A Modern Application – Autonomous Driving
AI-Based Pedestrian Detection

- Setup
  - An *autonomously operating vehicle* ...
  - ... is crossing an *intersection*

- AI functionality for detecting pedestrians
  - *Camera images* processed by CNN

- Output
  - Segmentation mask
  - Bounding box detections

---

Synthetically generated intersection and corresponding semantic segmentation
Project KI Absicherung - https://www.ki-absicherung.vdali.de
Can I Use a State-of-the-Art ML Model in an Automated Vehicle?

- Well, ML models (safety concerns)
  - don't work well on unseen data
  - are not robust to domain changes
  - may overfit to irrelevant correlations
  - are overconfident in their predictions

Syntetically generated intersection and corresponding semantic segmentation
Project KI Absicherung - https://www.ki-absicherung.vdali.de
Can I Use a State-of-the-Art ML Model in an Automated Vehicle?

- What can be done?
  - There are many attempts / research directions to alleviate these concerns.
  - One of our contributions is to investigate some of these methods more deeply.
  - Another one is to evaluate them w.r.t. to the safety concerns
  - and find a plausible argumentation that they are circumvented or kept at bay.
KI Absicherung is making the safety of Al-based function modules for highly automated driving verifiable.
The Project „KI Absicherung – Safe AI for Automated Driving“

Consortium lead: Volkswagen AG
Fraunhofer IAIS

Deputy consortium lead and scientific coordination:
Volkswagen AG

Budget: 41 Mio. €
Funding: 19.2 Mio. €

Project duration: 36 months
2019/01/07 - 2022/06/20

25 partners

Volkswagen AG | Wolfsburg
DLR | Braunschweig
Bergische Universität | Wuppertal
Fraunhofer IAIS | Sankt Augustin
Continental Automotive GmbH | Babenhausen
Opel Automobile GmbH | Rüsselsheim
HDI | Heidelberg
DFKI | Kaiserslautern
FZI | Karlsruhe (ext. Technologiepartner)
MackeVision Medien Design GmbH | Stuttgart
MackeVision | Abstatt
Robert Bosch GmbH | Abstatt
Hella Aglaia Mobile Vision GmbH | Berlin
Merantix GmbH | Berlin (ext. Technologiepartner)
Neurocat GmbH | Berlin (ext. Technologiepartner)
EICT GmbH | Berlin (ext. Partner Projektmanagement)
Valo Schalter und Sensoren GmbH | Kronach
QualityMinds GmbH | Nürnberg
ASTech GmbH | Gaimersheim
Audi AG | Ingolstadt
ELEKTRONISCHE FAHRWERSYSTEME GMBH* | Gaimersheim
BMW AG | München
Intel Deutschland GmbH | Neubiberg
Technische Universität | München
Fraunhofer IKS * | München
(Ext. Technology Solutions | München
(Ext. Technology Partner)

Consortium Lead  ⚫ OEMs  ⚫ Tier-1  ⚫ Technology provider  ⚫ Research  ⚫ External Partner  * In preparation
KI Absicherung
Main Goals

1. Methods for training and testing of AI-based functions

KI Absicherung develops and investigates means and methods for verifying AI-based functions for highly automated driving.

2. Safety argumentation

For the pedestrian detection use case, the project is developing an exemplary safety argumentation and methods for verifying a complex AI function.

3. Communication with standardization bodies on AI certification

The project’s results will be used in the exchange with standardization bodies to support the development of a standard for safeguarding AI-based function modules.
# KI Absicherung

## Main Goals

### Today’s Focus

<table>
<thead>
<tr>
<th>1. Methods for training and testing of AI-based functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>KI Absicherung develops and investigates means and methods for verifying AI-based functions for highly automated driving.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Safety argumentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>For the pedestrian detection use case, the project is developing an exemplary safety argumentation and methods for verifying a complex AI function.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Communication with standardization bodies on AI certification</th>
</tr>
</thead>
<tbody>
<tr>
<td>The project’s results will be used in the exchange with standardization bodies to support the development of a standard for safeguarding AI-based function modules.</td>
</tr>
</tbody>
</table>
From a Data-Driven AI Function to an Assurance Case

**Use Case: Pedestrian Detection**

- Process-related generation of synthetic learning, testing and validation data.

- Development of measures and methods that improve the AI function over a wide array of metrics.

- Development and validation of testing methods for these metrics.

- Stringent argumentation for the AI function and its Operational Design Domain (ODD).
From a Data-Driven AI Function to an Assurance Case

Use Case: Pedestrian Detection

- Process-related generation of synthetic learning, testing and validation data.
- Development of measures and methods that improve the AI function over a wide array of metrics.
- Development and validation of testing methods for these metrics.
- Stringent argumentation for the AI function and its Operational Design Domain (ODD).
Realistic Uncertainty Estimation
Know If You Know Nothing

- Local Uncertainty Estimation allows for a self-assessment of the neural network given its input, e.g., in order to detect out-of-distribution inputs.
- Increase safety by discarding uncertain predictions.
- Optimize your dataset by identifying data points with high uncertainty.
Realistic Uncertainty Estimation
Know If You Know Nothing

- **State-of-the Art:** Bayesian Networks, Deep Ensembles, MC Dropout
  - **Poorly calibrated:** Predictions are corrected by post processing
  - Yet, realistic local uncertainties are of minor quality

- **Our approach:**
  - Modify the loss function to provide realistic MC Dropout uncertainties
  - Formal understanding and proofs for uncertainty estimation in MC Dropout networks and deep ensembles
Teacher-Student-Methods
Gain Insight into the Inner Workings of a Neural Network

- Derive interpretable model (student) from a given black-box-model (teacher)
- Identify erroneous "explanations"
- Does the teacher suffer from the same problem?
- Enables analyses of the teacher model

Teacher-Student-Methods
Exploit Identified Insights

- Derive interpretable model (student) from a given black-box-model (teacher)
- Identify erroneous "explanations"
- Does the teacher suffer from the same problem?
- Enables analyses of the teacher model
- By these means we can construct semantic attacks

Binary classification: Is there a car in this image?

