Fraunhofer IKS

Fraunhofer Institute for Cognitive Systems IKS

AISafety 2023 Workshop (IJCAI)

Iwo Kurzidem, Simon Burton, Philipp Schleiss

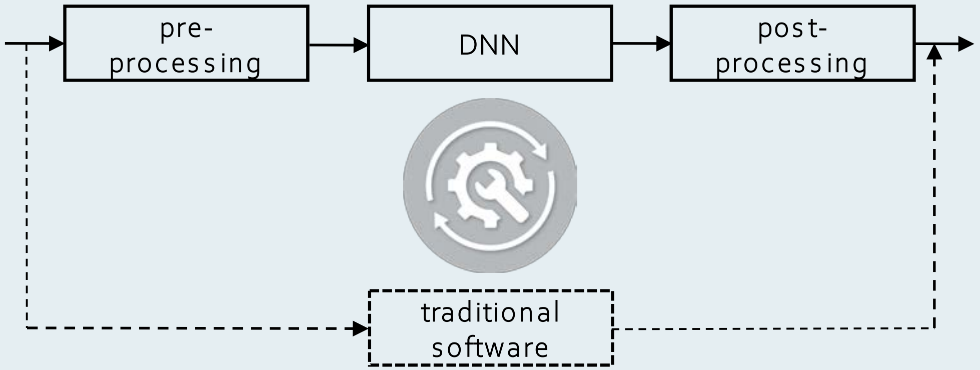
# AI for Safety: How to use Explainable Machine Learning Approaches for Safety Analyses

# ML for Safety: What is that?

## AI & Safety

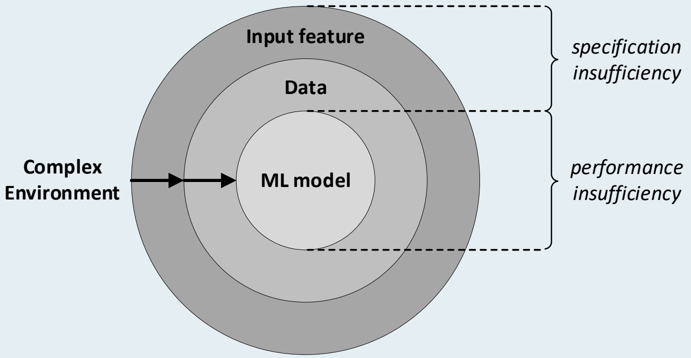
### Safety **of** ML

- Use **methods and measures to address** Machine Learning (ML) **insufficiencies** within the software (architecture)
- For instance: Uncertainty Calibration, Filtering, Robustification etc.




### ML **for** Safety

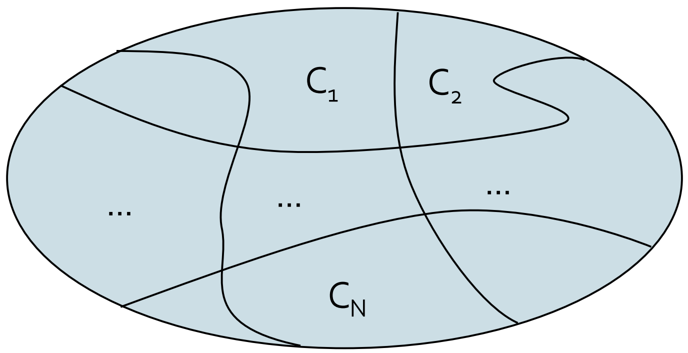
- Use (explainable) ML approaches **for safety analysis to identify** ML **insufficiencies** and link the **results to safety artifacts**
- For instance: Search-based Testing, Principal Component Analysis, Clustering etc.




