Towards Safe AI for Automated Driving

Fabian Hüger, Volkswagen & CARIAD

EDCC 2021, September 16, 2021

Agenda

1. Introduction – CARIAD
2. DNNs and Safety in Automated Driving
3. KI-Absicherung Project & Approach
4. Consequences

We deliver Volkswagen's answer to the digitalization of mobility

CARIAD *TRANSFORMS* →

- *Product Offers*
- *Business Models*
- *Customer Experience*
- *Supply Chains*
- *Vehicle Platforms*
- *Processes*
- ...

→

CARIAD does not only deliver software.

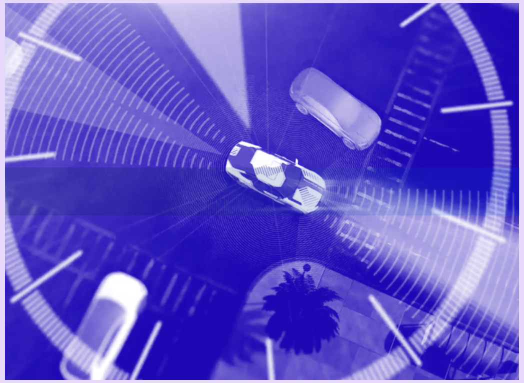
It reinvents mobility and the way Volkswagen works

Our Solutions

Our solutions are structured in technology domains and product enablers

OUR TECHNOLOGY DOMAINS

Intelligent Body & Cockpit



Autonomous Driving

Vehicle Motion & Energy

Digital Business & Mobility Services

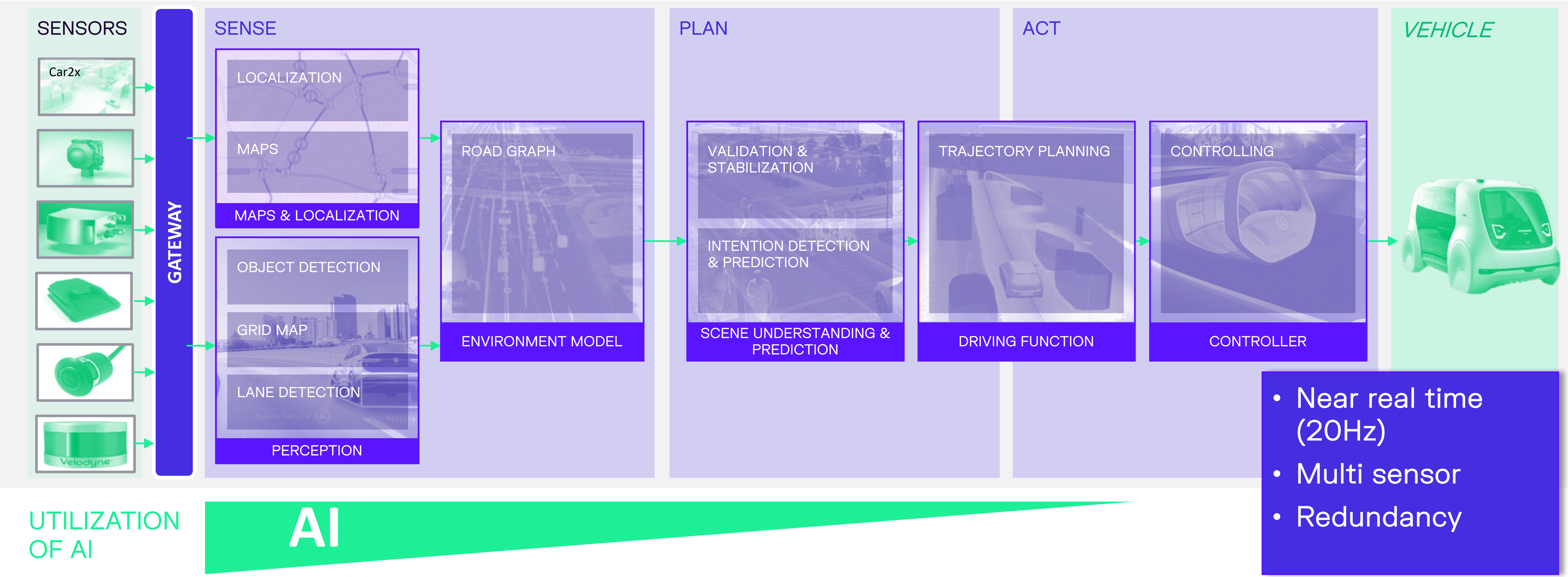
OUR PRODUCT ENABLERS

- *Vehicle Platform & Operating Systems*
- *Chief Security Office*
- *Architecture*
- *Integration*
- *PMT (Processes, Methods & Tools*
- *Embedded Quality*

Agenda

Automated Driving and AI

Processing chain of autonomous driving & the use of AI along



- Near real time (20Hz)
- Multi sensor
- Redundancy

Arguing Safety in Automated Driving Systems

AI goes safety critical

CENTRAL CHALLENGE

SAFETY

(FuSa + SOTIF)

Central Challenge in bringing highly automated driving on the road.

Argument on safe functioning needed to allow for acceptance & road permission

COMPLEXITY DRIVERS

Mere driving will not suffice to plausibilize safety – particularly challenging with respect to software updates over time. "Black-Box" approach seems impracticable

Handling complexity of the driving environment – open world, unknown unknowns, etc.

Need for continual safety monitoring & assurance – continuous monitoring

Arguing Safety in Automated Driving Systems
Standardization Activities

EXISTING STANDARDS

ISO 26262

- E/E failures
- Classification in ASIL-Levels
- No defined ML-specifics (in discussion for the 3rd edition)

ADDITIONAL NORMS & DOCUMENTS

SOTIF

ISO 21448

- **Behavioral safety** (describing performance limitations and triggering conditions alongside mitigation techniques)
- Highly relevant for non-fully specified perception systems for which DNNs seem to be standard

UL4600

- **Safety alongside development** process - Level-4 specific, more AI details
- Focus within the development process – reporting on design decisions with respect to raise resulting safety is key
- Consequently: yielding need for a strong **traceability of performance and safety evidence** to development decisions.

WORK IN PROGRESS

ISO Activities

+

ASAM working groups

Approaching standards for dependability of AI:

- ISO/IEC JTC1 SC42 activities (ISO TR5469, ISO/IEC TR 24029)
- ASAM working groups
- ISO TR 4804
- ISO TS 5083
- ISO NWIP Road Vehicles: Safety & AI

Agenda

1. Introduction – CARIAD
2. DNNs and Safety in Automated Driving
3. KI-Absicherung Project & Approach
4. Consequences

Acknowledgement: The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project "Methoden und Maßnahmen zur Absicherung von KI basierten Wahrnehmungsfunktionen für das automatisierte Fahren (KI-Absicherung)". The authors would like to thank the consortium for the successful cooperation.

KI ABSICHERUNG

Safe AI for Automated Driving

KI-Absicherung Project & Approach

The results, opinions and conclusions expressed in this publication are not necessarily those of Volkswagen Aktiengesellschaft.

