KI ABSICHERUNG

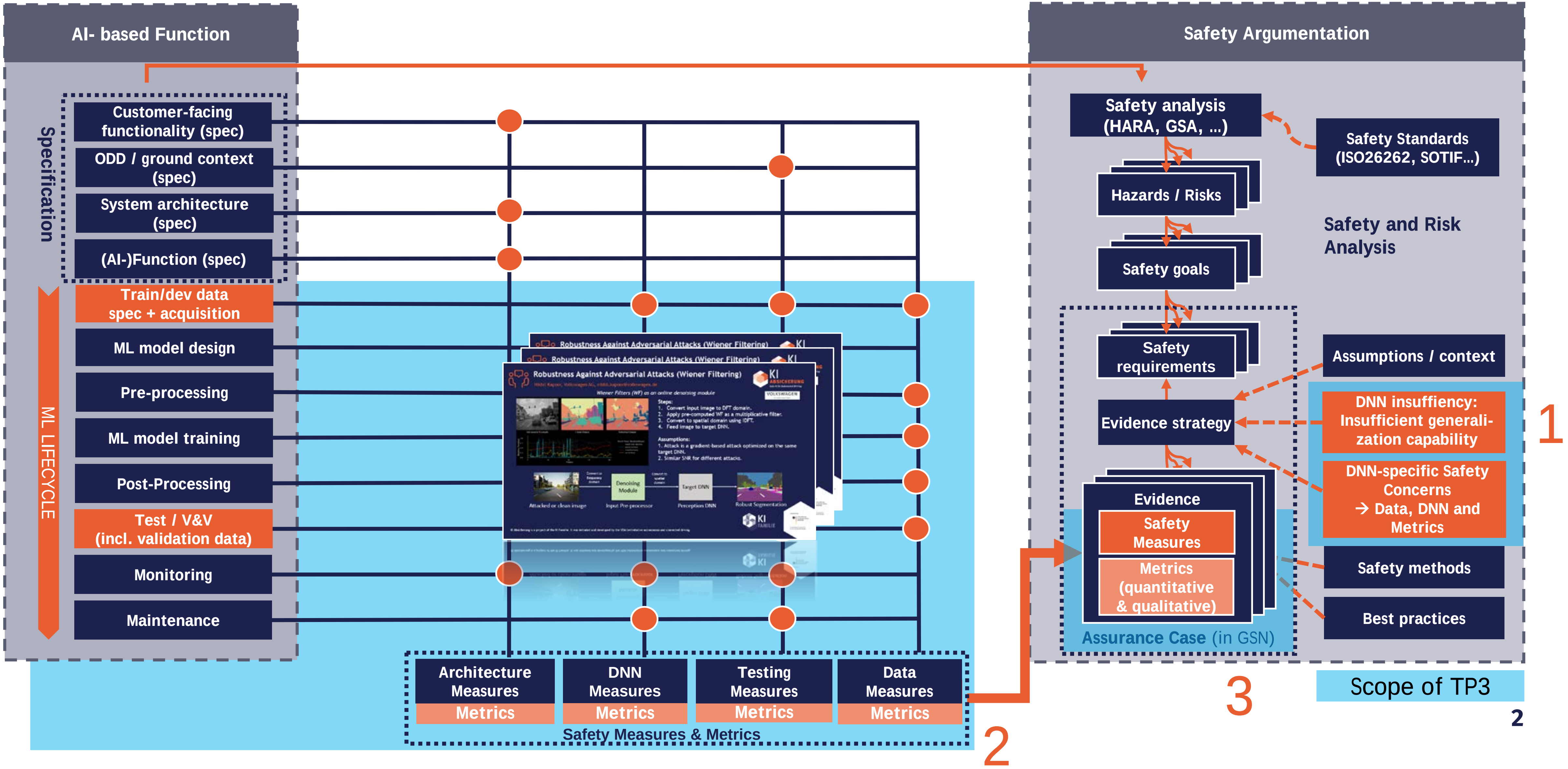
Safe AI for Automated Driving

11th March 2021, Online, Interim Presentation

Developing and evaluating measures and methods for the verification of the AI function

