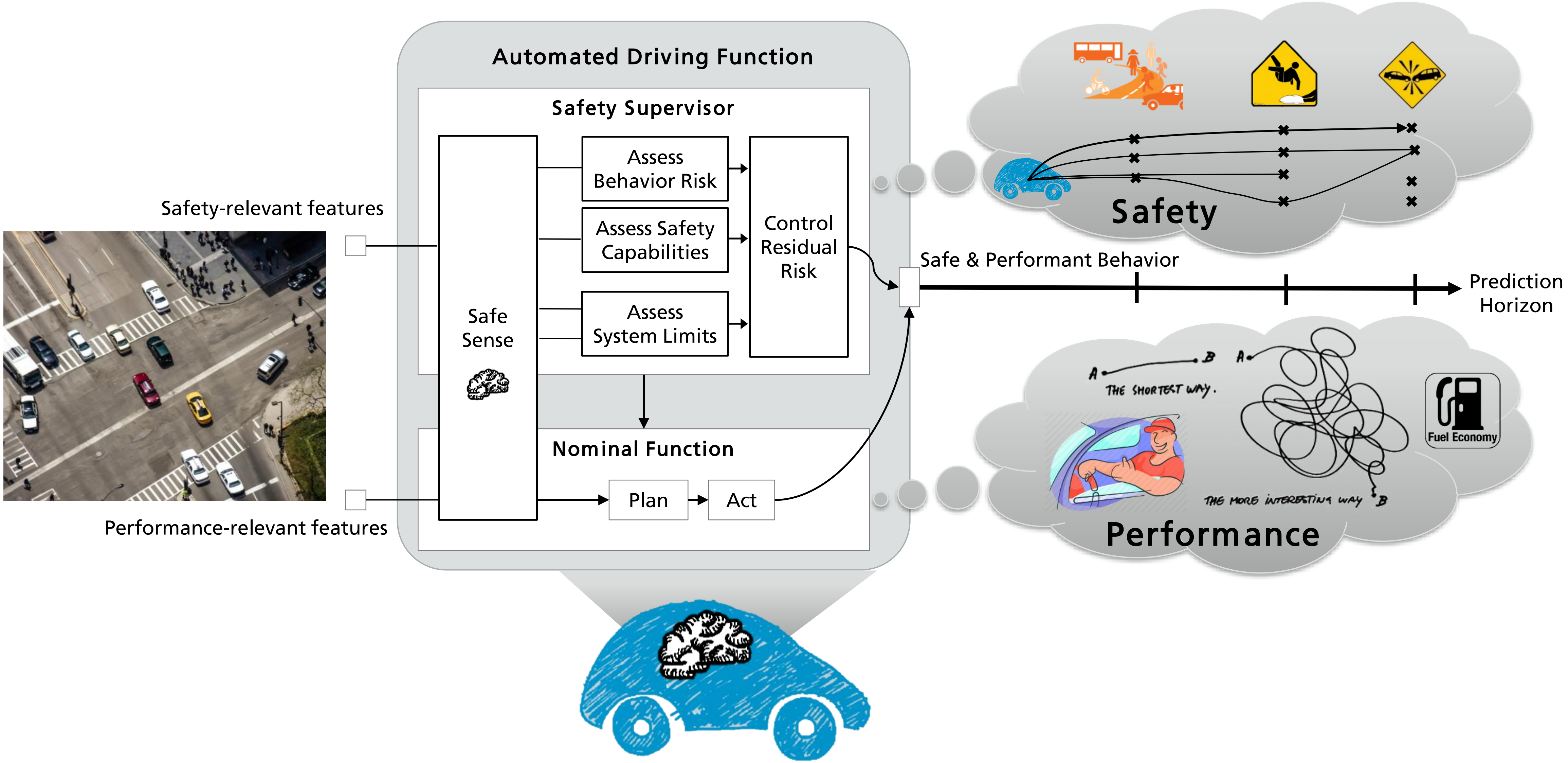
DYNAMIC RISK MANAGEMENT

DYNAMIC RISK MANAGEMENT VISION



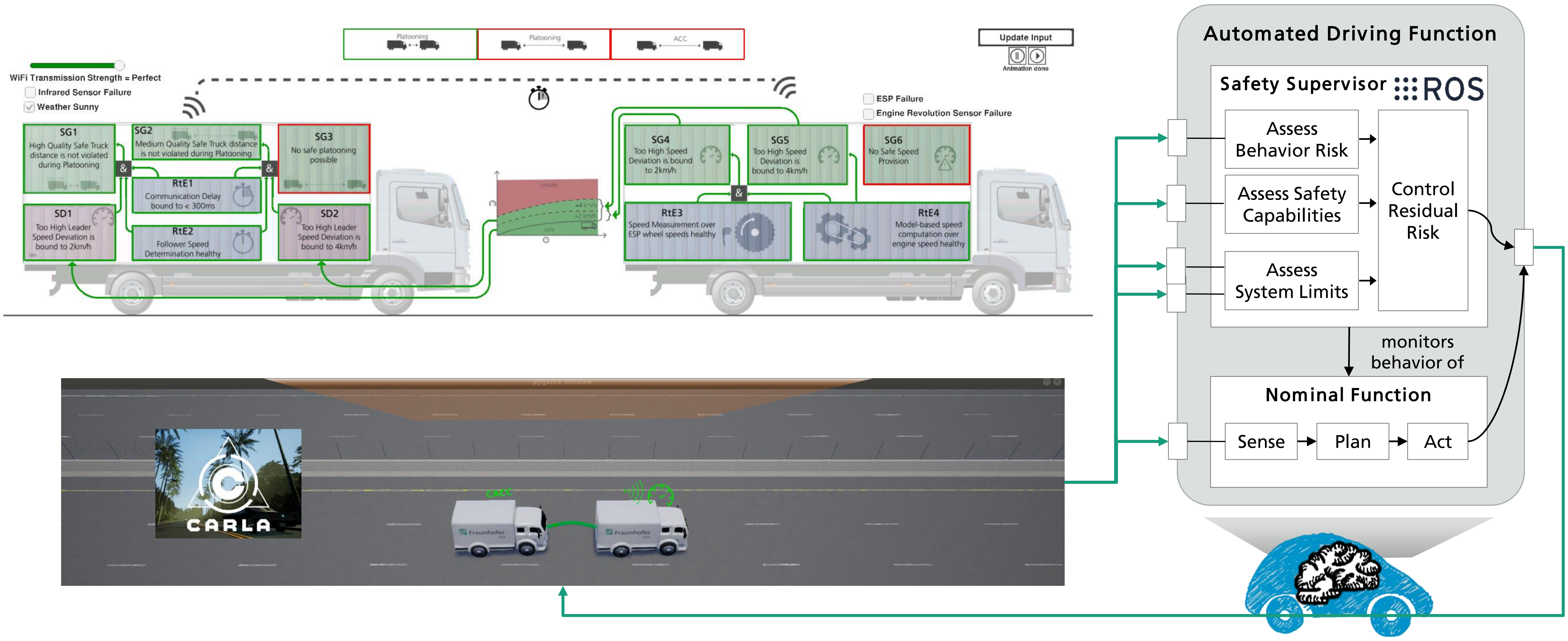
https://www.youtube.com/watch?v=HY9NrJHLxRI