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

www.ki-absicherung-projekt.de @KI_Familie KI Familie

Making the safety of AI-based function modules for highly automated driving verifiable

Pedestrian detection

Challenge

AI Land

Pixabay

Promising new technology with unimagined possibilities

Established safety processes cannot be applied

KI ABSICHERUNG

Safe AI for Automated Driving

Safety Land

Pixabay

Safe, trustworthy driving function

Industry consensus (Safe AI): Methodology for joint safety argumentation

Our Team: Experts from AI, Safety and Virtual Reality

OEMs

VOLKSWAGEN AKTIENGESELLSCHAFT | BMW GROUP

Tiers

BOSCH Invented for life | Continental | Valeo | ZF | EFS

Technology Provider

ASTech Automotive Safety Technologies | intel | luxoft A DXC Technology Company | MACKEVISION | MERANTIX | qualityminds

Research Institutes

DFKI Deutsches Forschungszentrum für Künstliche Intelligenz GmbH | DLR | Fraunhofer IAIS | Fraunhofer IKS | FZI

Universitys

Bergische Universität Wuppertal | Technische Universität München TUM | Universität Heidelberg

External Partners

BIT Technology Solutions | neurocat | understand.ai | eict

Consortium Lead: Volkswagen AG

Co - Lead: Fraunhofer IAIS

Budget: 41 Mil. €

BMWi Fundinng: 19.2 Mil. €

24 Partners

Duration: 36 month

01.07.2019 - 20.06.2022

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

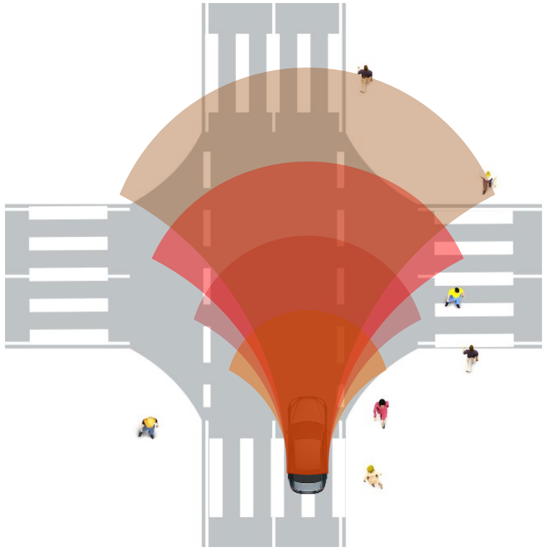
aufgrund eines Beschlusses des Deutschen Bundestages