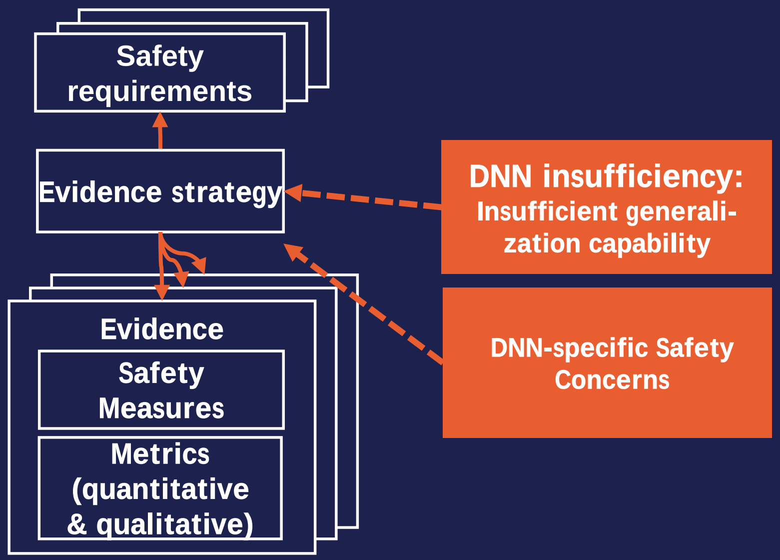
Dr. Fabian Hüger, Volkswagen AG

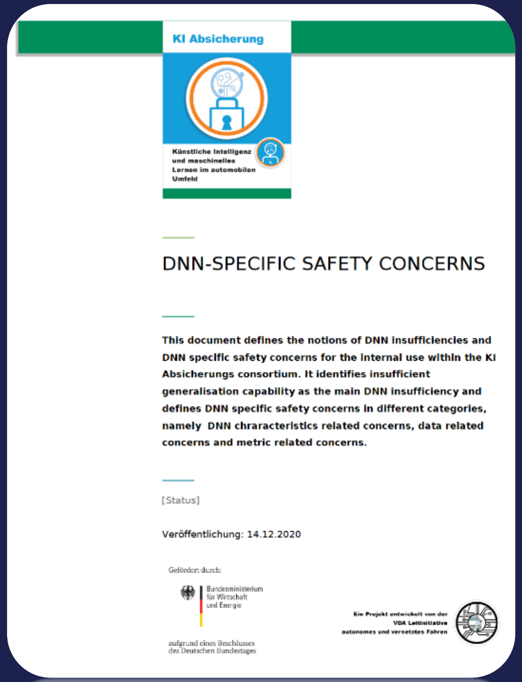
Methods and Measures in context of the KI Absicherung Big Picture



1

DNN-specific Safety Concerns





KI Absicherung

Künstliche Intelligenz und maschinelles Lernen im automobilen Umfeld

DNN-SPECIFIC SAFETY CONCERNS

This document defines the notions of DNN insufficiencies and DNN specific safety concerns for the internal use within the KI Absicherungs consortium. It identifies insufficient generalisation capability as the main DNN insufficiency and defines DNN specific safety concerns in different categories, namely DNN chraracteristics related concerns, data related concerns and metric related concerns.

[Status]

Veröffentlichung: 14.12.2020

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

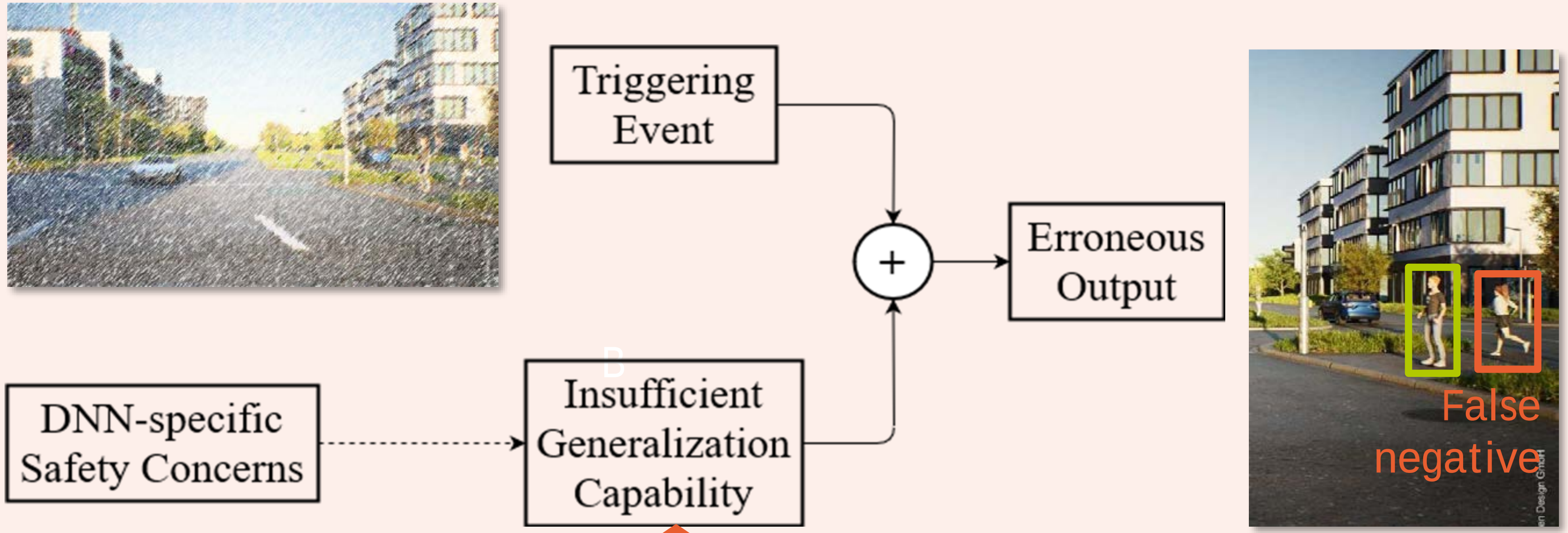
aufgrund eines Beschlusses des Deutschen Bundestages