Adapted from [1]

# Safety Artifacts: What kind?



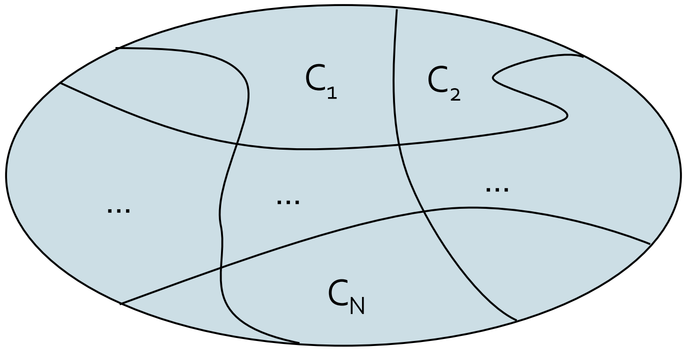

## Equivalence Classes of Equal Behavior [2]

*Definition: » [Equivalence] classes are identified based on the division of inputs and outputs, such that a representative test value can be selected for each [equivalence] class. «*

→ How is this useful for safety?
The identification and use of equivalence classes can **considerably reduce the required testing effort.**

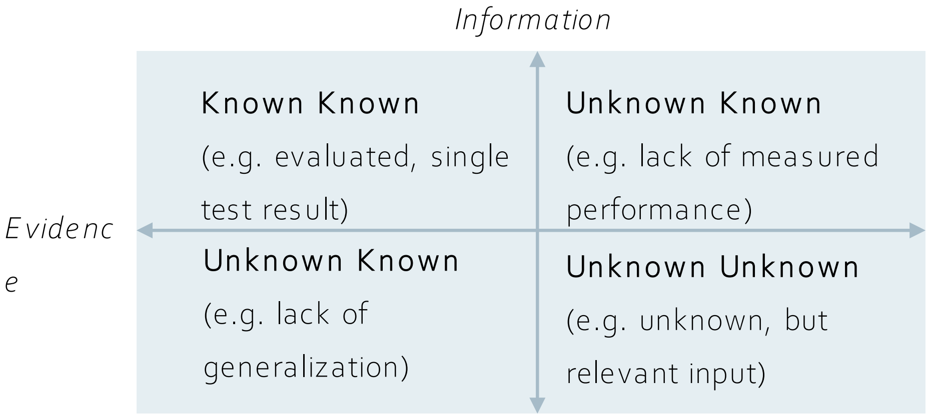
# Safety Artifacts: What kind?




## Equivalence Classes of Equal Behavior [2]

*Definition: » [Equivalence] classes are identified based on the division of inputs and outputs, such that a representative test value can be selected for each [equivalence] class. «*

→ How is this useful for safety?
The identification and use of equivalence classes can **considerably reduce the required testing effort.**




## Unknown Unknowns [3]

*Definition: » Unknown Unknowns are [...] known parameters of scenarios [that] can combine into unknown potential triggering conditions (e.g., combination of weather and traffic conditions). «*

→ How is this useful for safety?
The identification of unknown unknowns can potentially **reduce unsafe system behavior.**

# Safety Artifacts: How do we find them?

## Decision Trees (DTs): Mathematical Foundation

The basic concept of DTs is data partitioning according to:

Find highest *decrease in impurity* $\Delta(s,\ n)$ for data $S_n$ via

$$\Delta(s,n) = f_i(n) - \frac{S_{nL}}{S_n} * f_i(nL) - \frac{S_{nR}}{S_n} * f_i(nR),$$