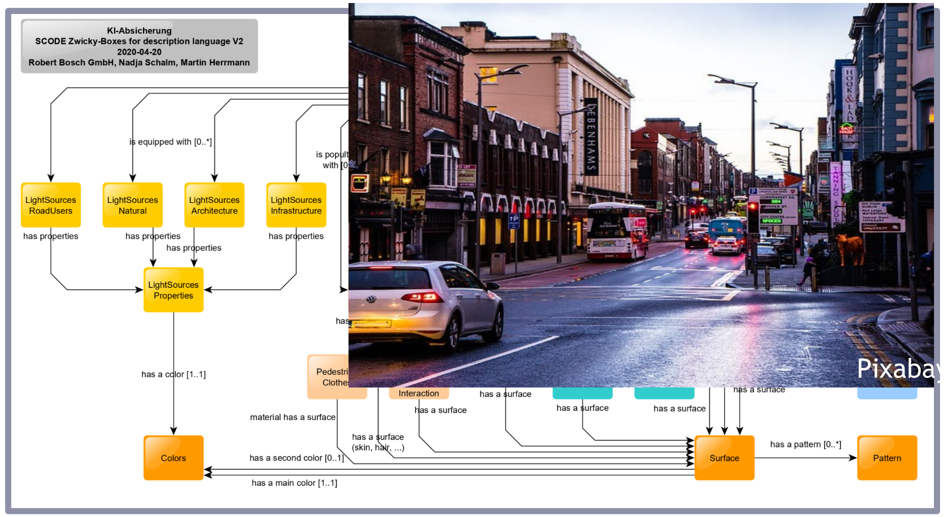
Our Approach: Specification

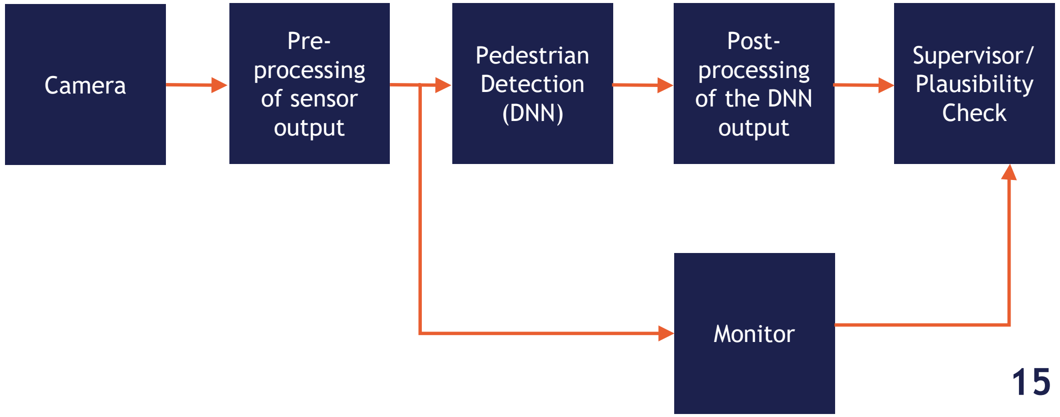
Function

ODD

System Architecture







Our Approach: AI Function Pedestrian detection

Function

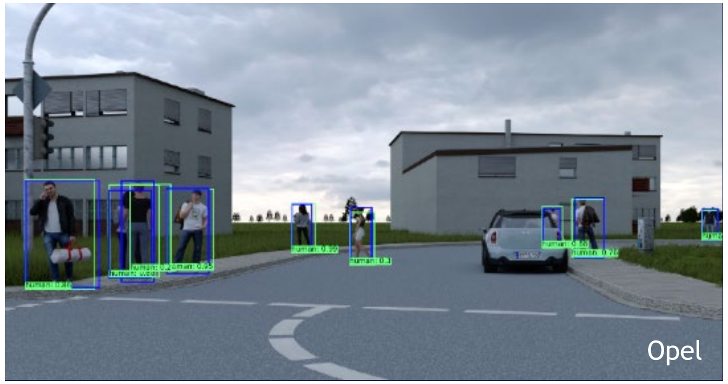
ODD

System Architecture

AI Function

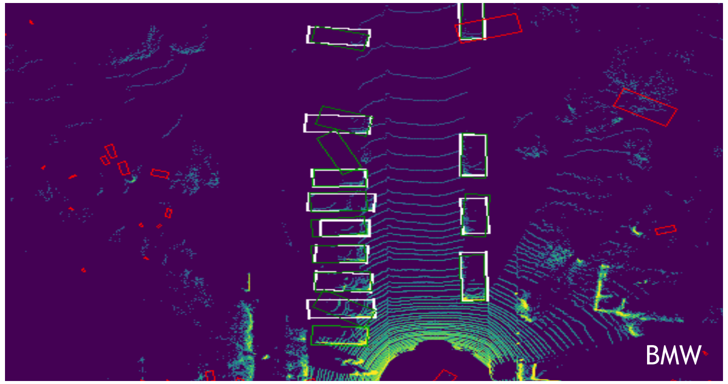
Semantic Segmentation

2D Bounding Box Detection

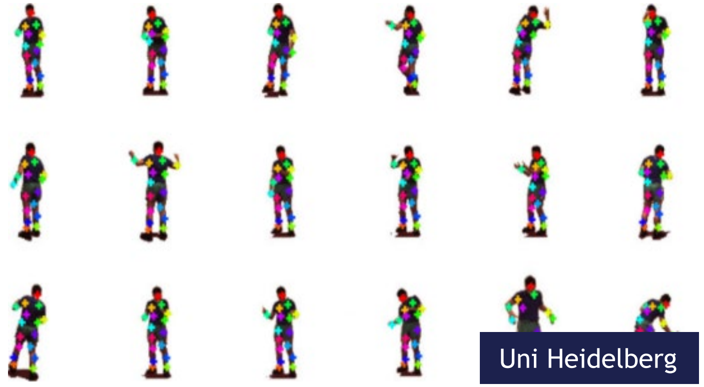


Instance Segmentation

3D Bounding Box Detection



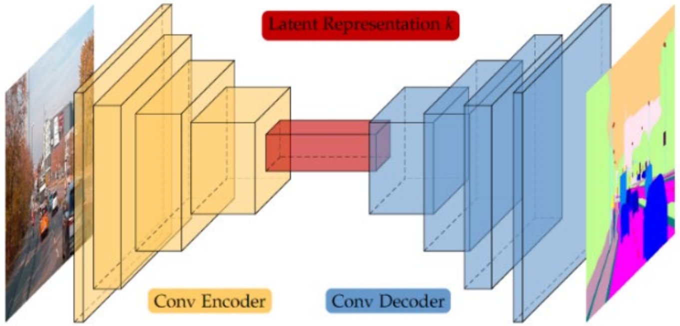
3D Pose estimation



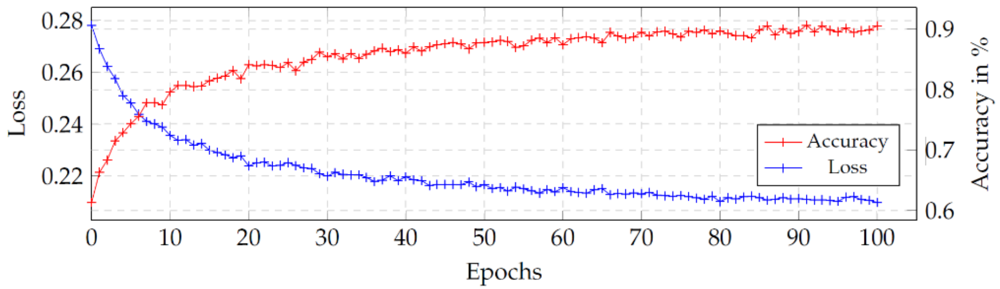
Our Approach: Synthetic Data and ML-Lifecycle

ML LIFECYCLE

- ML-Data
- Model design
- Pre-Processing
- Training
- Post-Processing
- Test/ V&V incl. Data
- Monitoring
- Maintenance