Ein Projekt entwickelt von der VDA Leitinitiative autonomes und vernetztes Fahren

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



Functional Insufficiency according to ISO/PAS 21448

Willers O, Sudholt S, Raafatnia S, Abrecht S (2020) Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks. SAFECOMP 2020 workshop: WAISE 2020

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-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-1.1	**UNRELIABLE CONFIDENCE INFORMATION** DNNs tend to be overconfident in their predictions under certain conditions or in general outputting unreliable confidence information.	**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-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-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-1.2.1	**LACK OF TEMPORAL STABILITY** Detection results rapidly changing in time whereas little change occurs in the ground truth	**SC-2.4**	**SPECIFICATION OF THE ODD** An incomplete or incorrect ODD specification leads to incomplete data records for training and testing.
SC-1.3	**INCOMPREHENSIBLE BEHAVIOUR** Inability to explain exactly how DNNs come to a decision.	**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-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-2.6**	**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-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-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.

Work-in-progress

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**

Functional Insufficiencies

DNN-characteristics-related concerns

Data-related concerns

Metric-related concerns

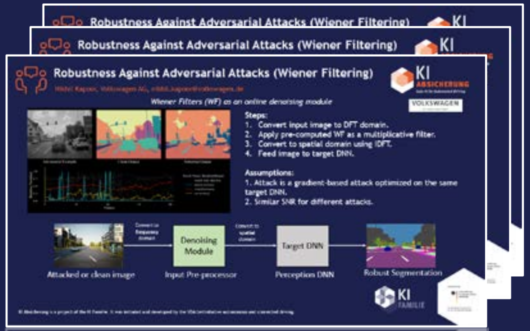
DNN-specific Safety Concerns

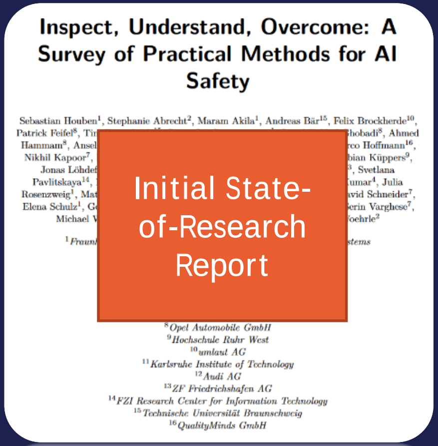
2

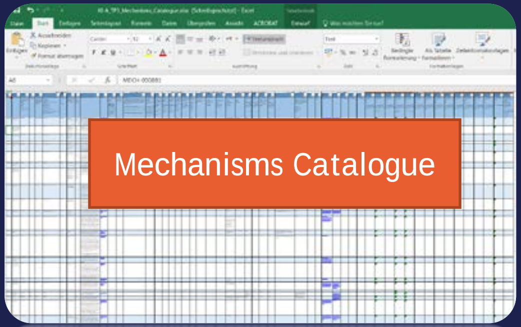
Exemplary Methods and Measures

Architecture Measures	DNN Measures	Testing Measures	Data Measures
Metrics	Metrics	Metrics	Metrics

Safety Measures & Metrics







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

Entropy Maximization and Meta Classification for Out-of-Distribution Detection in Semantic Segmentation

Addressed Safety Concerns: Unreliable confidence information

Enforce segmentation networks to output high prediction uncertainty on **Out-of-Distribution inputs** by means of a modified loss function