DRM RUNTIME ARCHITECTURE



Fraunhofer IESE

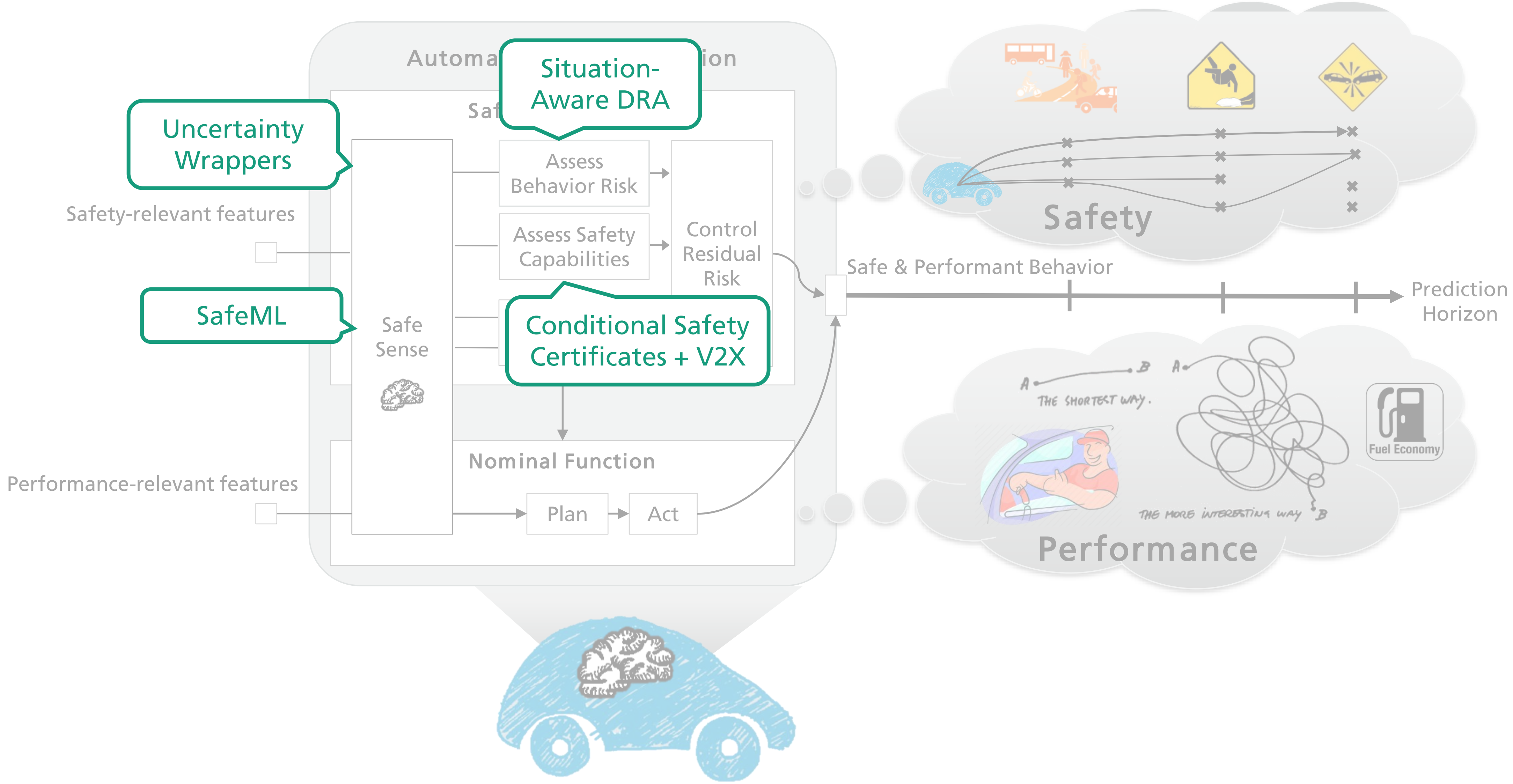
DYNAMIC RISK MANAGEMENT EXAMPLE



https://www.youtube.com/watch?v=Vdn-TCGxzgA

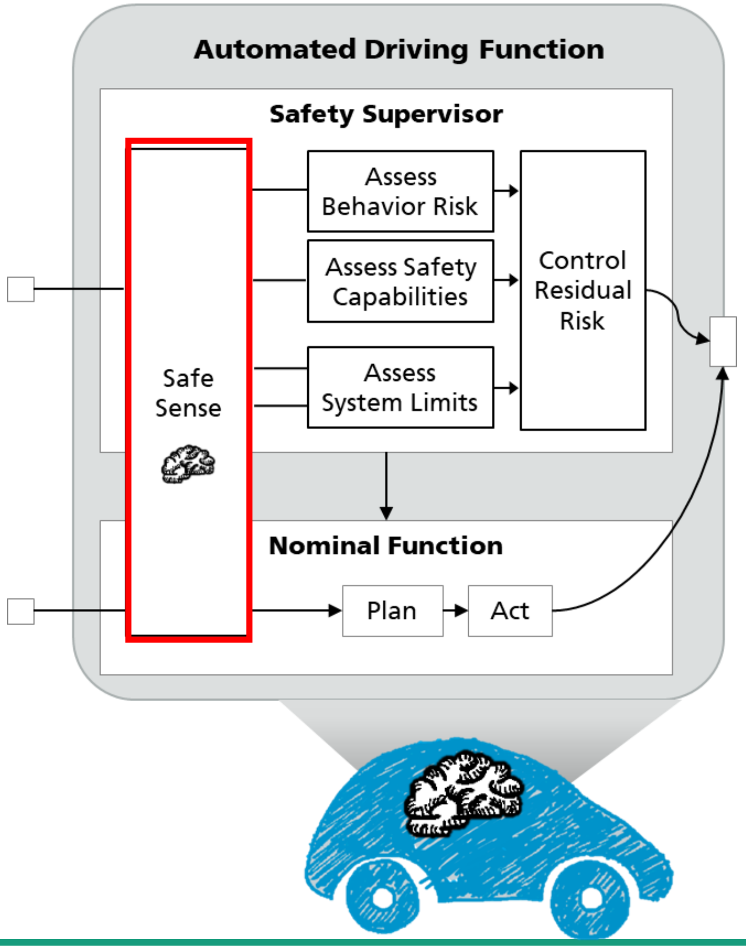
Fraunhofer IESE

FRAUNHOFER IESE TOPICS



Fraunhofer
IESE

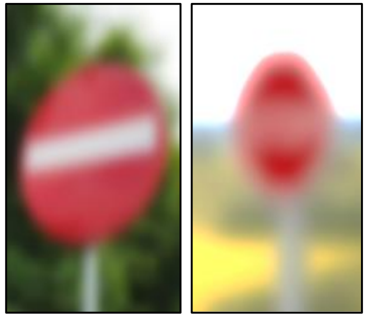
SAFE (ML-POWERED) SENSING



Fraunhofer IESE

Uncertainty Wrapper (Uw)

- **Challenge:** Uncertainty is inherent in data-based solutions and cannot be ignored