Volkswagen AG



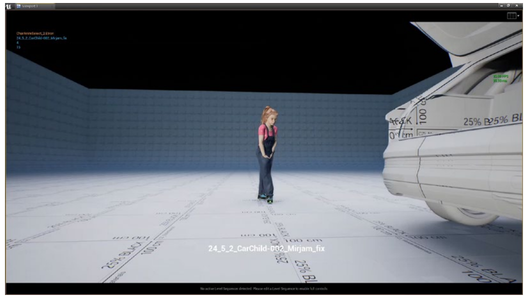
Volkswagen AG

Our Approach: ML-Lifecycle-Validation data

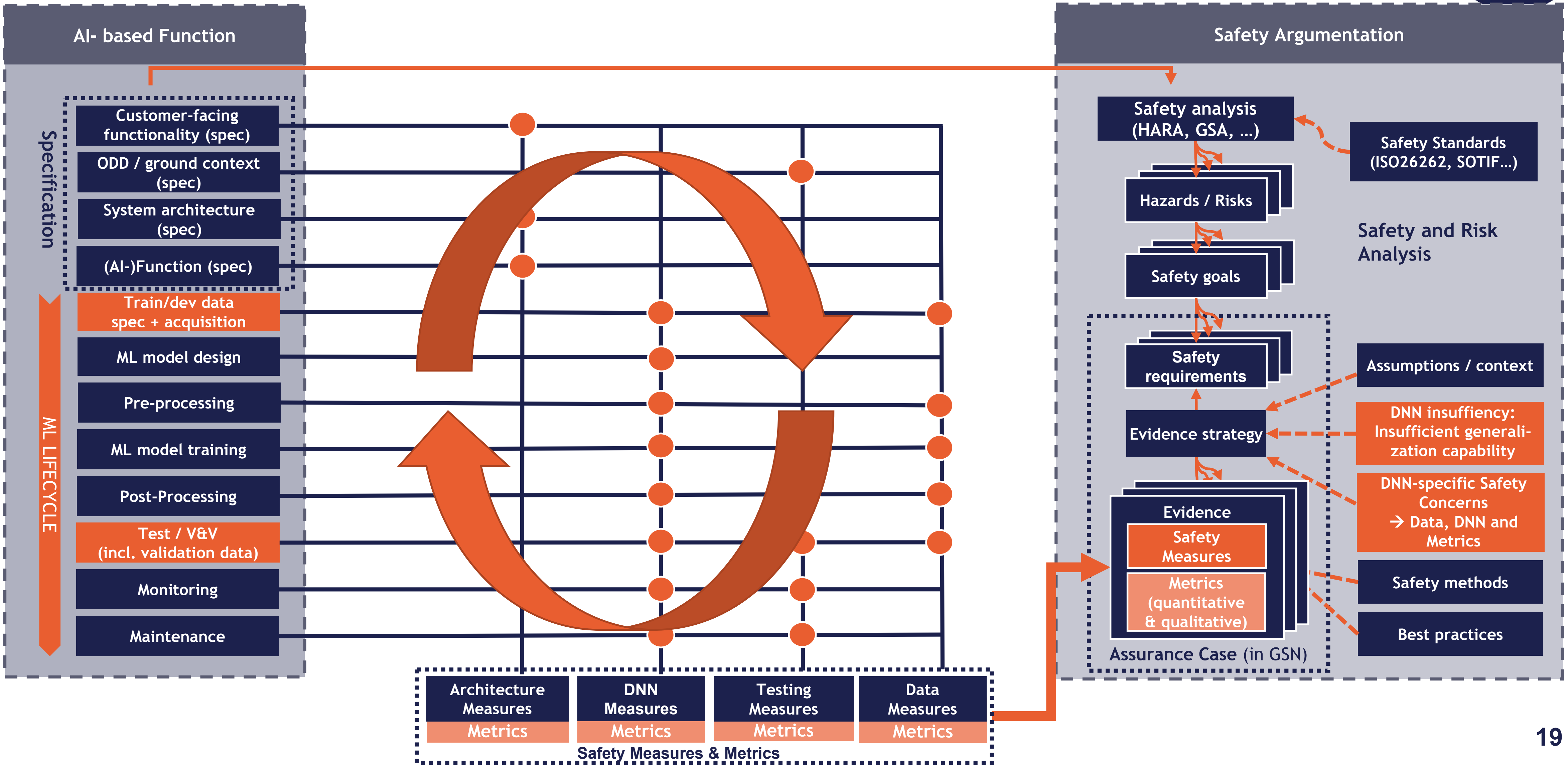
ML LIFECYCLE

- ML-Data
- Model design
- Pre-Processing
- Training
- Post-Processing
- **Test/ V&V incl. Daten**
- Monitoring
- Maintenance

Continuous process for identification, specification and generation of synthetic data



Our Approach: Big picture



Our Approach: DNN-specific Safety Concerns (1/2)

We define **DNN-specific Safety Concerns (SCs)** as underlying issues of DNN-based perception which may negatively affect the safety of a system.

e.g., fog, snow, camera issues

Triggering Event

Erroneous Output

DNN-characteristics-related concerns

Data-related concerns

Metric-related concerns

DNN-specific Safety Concerns

Insufficient Generalization Capability

Functional Insufficiency according to ISO/PAS 21448

DNN-specific Safety Concerns

FI-1 INSUFFICIENT GENERALIZATION CAPABILITY
Wrong outputs by an AI-based function that was trained on a limited database. Erroneous input to output mapping or wrong approximation.

SC-1.1 UNRELIABLE CONFIDENCE INFORMATION
DNNs tend to be overconfident in their predictions under certain conditions or in general outputting unreliable confidence information.

SC-1.2 BRITTLENESS OF DNNS
Non-robustness against common perturbations such as noise or certain weather conditions as well as targeted perturbations known as adversarial examples

SC-1.2.1 LACK OF TEMPORAL STABILITY
Detection results rapidly changing in time whereas little change occurs in the ground truth

SC-1.3 INCOMPREHENSIBLE BEHAVIOUR
Inability to explain exactly how DNNs come to a decision.

SC-1.4 INSUFFICIENT PLAUSIBILITY
AI based functions usually lack basic plausibility checks, which are intended to identify detections of the perception function that violate physical laws.

SC-2.1 DATA DISTRIBUTION IS NOT A GOOD APPROXIMATION OF REAL WORLD
The distribution of data used in the development should be a valid approximation of the ODD in the real world.

SC-2.2 INADEQUATE SEPARATION OF TEST AND TRAINING DATA
Test data might be correlated to training data which might induce overfitting on test data.

SC-2.3 DEPENDENCE ON LABELLING QUALITY
Labelling quality can directly affect the resulting model performance. Moreover, due to missing labelling quality, evaluation results might be misleading.

SC-2.3.1 MISSING LABEL DETAILS OR META-LABELS
Missing meta-labels or label details possibly leads to improper data selection or insufficient training objectives.

SC-2.4 SPECIFICATION OF THE ODD
An incomplete or incorrect ODD specification leads to incomplete data records for training and testing.

SC-2.5 DISTRIBUTIONAL SHIFT OVER TIME
A DNN is trained and tested at a certain point in time. Changes will occur naturally and therefore can potentially harm the performance of DNNs.

SC-2.6 UNKNOWN BEHAVIOUR IN RARE CRITICAL SITUATIONS
The long tail problem describes the fact that there exists an enormous amount of possibly safety-critical street scenes that have a low occurrence probability.

SC-3.1 SAFETY-AWARE METRICS
Some state-of-the-art metrics only evaluate the average performance of DNNs. Safety-aware metrics are required to sophistically evaluate the performance of DNNs.

Functional Insufficiencies

DNN-characteristics-related concerns

Data-related concerns

Metric-related concerns

Based on:

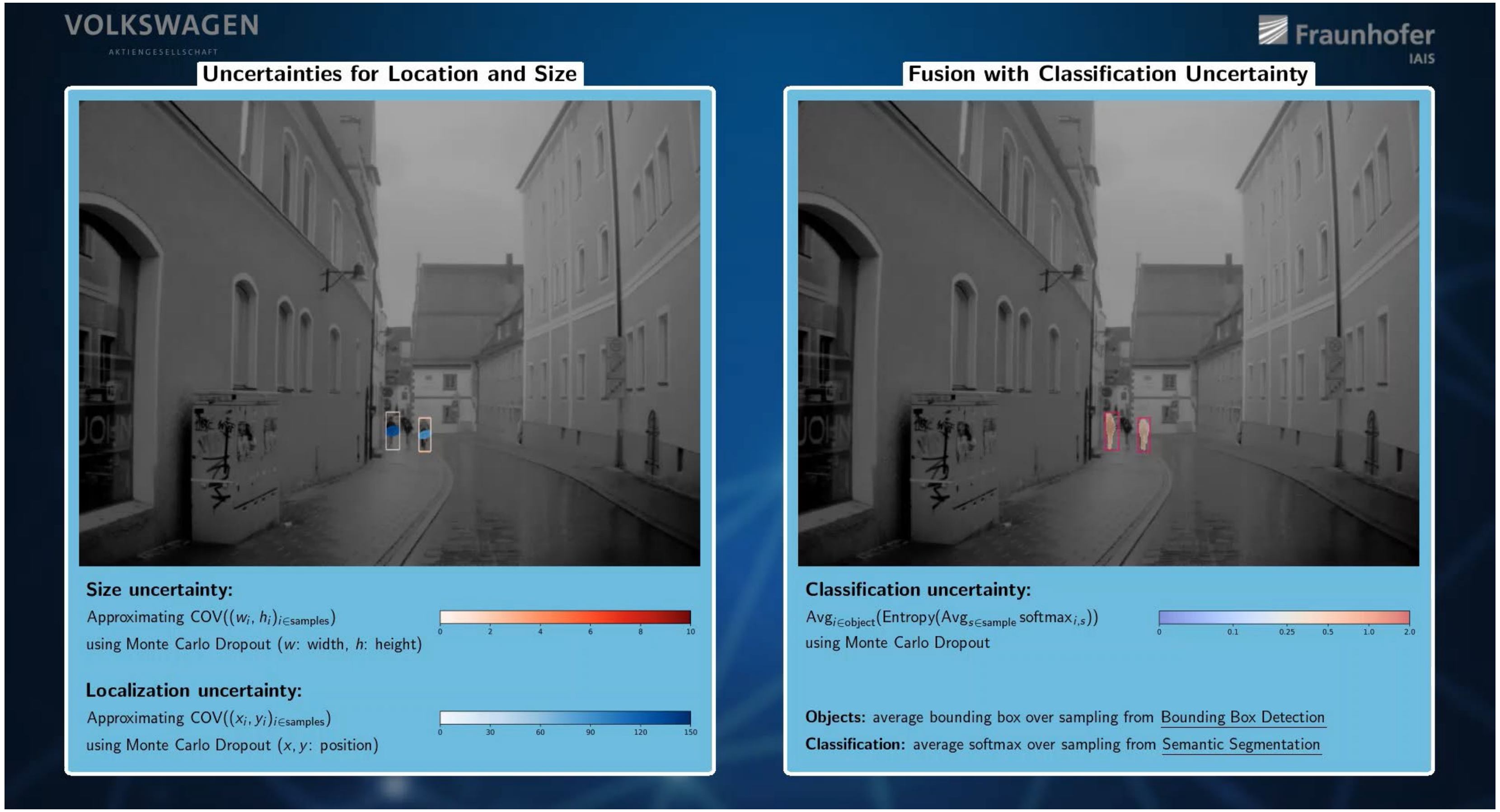
O.Willers, S. Sudholt, S. Raafatnia, S. Abrecht: Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks