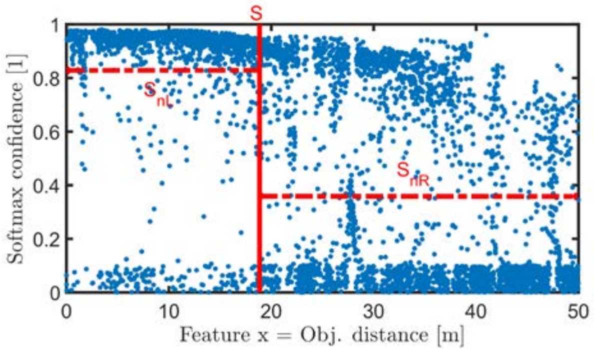
with *impurity function*

$$f_i(n) = \sum_{x,y\,\in\,L_n} (y_M - y_T)^2\,,$$

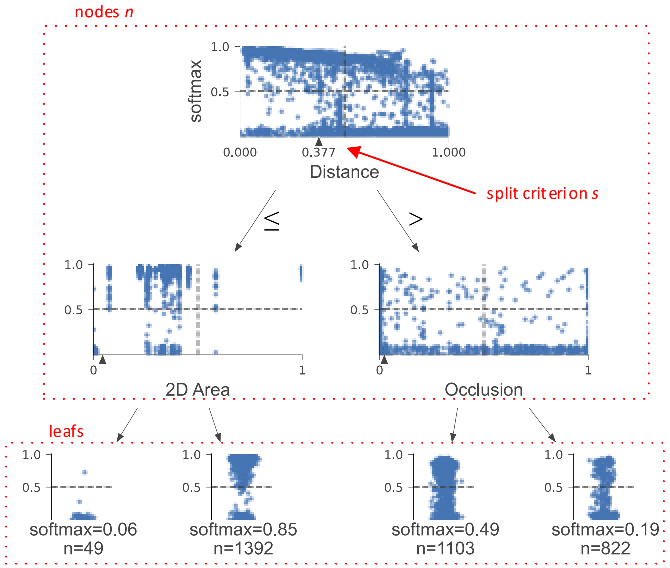
in order to **repeatedly partition the data into disjoint, smaller subsets**, such that each subset is **consistent with regards to its output**.

Hyperparameters:

- Threshold $\theta$ for the minimum decrease in impurity, i.e., $\Delta(s,\ n) < \theta$
- The minimum number of samples $S_{min}$ to allow further splits, i.e., $S_n > S_{min}$




*Example for single split of data*

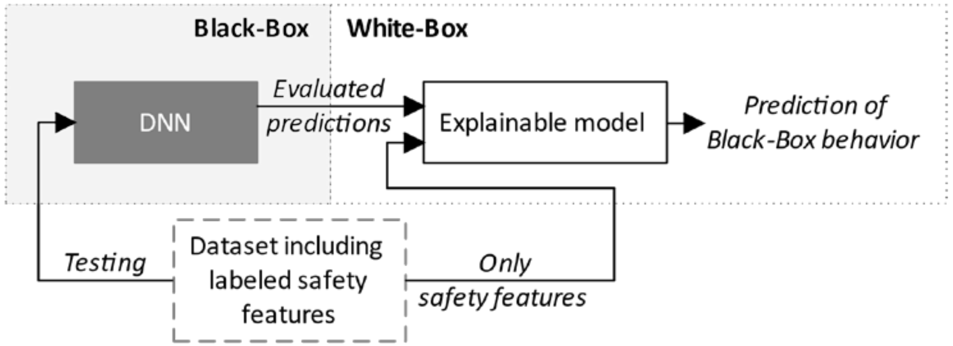



*Example for single decision tree*

# ML for Safety: How did we create the RF?

## Previous work [4]

Basic idea for safety assurance: **Build an introspective, explainable model** (so we understand *why* “something” is safe)