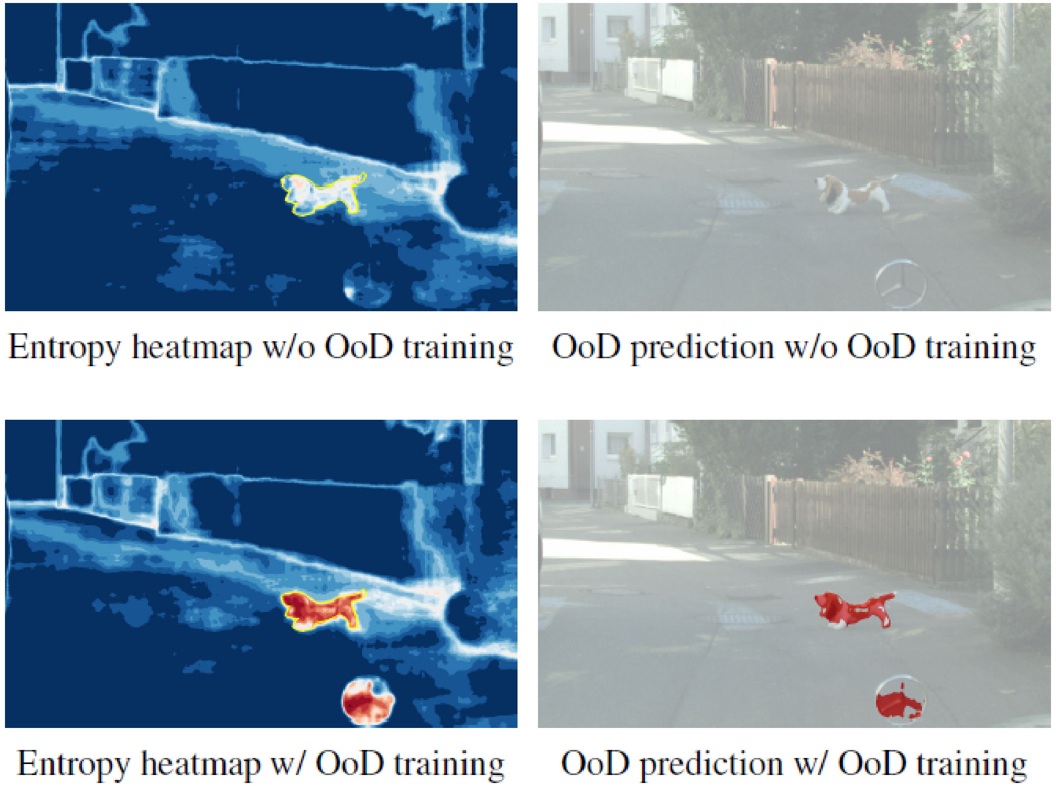


Figure 2: Comparison of softmax entropy heatmap and OoD prediction mask with our OoD training (*top row*) and without (*bottom row*). The yellow lines in the entropy heatmaps mark the annotation of the OoD object. The OoD object prediction is obtained by simply thresholding on the entropy heatmap (in this example at $t = 0.7$ yielding the red pixels in the OoD prediction masks).

Entropy Maximization and Meta Classification for Out-Of-Distribution Detection in Semantic Segmentation, R Chan et al.,arXiv preprint arXiv:2012.06575, 2020

Object Detection Uncertainty based on Gradient Information

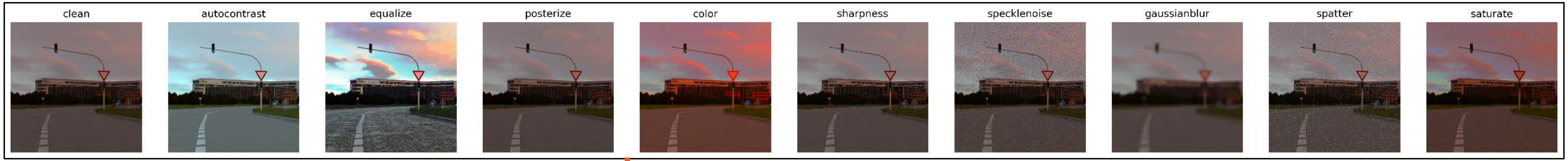
Addressed Safety Concerns: **Unreliable confidence information**

Tackling overconfidence via **novel online uncertainty mechanism** using gradient information

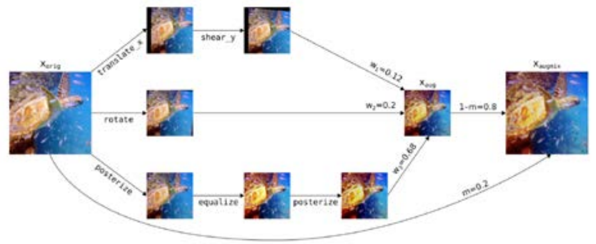
False Prediction at 0.7 confidence

Augmentation Training (AugMix)