T. Sämann, P.Schlicht, F. Hüger: Strategy to Increase the Safety of a DNN-based Perception for HAD Systems

G. Schwalbe, B. Knie, T. Sämann, T. Dobberphul, L. Gauerhof, S., V. Rocco: Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications

Our Approach: Identify, Measure and & Counteract „DNN-specific Safety Concerns" via MC dropout

Adressed Safety Concern:
Unreliable Confidence
via MC dropout



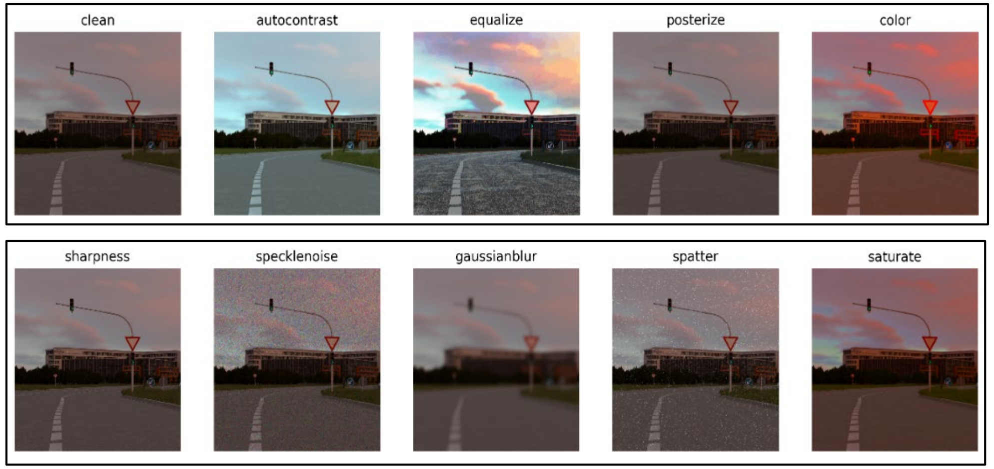
DNN-specific safety concern:
- Unreliable Confidence Information of DNNs

Method:
- Assessment of uncertainty: Stochastic evaluation of a multitude of model variations (Monte Carlo Dropout)
- Usage at design-time or run-time

Our Approach: Identify, Measure and & Counteract „DNN-specific Safety Concerns”

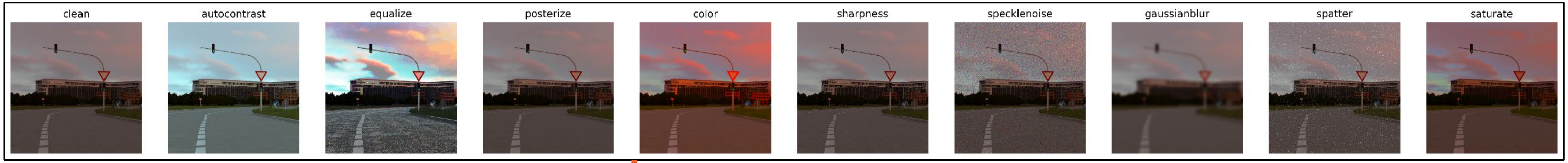
Adressed Safety Concern: Brittleness of DNNs

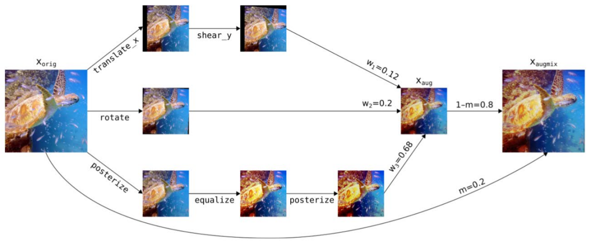
- Adressing **“Brittleness of DNNs” (Example)**
  - **Requirement:** Robustness = Performance even under reasonable perturbations (gained from ODD definition, data analysis and sensor specs)
  - **Metric:** Performance under corruption
  - **Methods (e.g.)**
    - Augmentation Training (**AugMix**)
    - From a Fourier-Domain Perspective on Adversarial Examples to a **Wiener Filter** Defense for Semantic Segmentation
  - **Evidence**: Effectiveness of measure via metric



Our Approach: Identify, Measure and & Counteract „DNN-specific Safety Concerns" via AugMix

Adressed Safety Concern: Brittleness of DNNs Corruption Robustness





AUGMIX: A SIMPLE DATA PROCESSING METHOD TO IMPROVE ROBUSTNESS AND UNCERTAINTY

Dan Hendrycks*
DeepMind
hendrycks@berkeley.edu

Norman Mu*
Google
normanmu@google.com

Ekin D. Cubuk
Google
cubuk@google.com

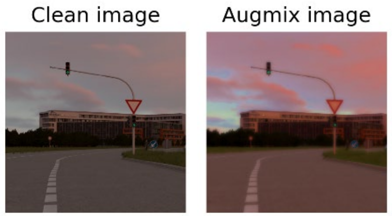
Barret Zoph
Google
barretzoph@google.com

Justin Gilmer
Google
gilmer@google.com

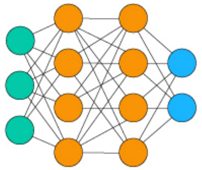
Balaji Lakshminarayanan†
DeepMind
balajiln@google.com

Combined using AugMix

+ Improved robustness
+ Improved generalization
+ Data efficient augmentation strategy