- **Approach:** "Uncertainty Wrapper" as a holistic, model-agnostic approach for the identification and situational reliable prognosis of uncertainty in AI-based components
- **Benefits**
  - Control of data management, model development and quality assurance
  - Expand the scope of action and reliably assure decision making at run-time when using the results of AI-based components
  - Setting up a convincing safety case (e.g., using GSM (goal-structuring notation) within the framework of Dynamic Risk Management



Outcome?

No Stop Sign

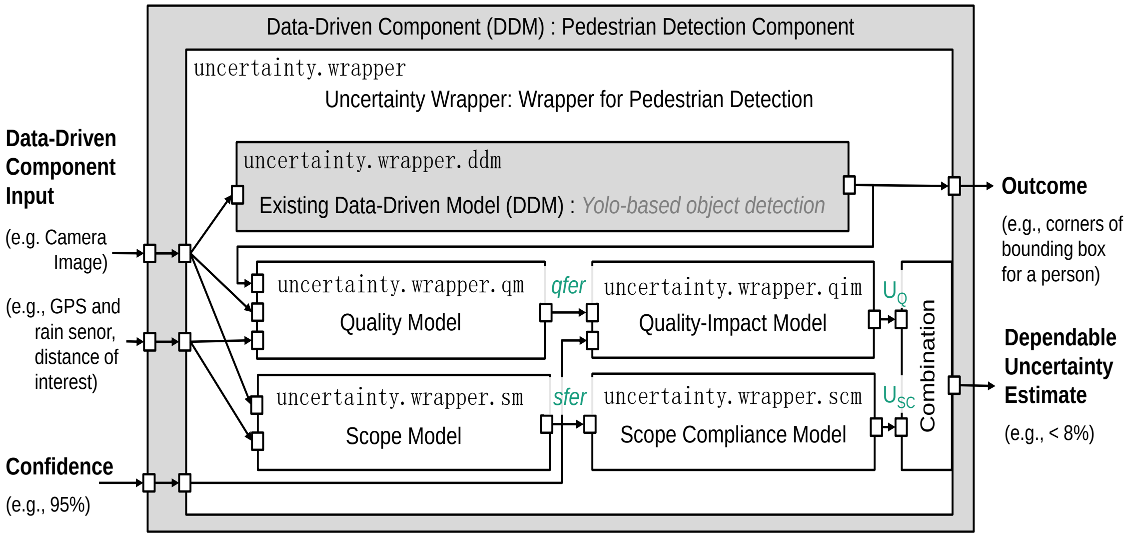
Uncertainty?

0.02 / 0.40

Confidence?