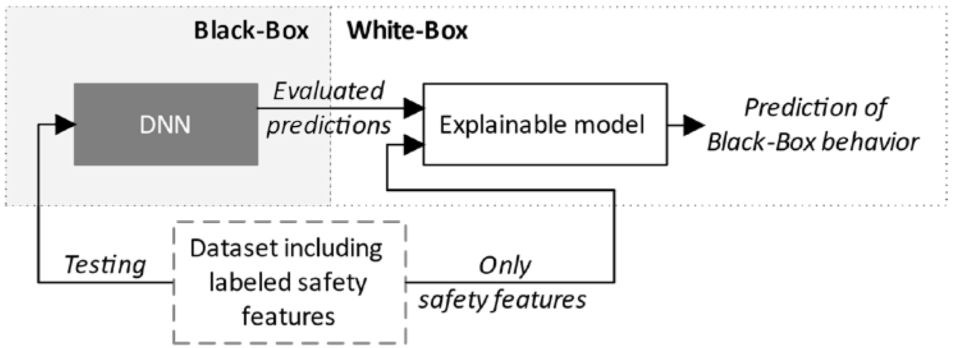


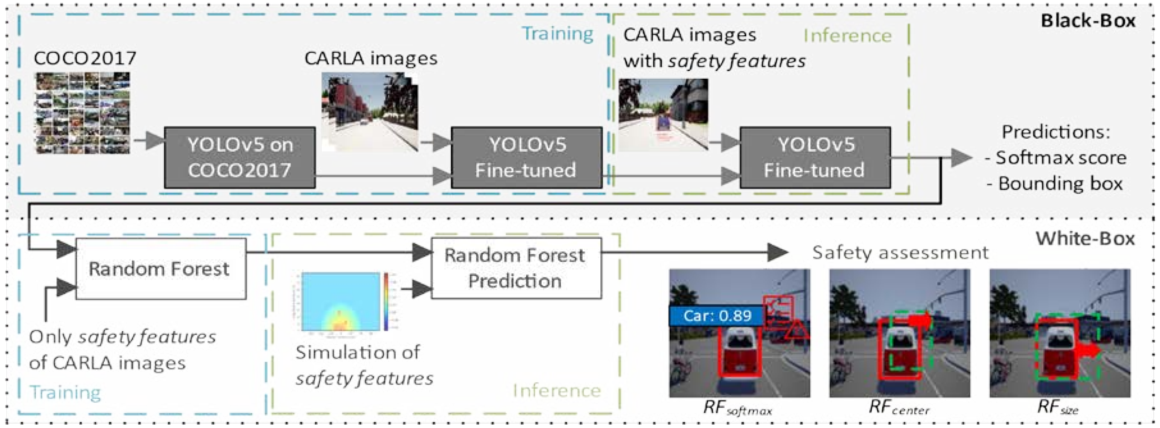

 

# ML for Safety: How did we create the RF?

## Previous work [4]

Basic idea for safety assurance: **Build an introspective, explainable model** (so we understand *why* "something" is safe)




ML components:

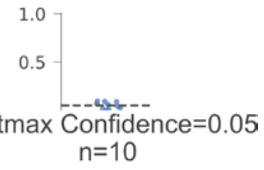
- **Black-Box**: A baseline YOLOv5 object detector is trained on COCO2017 data and fine-tuned on CARLA images.
- **White-Box**: A Random Forest (RF) is trained using the selected (safety) features and corresponding, evaluated YOLOv5 predictions.

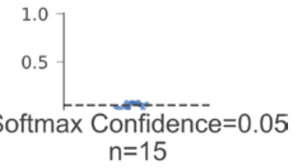
# ML for Safety: What did we find?

## Decision Tree Leaves

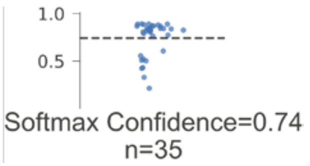
- 1. Leaves that show **little variance** in data and fulfill $S_n = S_{min}$
  Meaning: Desired result, best possible subset, given $\theta$ and $S_n$



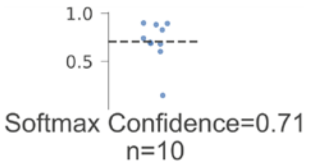

- 2. Leaves that show **little variance** in data and fulfill $S_n > S_{min}$
  Meaning: Early stopping, best possible subsets, $\Delta(s, n) < \theta$




- 3. Leaves that show **high variance** in data and fulfill $S_n > S_{min}$
  Meaning: Early stopping, inconsistent subsets, independent of $\theta$ and $S_n$




- 4. Leaves that show **high variance** in data and fulfill $S_n = S_{min}$
  Meaning: Impure result, prevent overfitting, given $\theta$ and $S_n$