Addressed Safety Concerns: Brittleness of DNNs



Combined using AugMix



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

Dan Hendrycks*
DeepMind
hendrycks@berkeley.edu

Norman Mu*
Google
normanmu@google.com

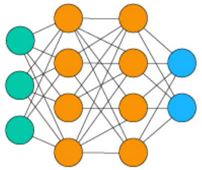
Ekin D. Cubuk
Google
cubuk@google.com

Barret Zoph
Google
barretzoph@google.com

Justin Gilmer
Google
gilmer@google.com

Balaji Lakshminarayanan†
DeepMind
balajiln@google.com

Training



DeepLabv3
ResNet 101
(KIA model by Intel)

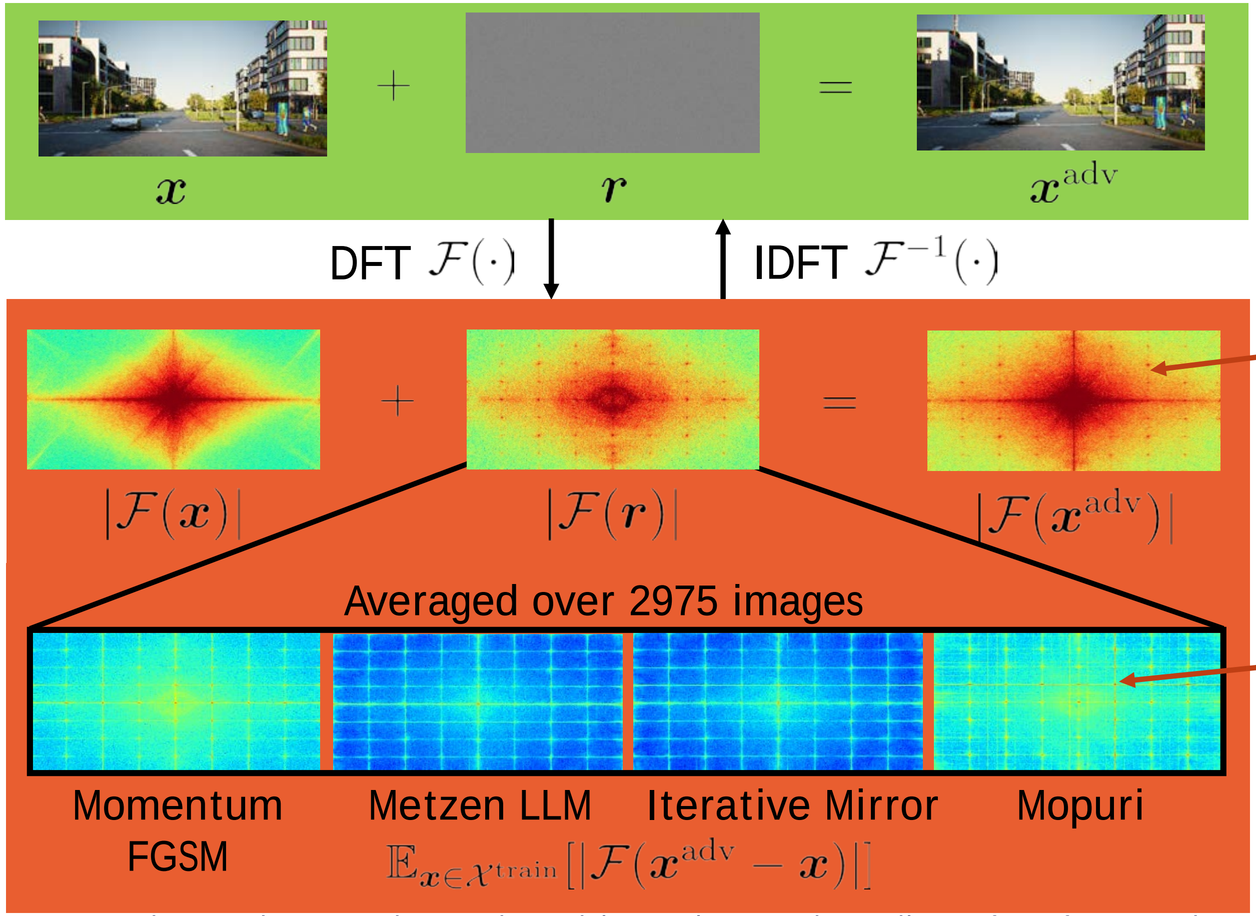
Evaluation on ***unseen*** **„real-world“** corruptions

+ Improved robustness
+ Improved generalization
+ Data efficient augmentation strategy