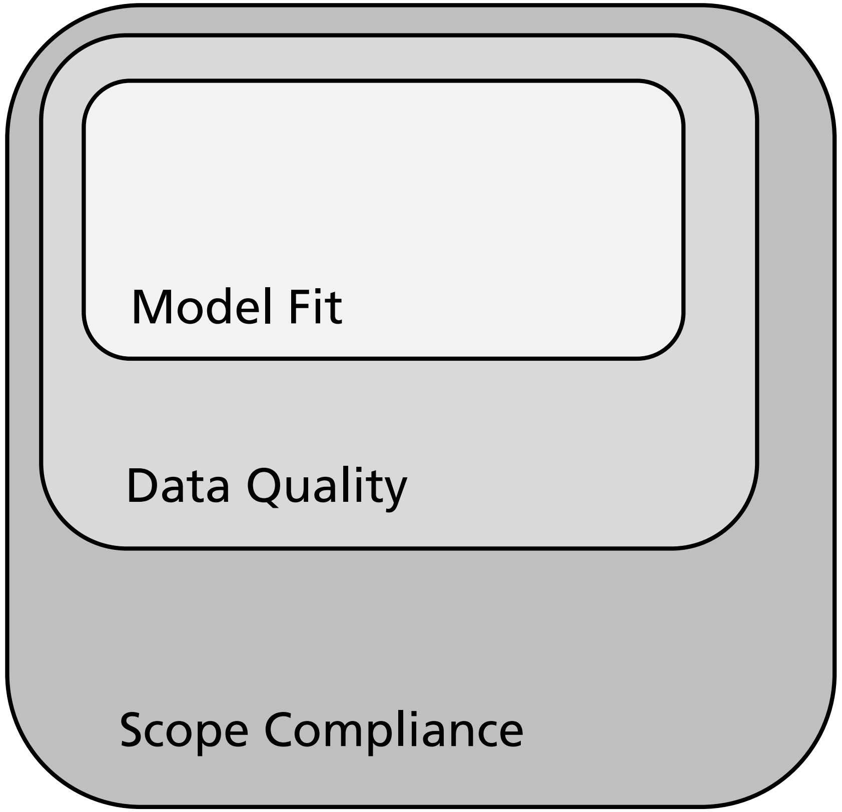
0.9999

Module developed in Python realizing scikit-learn estimator interface



Fraunhofer IESE

CAUSES FOR UNCERTAINTIES



Uncertainty caused by (inherent) limitations of the learned model

Additional

Uncertainty caused by data quality limitations during model application

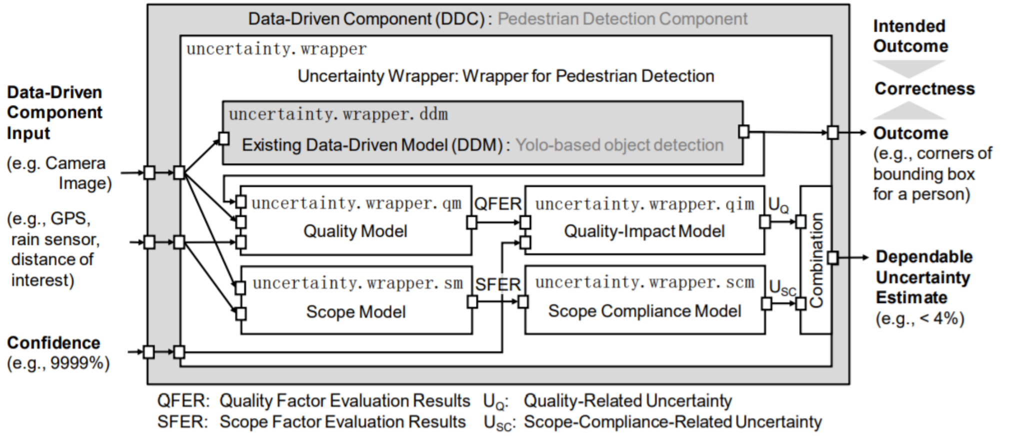
Additional

Uncertainty caused by mismatch between target/test context and application context

Fraunhofer IESE

Developing Uncertainty Wrappers

- Require representative dataset of ML model under control
  - Intended function and outcomes should be known i.e. supervised learning and labeled dataset
- Definition of correctness per each outcome known
- Quality Impact model specification
  - Determines how input quality across each input feature affects uncertainty of ML model outcome
- Scope compliance model specification
  - Specifies how to test whether we're inside or outside target application scope
  - Governed by scope factor models, which can be external ML models as well
- Available as Python library, compliant to scikit-learn interface
- Can be integrated into ML QA process



Quality Impact Model (Decision Tree)

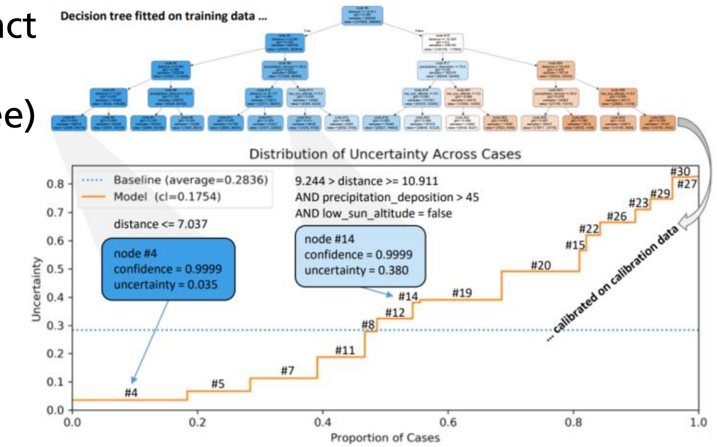


Fig. 3 A calibrated decision-tree-based quality impact model with confidence=.9999 and its evaluation, visualizing uncertainty estimates and certainty loss (cl) in comparison to DDM baseline.

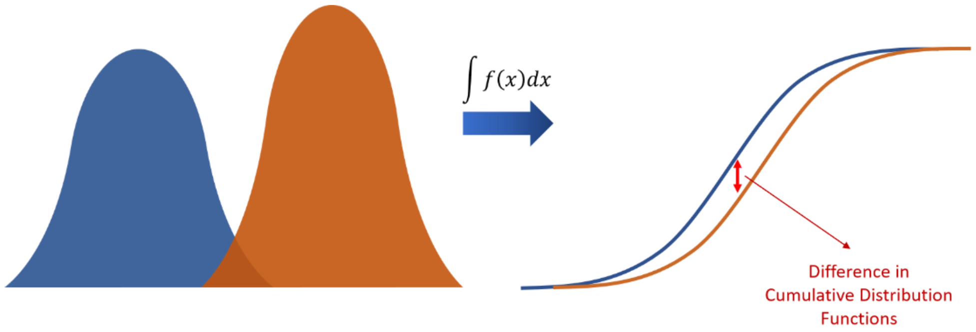
Source: http://klaes.org/Z-files/Klaes-2020-WAISE.pdf