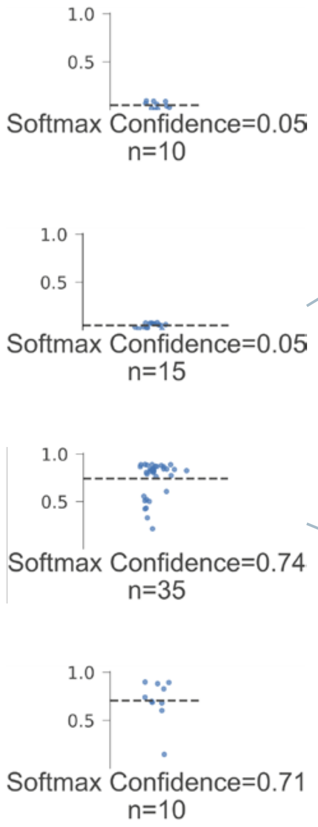


# Safety Artifacts: How did we find them?

## Decision Tree Leaves

- 1. Leaves that show **little variance** in data and fulfill $\boldsymbol{S_n = S_{min}}$
  Meaning: Desired result, best possible subset, given $\theta$ and $S_n$
- 2. Leaves that show **little variance** in data and fulfill $\boldsymbol{S_n > S_{min}}$
  Meaning: Early stopping, best possible subsets, $\Delta(s,n) < \theta$
- 3. Leaves that show **high variance** in data and fulfill $\boldsymbol{S_n > S_{min}}$
  Meaning: Early stopping, inconsistent subsets, independent of $\theta$ and $S_n$
- 4. Leaves that show **high variance** in data and fulfill $\boldsymbol{S_n = S_{min}}$
  Meaning: Impure result, prevent overfitting, given $\theta$ and $S_n$




### Safety Artifacts

Does this mean we found a general area of equivalent behavior, as the data „naturally" converges?

Does this mean the provided data does not allow a disentanglement with the contained information (given the inputs, data points and model)?

Fraunhofer IKS

# Safety Artifacts: Equivalence Classes of Equal Behavior

## Identification and Validation of Equivalence Classes of Equal Behavior

**Idea**: If a leaf contains more samples than $S_{min}$ a split could have been possible, however, it was not required as $\theta$ has not been exceeded, so all samples have the same output.

Identification:

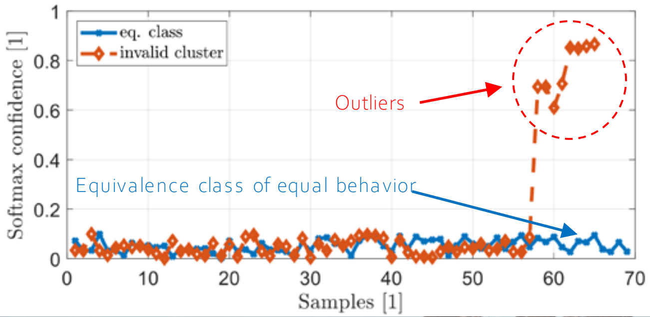
- Search for leaves that fulfill $S_n > S_{min}$ and $\Delta(s, n) < \theta$
- **Aggregate all split criteria** $s$ along the path from origin this very leaf

Validation:

1. Check **validity** of the identified equivalence class **within the complete data-set** (training and test)
2. Check identified equivalence class **against system**

→ Almost all the identified equivalence classes converge on a combination of factors representing technical limitations of the system, such as robustness against noise or maximum detection distance.

1




2


| Input Feature | Interval | Unit |
|---|---|---|
| Object distance | all | [m] |
| Object area | x ≤ 3.6233 | [m²] |
| Object occlusion | all | [%] |
| Noise variance | 74 ≤ x | [%] |

Fraunhofer IKS

# Safety Artifacts: Unknown Unknowns?

## Identification of Root Cause by Process of Elimination

**Idea**: Show by process of elimination that the only possible explanation for the existence of inconsistent clusters are unknown unknowns.