Student model considers traffic beacon an important hint
Semantic attack: Add traffic beacons to an image

Student model predicts a car in this image
Teacher model (a ResNet) does so, too.
Assess Test Data Completeness
Find Situations with Systematically Low Performance

- Find correlations among **semantic concepts** (e.g., position / size of pedestrians)
  - and poor model performance
  - or pronounced and distinct safety concerns

- Reveal situations with poor prediction performance

- Reveal poor training procedures
  - E.g., with **Neuron Coverage** (Percentage of neurons that are sufficiently activated by at least one test example)

---

Evaluation of Dependencies between Neural Networks and Data
Visual Interactive Analysis of Semantic Features

- **Goal**: finding correlated insufficiencies and gaining insight into the decision of networks

- Understanding semantic concepts of the data is the key to identifying & distinguishing outliers from systematic weaknesses (like shortcuts or data flaws)

- But: automated analysis of semantics is difficult

- Those **semantic features** are examined best visually by humans
Evaluation of Dependencies between Neural Networks and Data
Finding Semantic Clusters in a Visual Interactive Interface

- Specific focus on enabling the human expert to:
- Interactively analyze the **KPIs w.r.t. robustness**
- Inspect image data sets to **gain insights**
  - E.g. into important image parts, hard or underrepresented images/image scenes ("corner cases"), unusual object appearances, data flaws etc.
Enabling the human to **understand semantic concepts of the data** with additional information
- E.g. metadata, histogram data

**Identify semantic clusters**
- Use VA to develop metrics incorporating **human semantic understanding** and DNN performance measures
- E.g. by textual and visual querying (“query by example”) and filtering
- E.g. tagging, sorting and searching images
These insights can then in turn be used to enhance
- The training methods of the neural networks
- The data set generation
- Establishing a feedback loop between data generation, neural network training and analyses of both
Beyond Absicherung
Typical Challenges of the Individual Audit Areas

- **Ethics & Law**: Key questions concerning ethical issues
- **Fairness**: Historically unbalanced data
- **Autonomy & Control**: Appropriate degree of autonomy
- **Transparency**: Incomprehensibility of results from neural networks

**Today’s Focus**

- **Reliability**: Robustness of results processed by AI-systems
- **Safety & Security**: Safety risks due to probabilistic output from AI component
- **Privacy**: New types of personal data through AI
Certifying Artificial Intelligence
Whitepaper Points out Audit Areas

- Collaboration of experts from Fraunhofer IAIS, Univ. Bonn and Univ. Cologne from the fields of
  - Machine Learning
  - Law
  - Ethics
  - IT Security
- Interdisciplinary initiative funded by the competence platform KI.NRW
- Audit areas for trustworthy AI
- www.iais.fraunhofer.de/ki-zertifizierung

Publication with high international attention
Advanced Trainings
Data Scientist Specialized in Trustworthy AI

- Advanced training offered by Fraunhofer IAIS
  „Data Scientist Specialized in Trustworthy AI“
  - Audit areas of trustworthy AI
  - Methods for assessing and verifying AI applications
- Project „KI-Absicherung“
  VDA Leitinitiative
  “Autonomous and Connected Driving“
  www.ki-absicherung.vdali.de
- Point of contact: PD Dr. Michael Mock
  michael.mock@iais.fraunhofer.de

www.bigdata.fraunhofer.de/datascientist
THANK YOU FOR YOUR ATTENTION

Dr. Sebastian Houben | Fraunhofer IAIS | sebastian.houben@iais.fraunhofer.de
The Connected Car and Autonomous Driving, October 26th, 2020
Disclaimer

Copyright © by
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.
Hansastraße 27 c, 80686 Munich, Germany

All rights reserved.

Responsible contact: Dr. Sebastian Houben, PD Dr. Michael Mock
E-mail: Sebastian.houben@iais.fraunhofer.de

All copyrights for this presentation and their content are owned in full by the Fraunhofer-Gesellschaft, unless expressly indicated otherwise.

Each presentation may be used for personal editorial purposes only. Modifications of images and text are not permitted. Any download or printed copy of this presentation material shall not be distributed or used for commercial purposes without prior consent of the Fraunhofer-Gesellschaft.

Notwithstanding the above mentioned, the presentation may only be used for reporting on Fraunhofer-Gesellschaft and its institutes free of charge provided source references to Fraunhofer’s copyright shall be included correctly and provided that two free copies of the publication shall be sent to the above mentioned address.

The Fraunhofer-Gesellschaft undertakes reasonable efforts to ensure that the contents of its presentations are accurate, complete and kept up to date. Nevertheless, the possibility of errors cannot be entirely ruled out. The Fraunhofer-Gesellschaft does not take any warranty in respect of the timeliness, accuracy or completeness of material published in its presentations, and disclaims all liability for (material or non-material) loss or damage arising from the use of content obtained from the presentations. The afore mentioned disclaimer includes damages of third parties.

Registered trademarks, names, and copyrighted text and images are not generally indicated as such in the presentations of the Fraunhofer-Gesellschaft. However, the absence of such indications in no way implies that these names, images or text belong to the public domain and may be used unrestrictedly with regard to trademark or copyright law.