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

Wiener Filters (WF) as an online denoising module

Addressed Safety Concerns:
Brittleness of DNNs



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

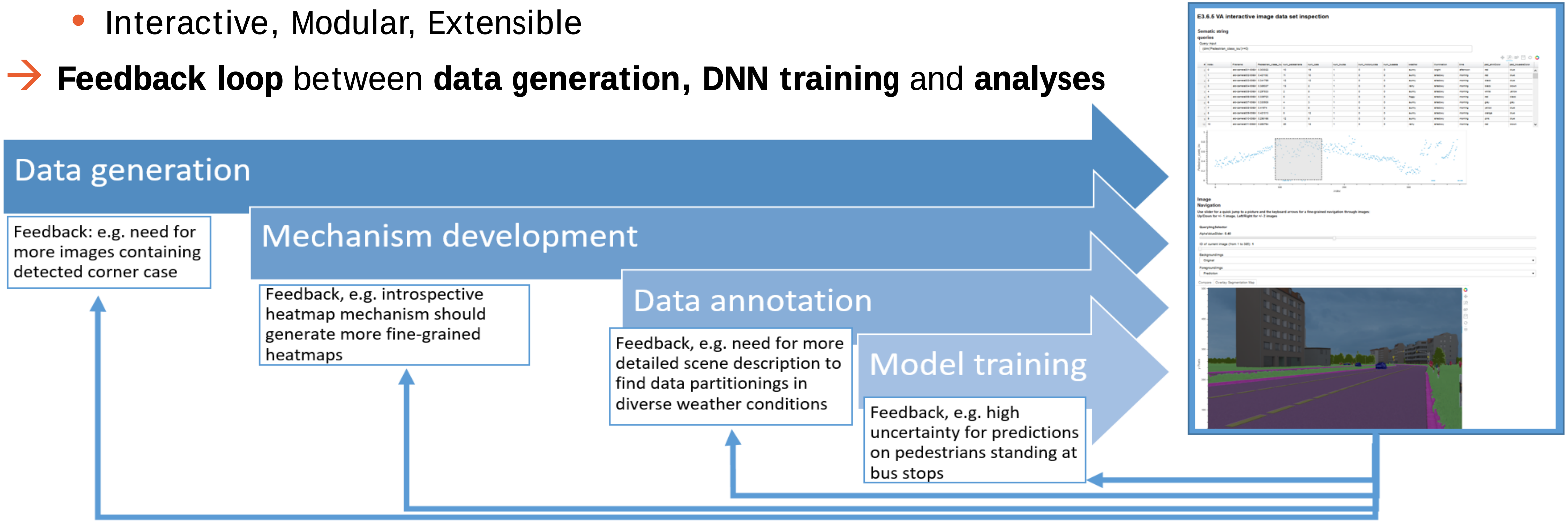
From a Fourier-Domain Perspective on Adversarial Examples to a Wiener Filter Defense for Semantic Segmentation, N. Kapoor et al. *https://arxiv.org/abs/2012.01558*

Semantic Analysis of DNN Predictions with Visual Analytics

Addressed Safety Concerns: Incomprehensible Behavior

- Development of a **visual interactive interface**
  - Inspection of **DNN predictions** and **data sets** w.r.t. pre-computed **meta data (semantics)**
  - Interactive, Modular, Extensible

→ **Feedback loop** between **data generation, DNN training** and **analyses**



Heatmap-based Attention Consistency Validation

Adressed Safety Concerns:
Insufficient Plausibility

Detection of implausibilities between detections and attention

Further exemplary mechanisms

- Mixture of Experts
- Domain Randomization in Optimized Dataset Selection
- MC Dropout
- Uncertainties For Anomaly Detection
- Hybrid Learning using Concept Enforcement
- Active Learning
- Adversarial Training
- Hybrid and robustness-focused Compression
- ...