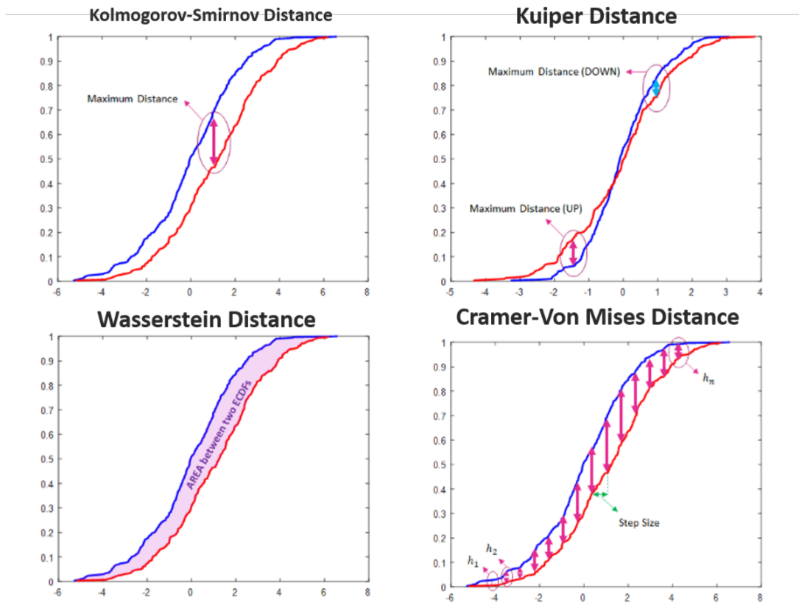
Fraunhofer IESE

SafeML

https://github.com/ISorokos/SafeML

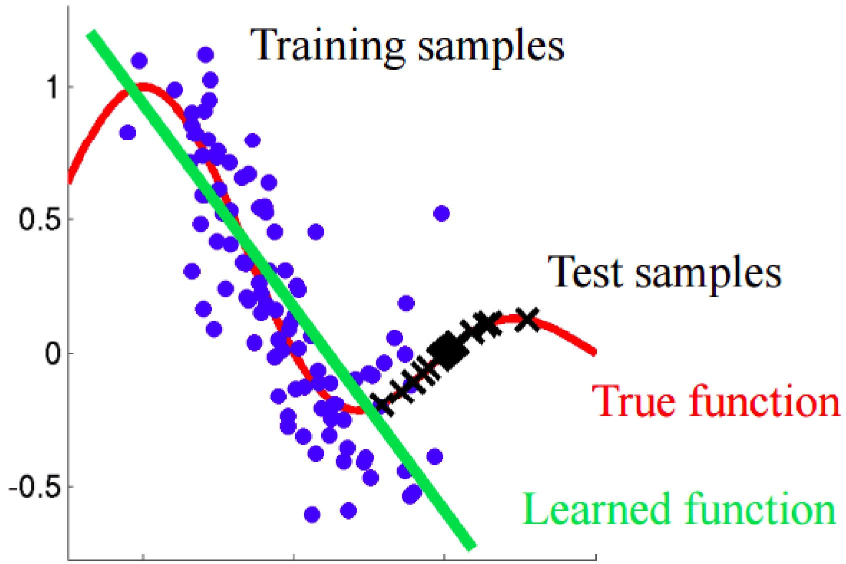
- **Challenge:** How do we know we're operating in the intended context ?
- **Our Approach:** SafeML uses statistical distance measures to evaluate 'how far' from our trained context are we currently operating in. If exceeding user-specified thresholds, alternative actions can then be employed.
- **Customer Benefits**
  - Monitor uncertainty of operational context compliance
  - Maintain safe state by not trusting ML when out of intended context





Dataset Shift

- Multiple definitions / similar terms over time
  - Dataset/Concept shift/drift
- Common theme
  - The data you originally trained with no longer applies
  - Can happen during training, but also during operation
- Specific topics include
  - Shift detection
  - Shift explanation/analysis
  - Shift response