Training

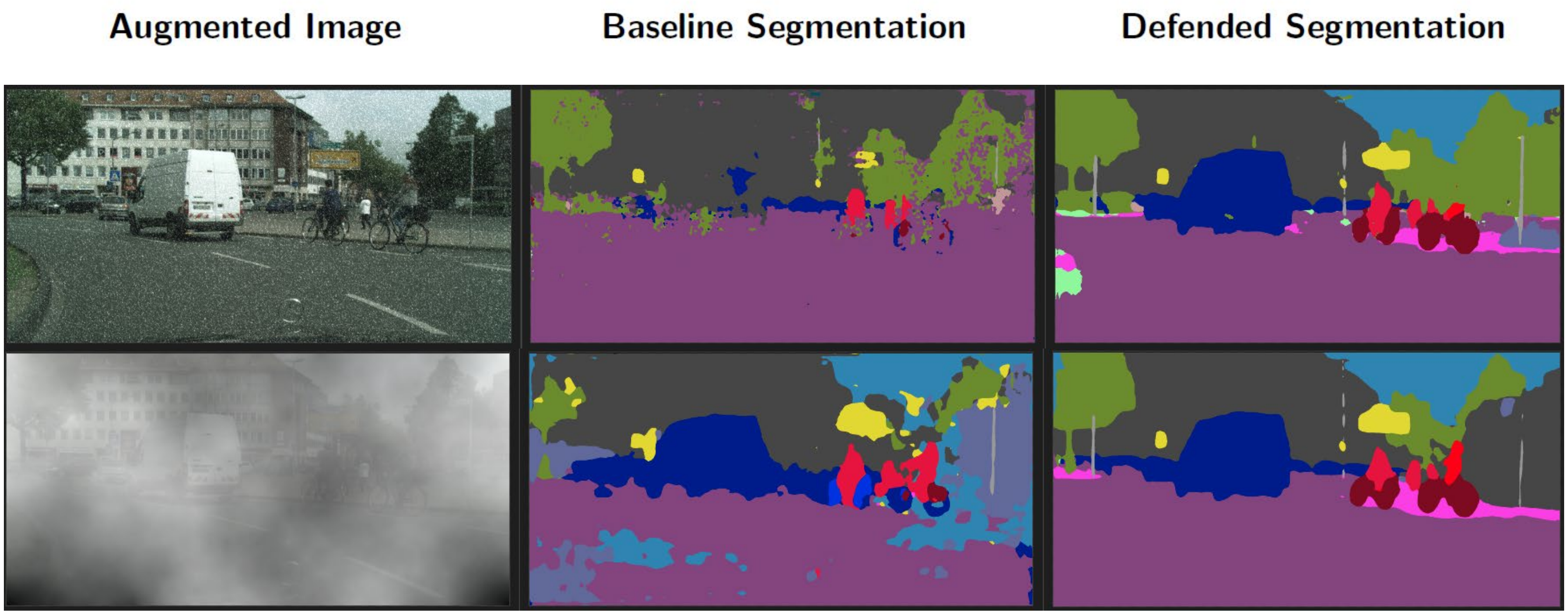


Evaluation on 14 *unseen* „real-world" corruptions

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, D. Hendrycks et al, https://arxiv.org/abs/1912.02781

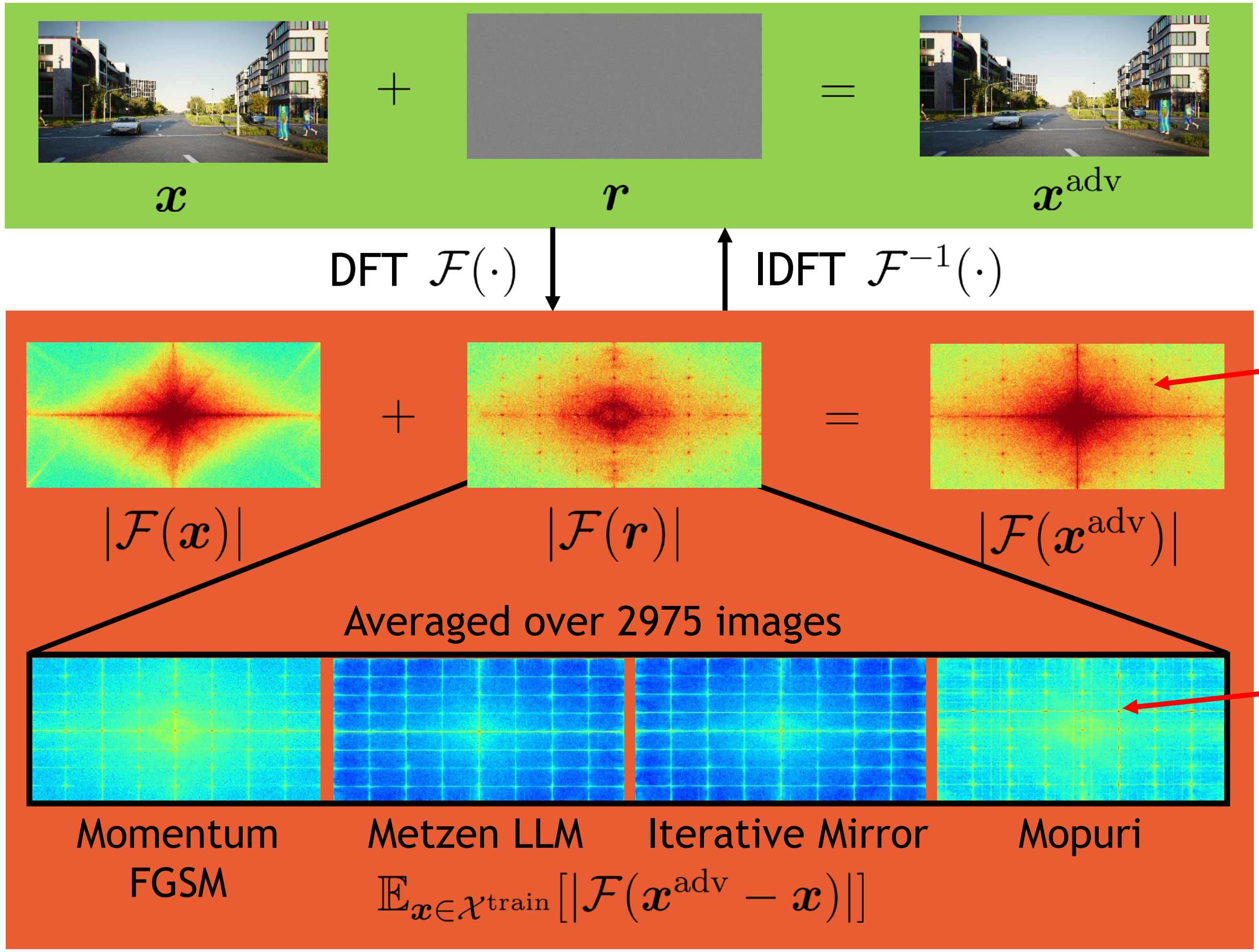
Our Approach: Identify, Measure and & Counteract „DNN-specific Safety Concerns" via AugMix

Adressed Safety Concern:
Brittleness of DNNs
Corruption Robustness



Our Approach: Identify, Measure and & Counteract „DNN-specific Safety Concerns"

Adressed Safety Concern:
Brittleness of DNNs
Adversarial Attacks



Adversarial examples are imperceptible in the spatial domain

Strong visible artifacts in the frequency domain

These artifacts are image-type and attack-type independent

Spatial domain
Frequency domain

From a Fourier-Domain Perspective on Adversarial Examples to a Wiener Filter Defense for Semantic Segmentation, N. Kapoor et al. *https://arxiv.org/abs/2012.01558*

Our Approach: Identify, Measure and & Counteract „DNN-specific Safety Concerns” via Wiener Filters

Adressed Safety Concern:
Brittleness of DNNs
Adversarial Attacks

Wiener Filters (WF) as an online denoising module

Steps:

1. Convert input image to DFT domain.
2. Apply pre-computed WF as a multiplicative filter.
3. Convert to spatial domain using IDFT.
4. Feed image to target DNN.