Identification:

- Search for leaves that fulfill $\boldsymbol{S_n} > \boldsymbol{S_{min}}$ and $\Delta(\boldsymbol{s},\boldsymbol{n}) \geq \boldsymbol{\theta}$
- Check if their existence can be **explained by other causes in the ML development cycle**; if not, possible unknown unknown.


**Process of Elimination**:

1. Use different ML methodologies
   → Must be explainable method
2. Investigate data
   → Check: Balanced distribution, bias, accurateness,
   3 etc.
3. Inspect input features
   → Find: „Noticeable abnormalities" (e.g., contradictions)

Fraunhofer IKS

# Safety Artifacts: Unknown Unknowns?

## Identification of Root Cause by Process of Elimination

**Idea**: Show by process of elimination that the only possible explanation for the existence of inconsistent clusters are unknown unknowns.

Identification:

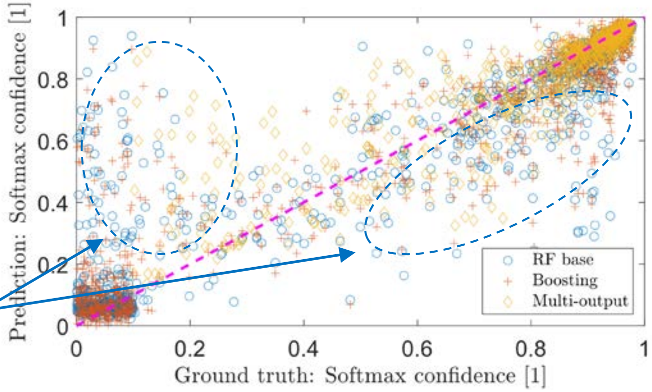
- Search for leaves that fulfill $S_n > S_{min}$ and $\Delta(s,n) \geq \theta$
- Check if their existence can be **explained by other causes in the ML development cycle**; if not, possible unknown unknown.


Process of Elimination:

1. Use different ML methodologies
   → Must be explainable method
2. Investigate data
   → Check: Balanced distribution, bias, accurateness, etc.
3. Inspect input features
   → Find: „Noticeable abnormalities" (e.g., contradictions)

1

Changing the model does not resolve inhomogeneous clustering, similar performance for all tested models.
→ Remaining possibilities: specification insufficiencies in data and/or input.



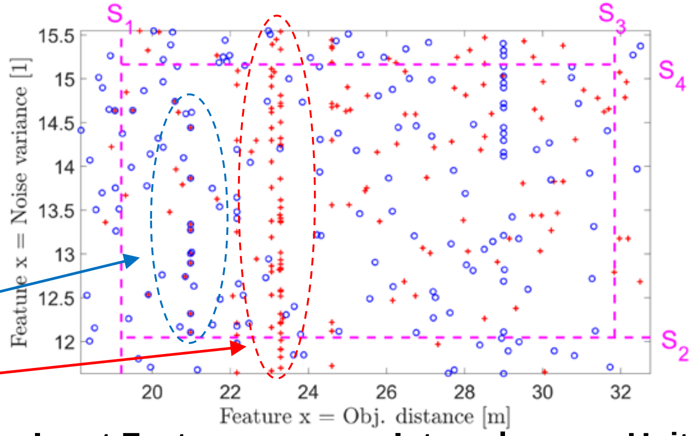

2 & 3

Inspection of data revealed:

- No visible data imbalance or bias

but

- Noticeable „contradictions" and
- Low confidence region (invariant to noise).




| Input Feature | Interval | Unit |
|---|---|---|
| Object distance | 18.85 ≤ x ≤ 31.25 | [m] |
| Object area | 2.018 ≤ x | [m²] |
| Object occlusion | all | [%] |
| Noise variance | 62 ≤ x ≤ 78 | [%] |

# Safety Artifacts: Unknown Unknowns!

## Identification and Mitigation of Unknown Unknowns

**Idea**: Show by process of elimination that the only possible explanation for the existence of inconsistent clusters are unknown unknowns.