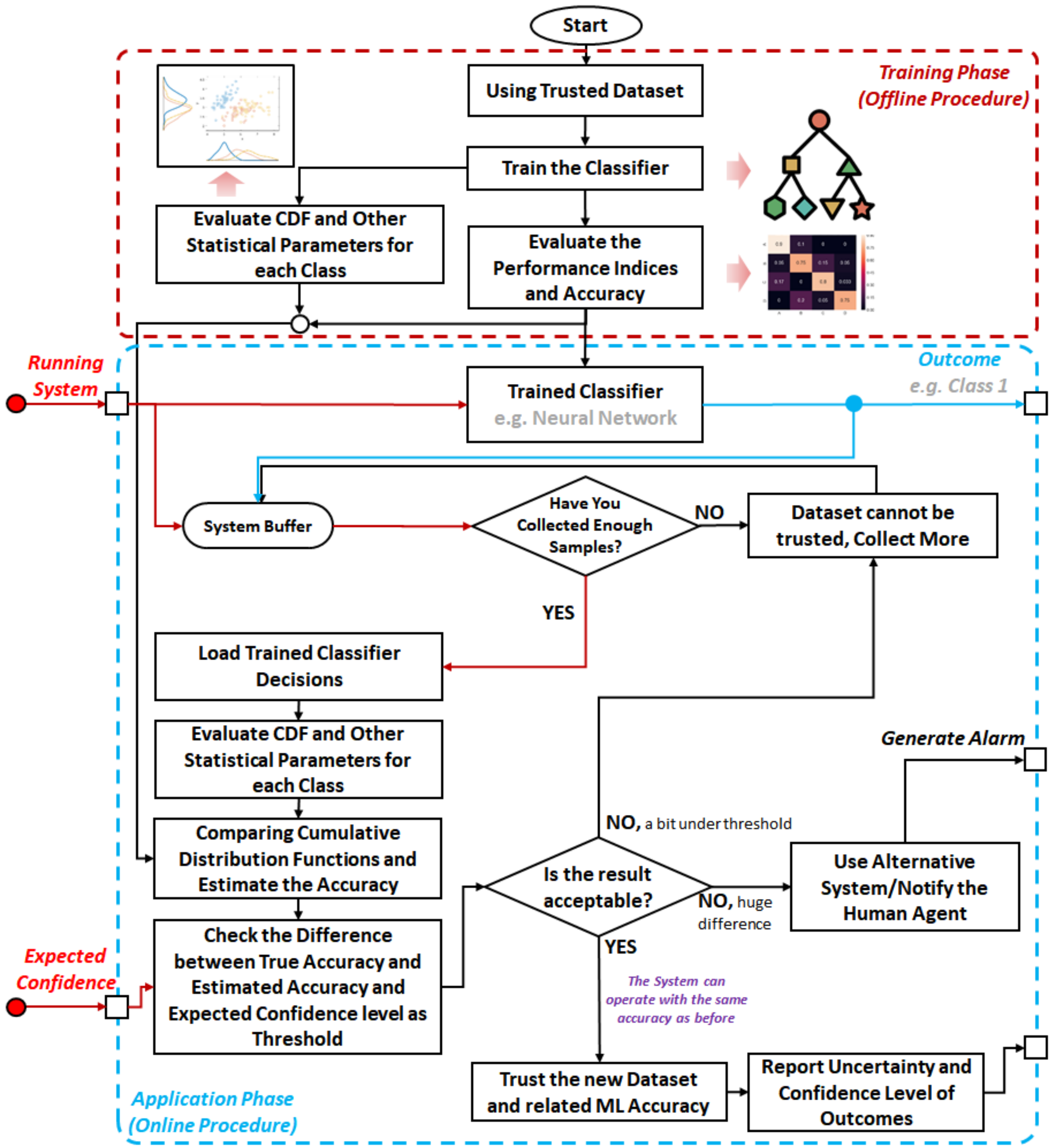


Covariate shift: input distribution changed

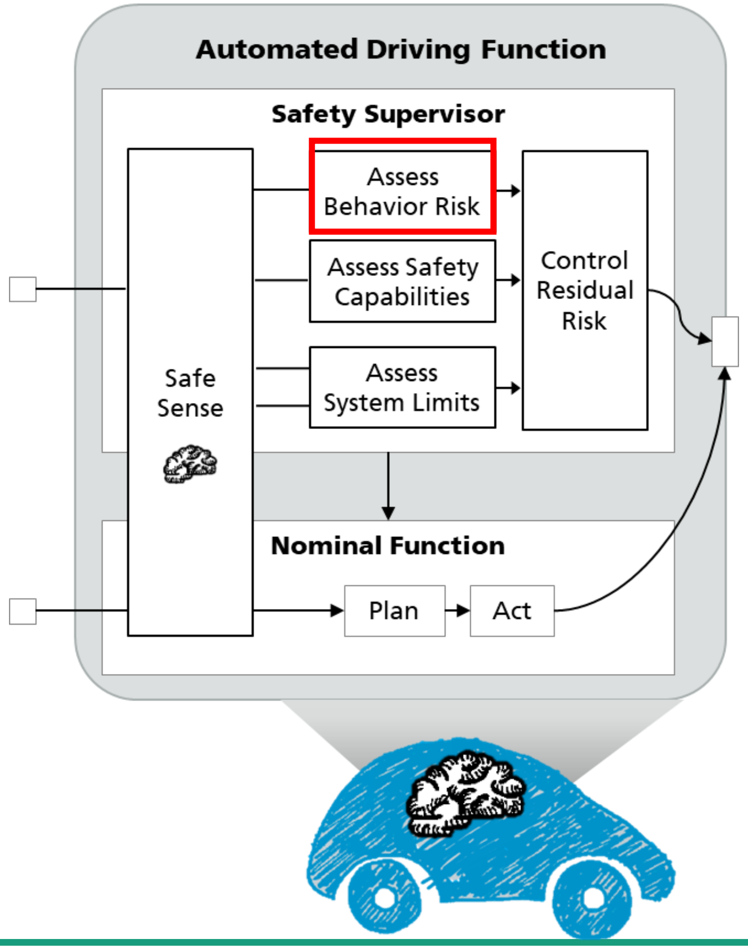
https://www.section.io/engineering-education/correcting-data-shift/

SafeML: Example Workflow

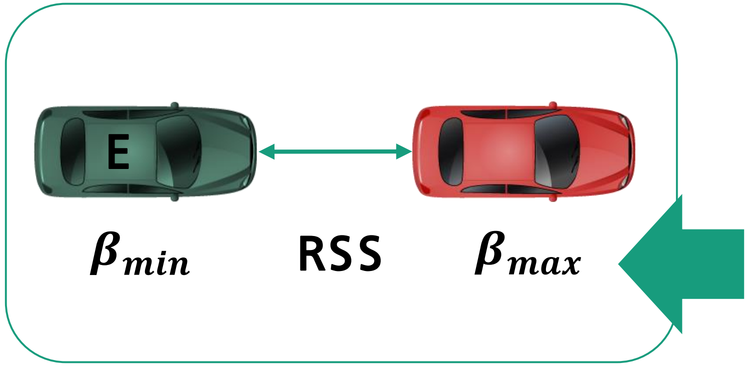
- Two stages:
  - Setup during ML Training
  - Deploy during ML Operation
- During training, store ECDF descriptors
- During operation
  - Sample from operational data
  - Form operational ECDF
  - Compare with stored
  - If distance > threshold
    -> alarm/user intervention/…