3

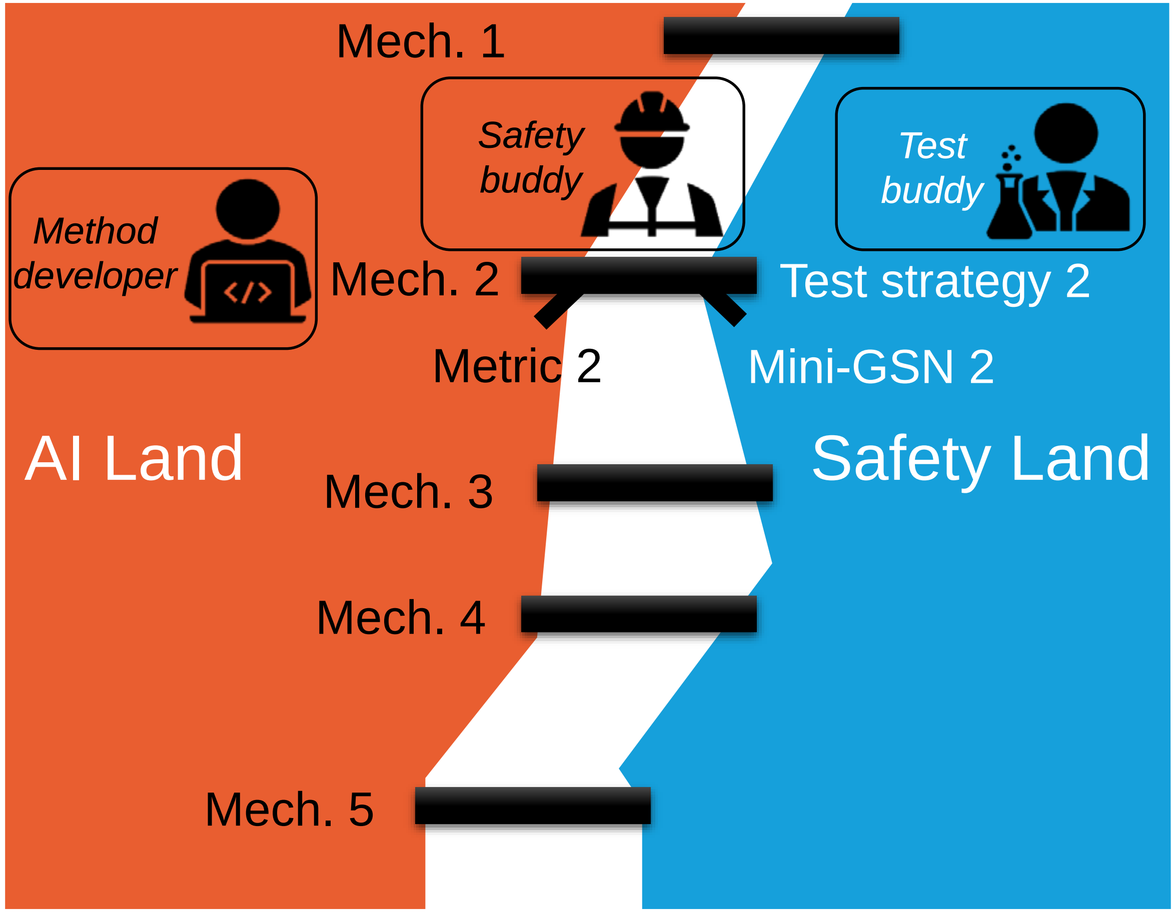
Injecting Mechanisms into the Safety Argumentation: Evidence Workshops

Developing and evaluating measures and methods for the verification of the AI function

Evidence workshops from P4

How to build the **big bridge** between
AI Land and Safety Land?

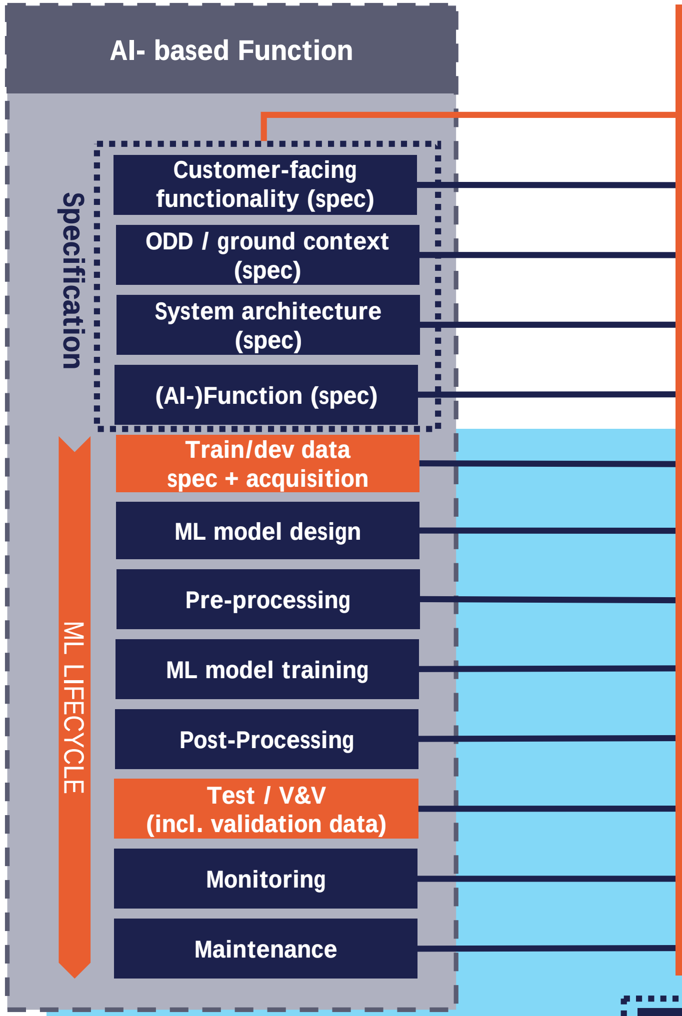
Evidence workshops were conducted to streamline and integrate the mechanisms into the safety argumentation in TP4