Identification:

- Inspect input features

  Identify: „Noticeable abnormalities" (e.g., contradictions)

Mitigation:

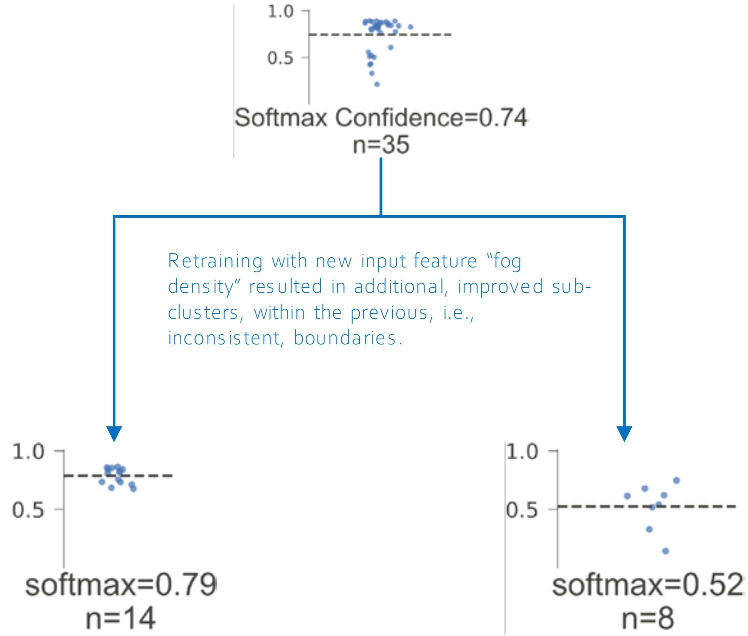
- Introduce new input feature

  Retrain Model with updated input feature

  Check the leaf(s) that fall within the previously identified, inconsistent cluster

→ Please be aware that the new leaves can still result in any of the basic cases for DT leaves (as shown on slide 5), so the analysis might not end conclusively every time.

Discovered that *carla.WeatherParameters.fog_density* has a nonzero value for all low confidence cases within this cluster.
→ Included parameter as new input feature.



# What has been done and what is left to do

## Summary & Future Scope

- Developed an approach to use explainable ML for safety analyses of "Equivalence Classes of Equal Behavior" and "Unknown Unknowns"
- Equivalence Classes are derived from "naturally" converging data clusters after training
  - Successful validation (against collected data and system behavior) indeed indicate an identified "Equivalence Class of Equal Behavior"
- The starting point for Unknown Unknowns are inconsistent DT leaves that do not exceed the defined thresholds
  - By process of elimination the only possible explanation for their existence is an unknown unknown
  - Identification of this unknown unknown and subsequent integration into the development cycle can mitigate their effect

- So far, we were able to identify <u>one</u> unknown unknown by disentangling <u>one promising</u> inconsistent data cluster
- Identified Equivalence Classes cannot always be interpreted to be meaningful
- The requirement of explainable ML limits the applicability of this approach

# Contact

Iwo Kurzidem
Systems Safety Engineering
Tel. +49 89 547088 - 350
iwo.kurzidem@iks.fraunhofer.de

Fraunhofer Institute for Cognitive Systems IKS
Hansastraße 32
D-80686 Munich
www.iks.fraunhofer.de/en

Fraunhofer
IKS

Fraunhofer Institute for Cognitive Systems IKS

## References:

[1] S. Burton, B. Herd, “Addressing uncertainty in the safety assurance of machine-learning”, Frontiers in Computer Science Hypothesis and theory article, 2023

[2] International Organization for Standardization, “Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation (ISO/TR 4804:2020)”, 2020.

[3] International Organization for Standardization, “Safety Of The Intended Functionality - SOTIF (ISO/-PAS 21448)”, 2019.

[4] Kurzidem, Iwo, et al. "Safety Assessment: From Black-Box to White-Box." 2022 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 2022.