FOCUS ON DYNAMIC RISK ASSESSMENT

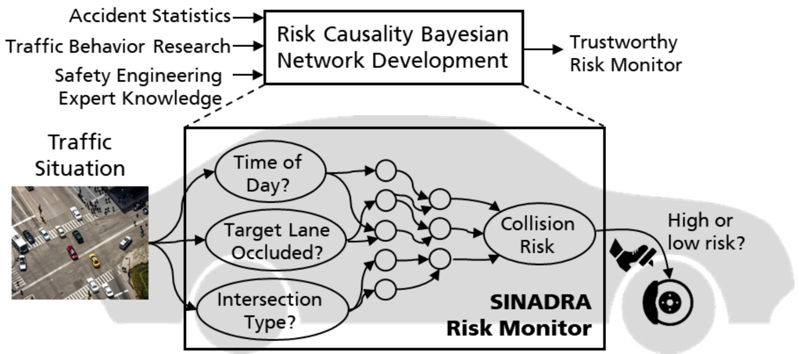


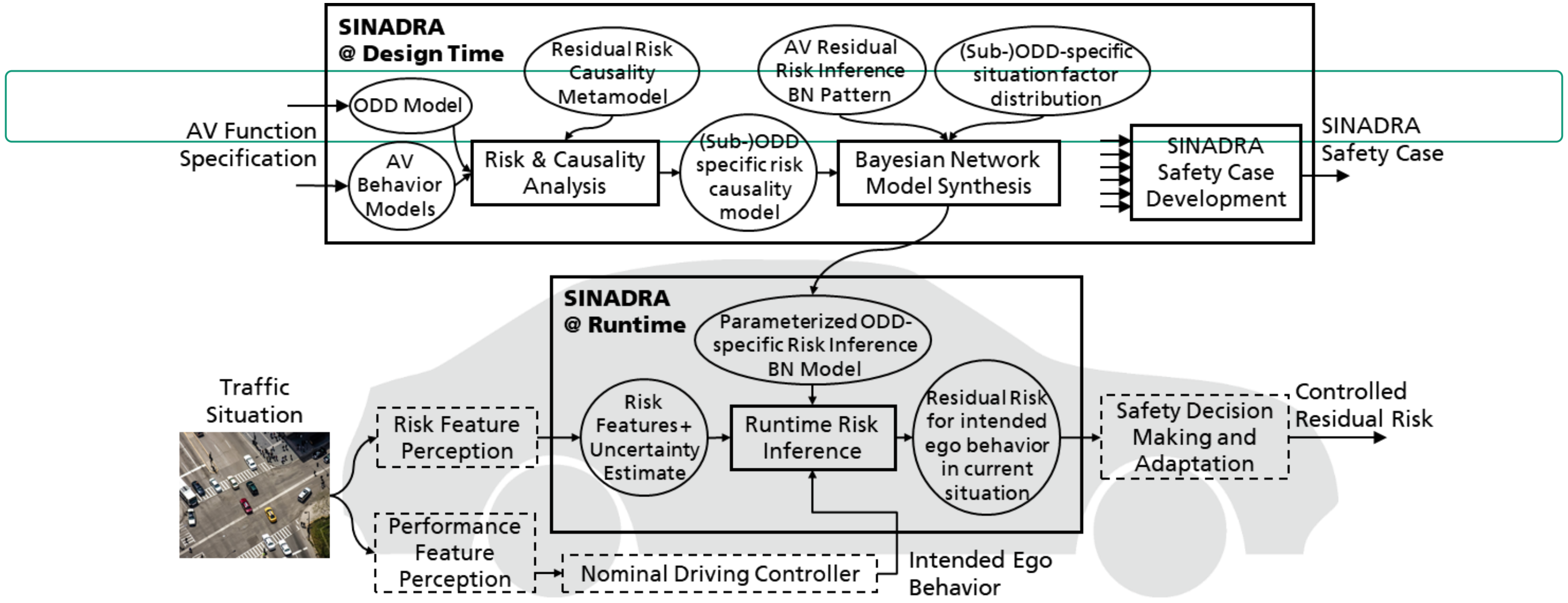
Dynamic Risk Assessment Research @ IESE



Situation-aware Dynamic Risk Assessment of Autonomous Vehicles (SINADRA)

- How can kinematic-based risk metrics be extended with situational awareness?
- How to quantify relationship between feature presence and risk?
- How can perception uncertainties be propagated to risk estimate?
- Formal relation to design time safety engineering (HARA) and safety case