Attacked or clean image

Convert to **frequency** domain

Denoising Module

Input Pre-processor

Convert to **spatial** domain

Target DNN

Perception DNN

Robust Segmentation

From a Fourier-Domain Perspective on Adversarial Examples to a Wiener Filter Defense for Semantic Segmentation, N. Kapoor et al. *https://arxiv.org/abs/2012.01558*

Our Approach: Explore Mechanisms!

- Heatmap-based Attention Consistency Validation
- Mixture of Experts
- Domain Randomization in Optimized Dataset Selection
- MC Dropout
- Uncertainties For Anomaly Detection
- Hybrid Learning using Concept Enforcement
- Active Learning
- Adverserial Training
- Hybrid and robustness-focussed Compression
- …

Approx 80 Mechanisms are developed and evaluated

Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety

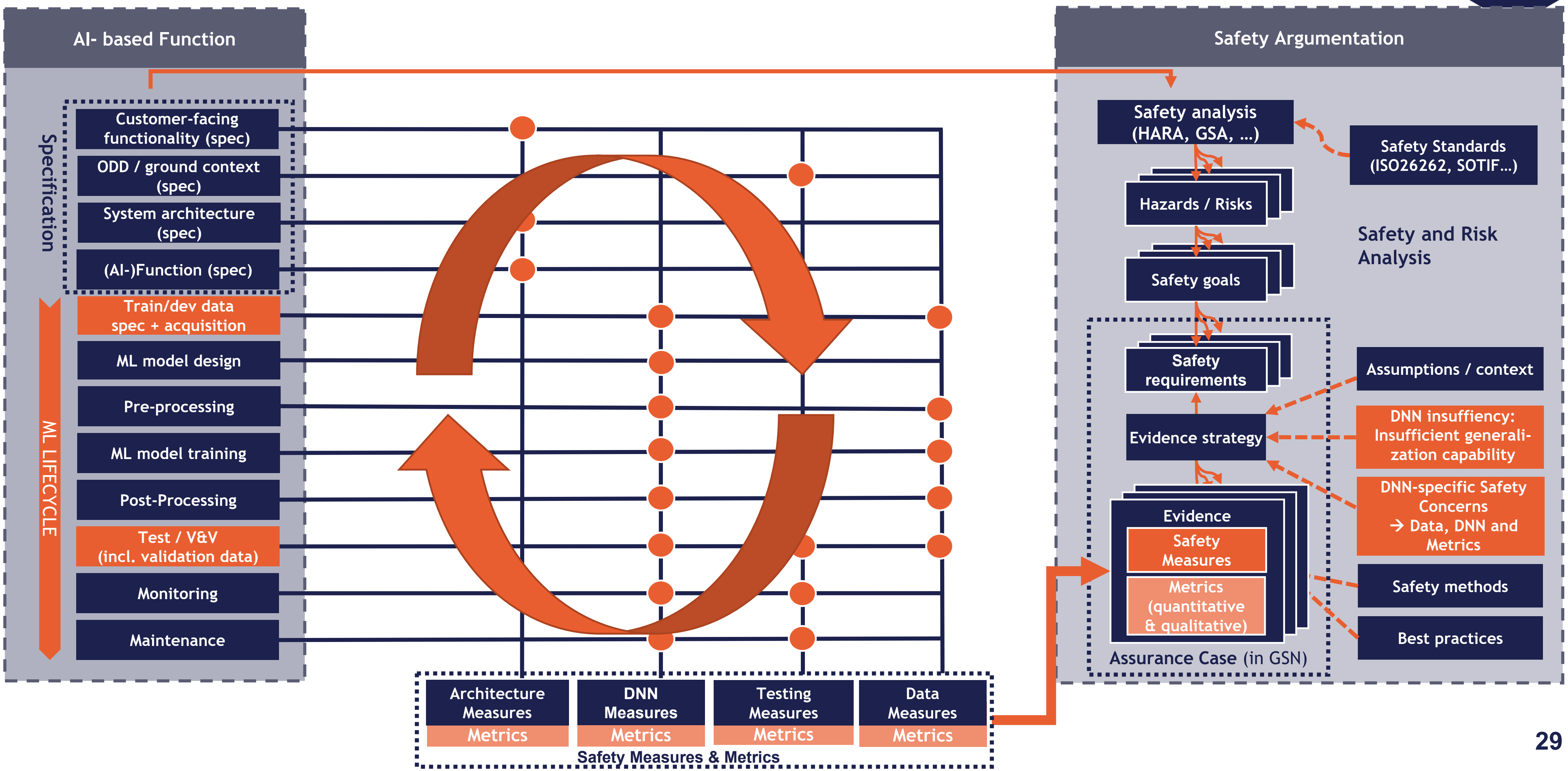
Sebastian Houben[1], Stephanie Abrecht[2], Maram Akila[1], Andreas Bär[15], Felix Brockherde[10], Patrick Feifel[8], Tim Fingscheidt[15], Sujan Sai Gannamaneni[1], Seyed Eghbal Ghobadi[8], Ahmed Hammam[8], Anselm Haselhoff[9], Felix Hauser[11], Christian Heinzemann[2], Marco Hoffmann[16], Nikhil Kapoor[7], Falk Kappel[13], Marvin Klingner[15], Jan Kronenberger[9], Fabian Küppers[9], Jonas Löhdefink[15], Michael Mlynarski[16], Michael Mock[1], Firas Mualla[13], Svetlana Pavlitskaya[14], Maximilian Poretschkin[1], Alexander Pohl[16], Varun Ravi-Kumar[4], Julia Rosenzweig[1], Matthias Rottmann[5], Stefan Rüping[1], Timo Sämann[4], Jan David Schneider[7], Elena Schulz[1], Gesina Schwalbe[3], Joachim Sicking[1], Toshika Srivastava[12], Serin Varghese[7], Michael Weber[14], Sebastian Wirkert[6], Tim Wirtz[1], and Matthias Woehrle[2]