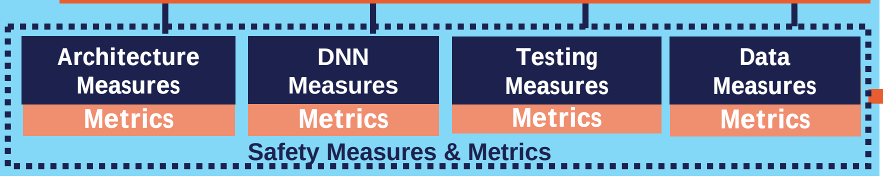
4

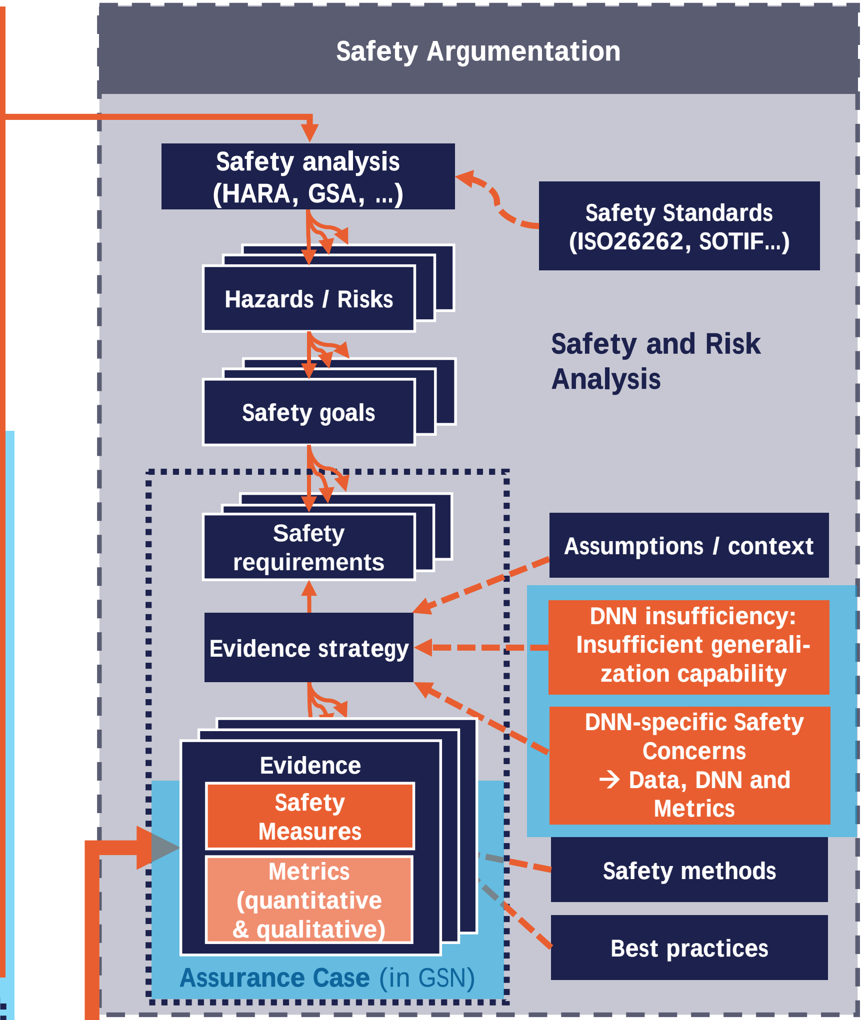
Summary & Outlook

Summary and Outlook



- Identification of DNN-specific Safety Concerns
- Development of Methods and Measures (Mechanisms)
- Comparative Benchmarks
- Combinations of Mechanisms
- Mechanisms for 3D Object Detection
- Evaluation of the Mechanisms w.r.t. Safety Requirements and Concerns
- Application of the Mechanisms in the Assurance Case





Scope of TP3

SAIAD Workshop 2021

https://sites.google.com/view/saiad2021

Submission Deadline: March 15, 2021, Anywhere on Earth (UTC-12)

KI ABSICHERUNG

Safe AI for Automated Driving

Dr. Fabian Hüger, Volkswagen AG

fabian.hueger@volkswagen.de

KI Absicherung ist ein Projekt der KI Familie und wurde aus der VDA Leitinitiative autonomes und vernetztes Fahren heraus entwickelt.

KI FAMILIE

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

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