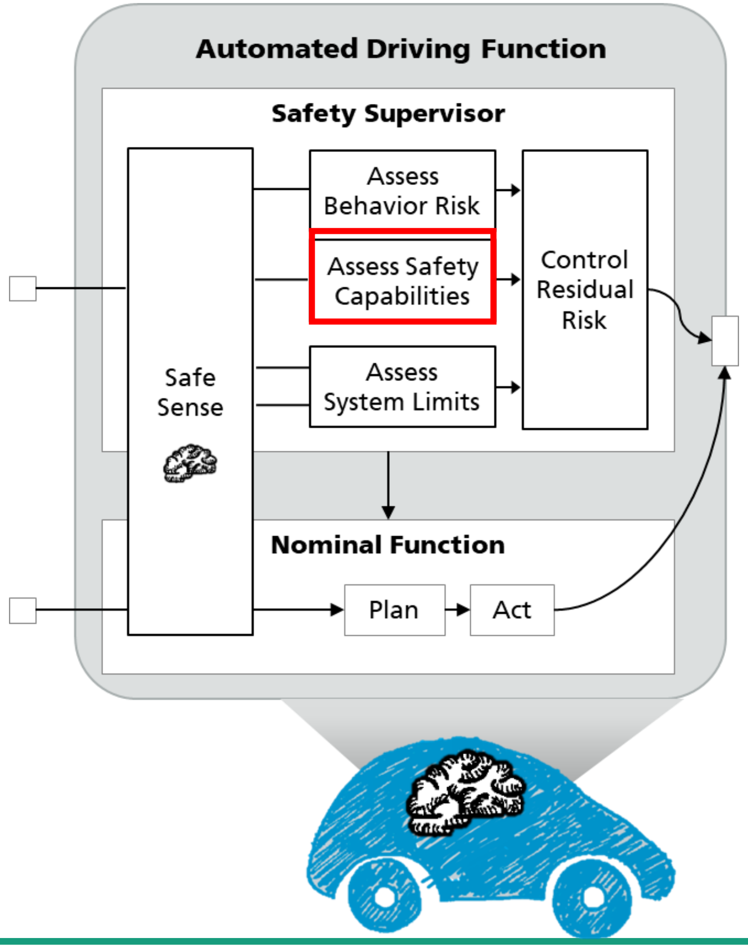




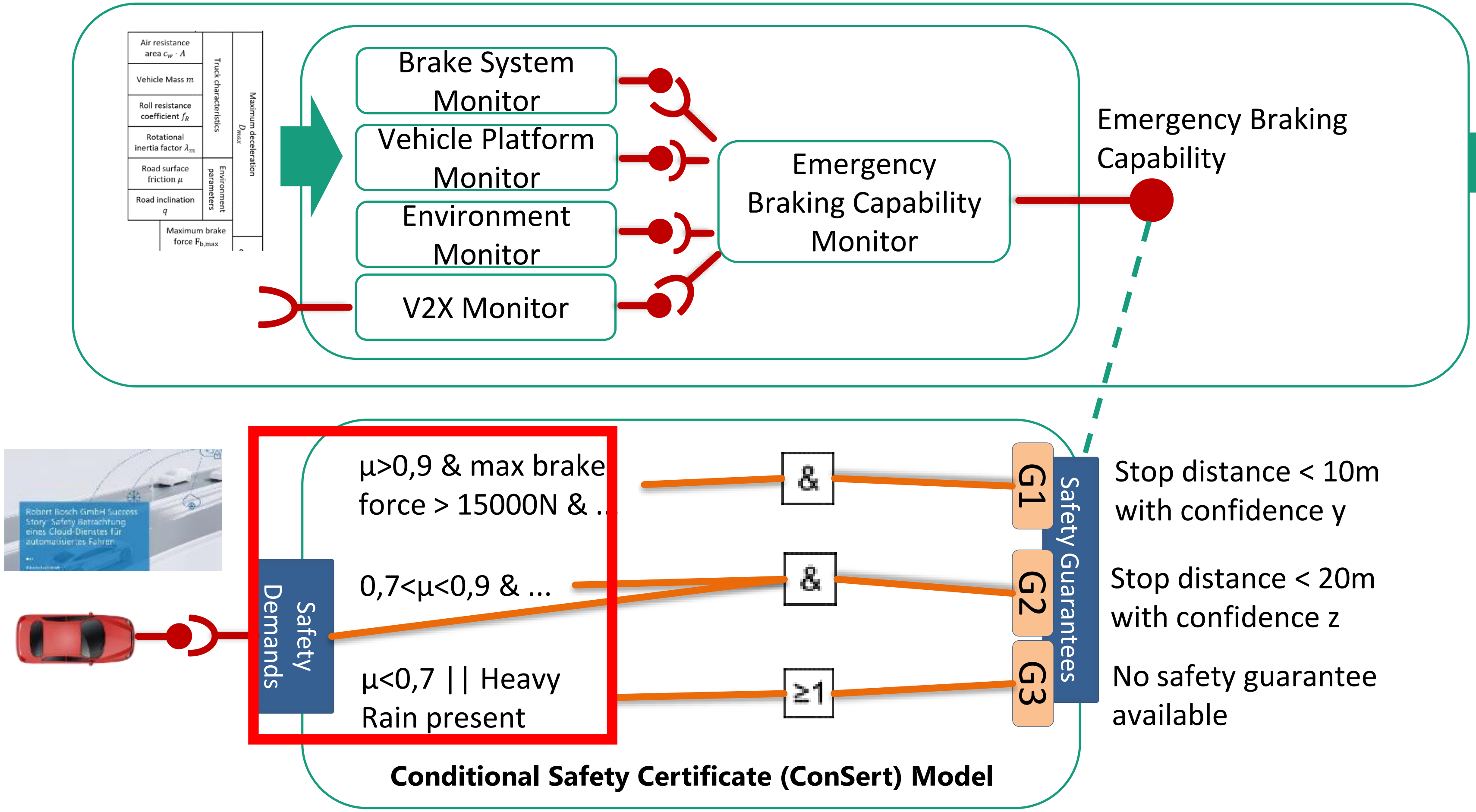
https://www.youtube.com/watch?v=fso4pAIcoUw

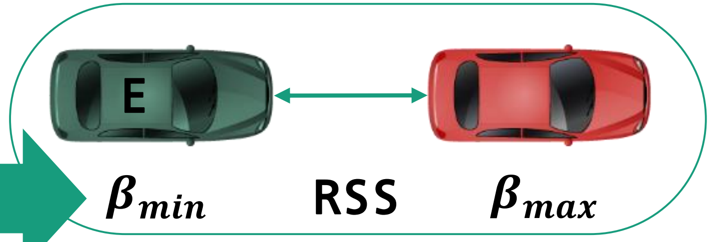
Fraunhofer IESE

FOCUS ON DYNAMIC CAPABILITY ASSESSMENT CONSERTS



Dynamic Safety Capability Assessment Research @ IESE - ConSerts





- ConSerts are preassured modular safety concept variant runtime models
- Algorithms for runtime composition and safety guarantee evaluation
- Supports heterarchically structured systems
- Evaluation and Improvement since 2013 in various research and industry projects and domains: automotive, agricultural, smart manufacturing, medical

Fraunhofer IESE

SUMMARY

- Using ML components in (safety-critical) systems has huge potential, quality assurance (and safety assurance in particular) is a big challenge
- There is no single silver bullet for assuring safety of systems with ML-components, a specific concept is always required
- There is no commonly accepted state of the practice or even a sound understanding with respect to suitable engineering methods, techniques and tools
- This talk gave an (selective) overview on challenges and solution ideas along an envisioned integrated safety and ML engineering lifecycle
  - General solution approaches, recommendations, DRM, dealing with uncertainty!

Thank you for your interest

Contact:

Dr. Daniel Schneider
daniel.schneider@iese.fraunhofer.de
Tel.: +49 (0) 631 / 6800-2187
Fax.: +49 (0) 631 / 6800-9-2187
Mobile: +49 (0) 151 / 649 530 70