[1] *Fraunhofer Institute for Intelligent Analysis and Information Systems*
[2] *Robert Bosch GmbH*
[3] *Continental AG*
[4] *Valeo S.A.*
[5] *University of Wuppertal*
[6] *Bayerische Motorenwerke AG*
[7] *Volkswagen AG*
[8] *Opel Automobile GmbH*
[9] *Hochschule Ruhr West*
[10] *umlaut AG*
[11] *Karlsruhe Institute of Technology*
[12] *Audi AG*
[13] *ZF Friedrichshafen AG*
[14] *FZI Research Center for Information Technology*
[15] *Technische Universität Braunschweig*
[16] *QualityMinds GmbH*

Survey: available at
https://www.ki-absicherung-projekt.de/

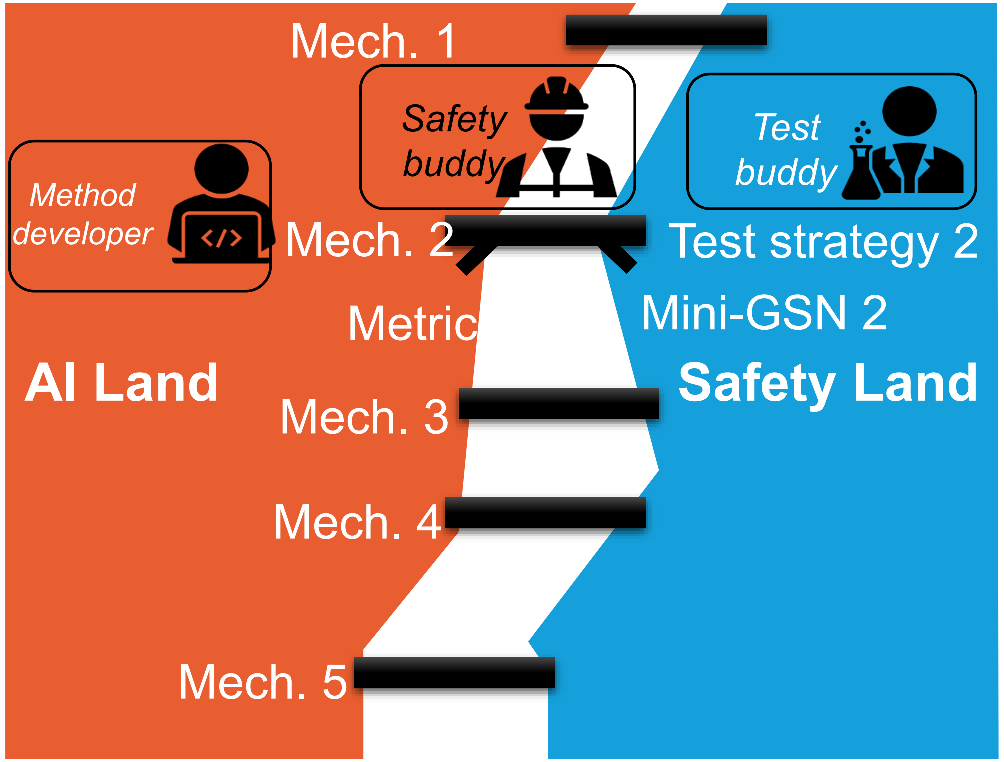
M. Mock et al.: An Integrated Approach to a Safety Argumentation for AI-based Perception Functions in Automated Driving, WAISE @SafeCOMP 2021)

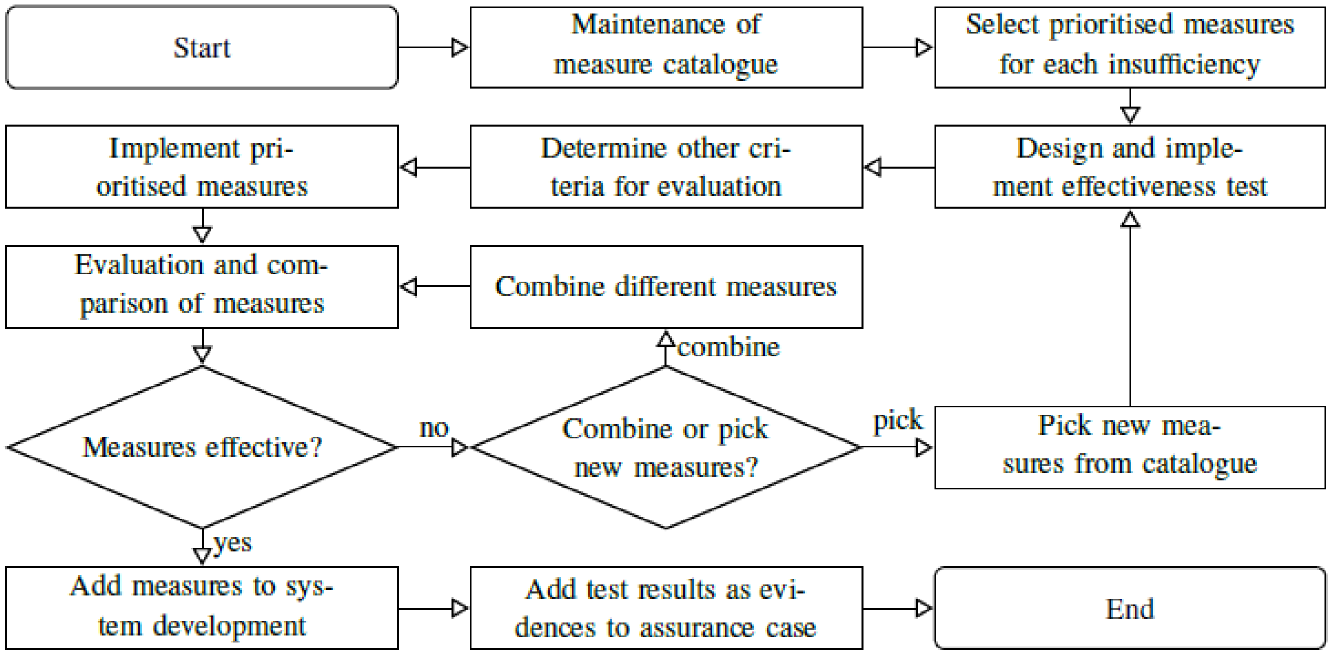
Our Approach: Summary



Our Approach: Evidence Workstreams

Empowering experts from safety engineering and ML to produce measures and evidences





S. Burton et al (2022): Evidences for the Assurance Case, to be published

Agenda

1. Introduction – CARIAD
2. DNNs and Safety in Automated Driving
3. KI-Absicherung Project & Approach
4. **Summary**

Summary

Findings & Consequences

- Safe AI is a central challenge for highly automated driving
- KI-Absicherung provides an approach for Safe AI
- Approach may serve as template for the industry and beyond
- Deep integration of AI-specifics into development PMT is necessary (continuous assurance of AI)

Contact:

Fabian Hüger

Artificial Intelligence Safety
@Volkswagen CARIAD

Contact: **fabian.hueger@volkswagen.de**

https://scholar.google.de/citations?user=ISPOi1UAAAAJ

www.ki-absicherung-projekt.de **@KI_Familie** **KI Familie**

Thank you!

QUESTIONS?