Fraunhofer IKS

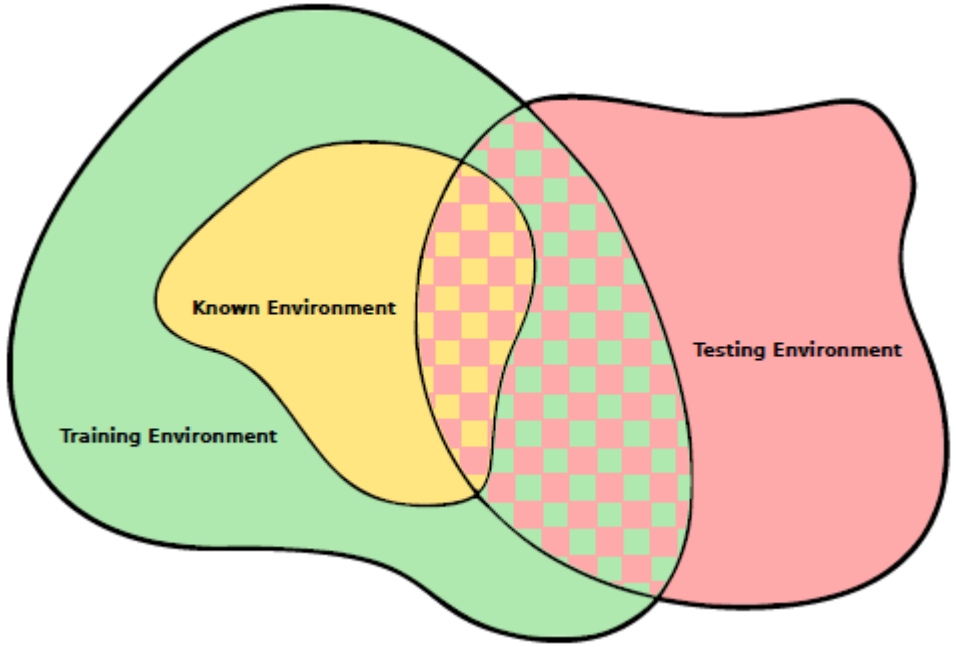
Fraunhofer Institute for Cognitive Systems IKS

14.02.2023 – Dirk Eilers

Safety Assurance with Ensemble-based Uncertainty Estimation and overlapping alternative Predictions in Reinforcement Learning

Dirk Eilers, Simon Burton, Felippe Schmoeller da Roza and Karsten Roscher

In-distribution (ID) vs. out of Distribution (OOD)



- Agent will act certain in the known environment and generalize to some extend
- There will be uncertain predictions outside the known environment -> Epistemic Uncertainty

Problem outline

Epistemic Uncertainty Estimation

- A Reinforcement Learning (RL) model will predict an action in whatever state it will find itself in
- Even when the model is unsure about the situation, the output will not indicate that
- This becomes a problem especially in applications with safety critical constraints
- Epistemic Uncertainty Estimation can detect unseen states during training (OOD)

Problem:

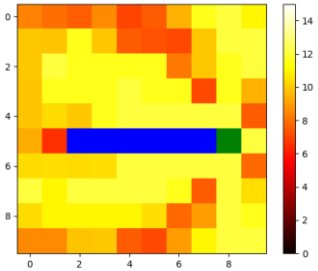
- In states with multiple possible and equally valid actions an ensemble of models might indicate uncertainty even though the state is safe
  -> uncertainty from alternative valid actions can "overlap" with uncertainty from unseen states
- For the **Safety Assurance argument,** it is difficult to demonstrate the absence of unreasonable risk of unsafe actions coming from these uncertainties

Uncertainty Estimation on Gridworld scenario

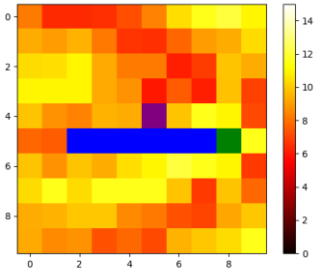
Delta to ID

- Simplified Gridworld scenario with known hazards (blue), goal in green and unknown obstacle (purple)
- Bootstrapping to train an ensemble of neural networks on different subsets of the available data
- “Compare" the given OOD situation to a nearest ID situation to identify uncertainty from unknown's vs the “valid" uncertainty from alternative actions.
- Results: IDD will „isolate“ the hotspot around the unknown obstacle to help indicate for a backup policy

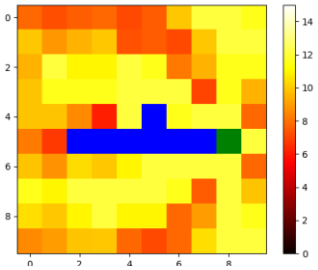
- ACV for scenario with known hazards



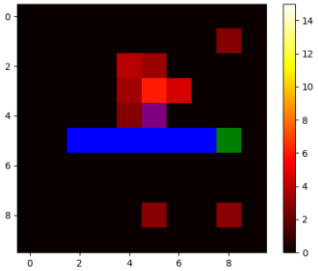
- Scenario with additional unknown obstacle



- Scenario with ID-equivalent known hazard

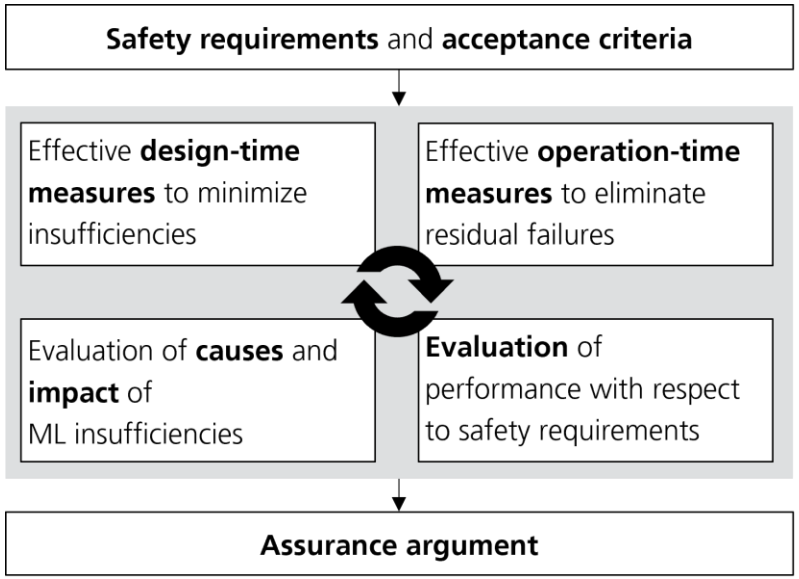


- IDD for scenario with unknown obstacles

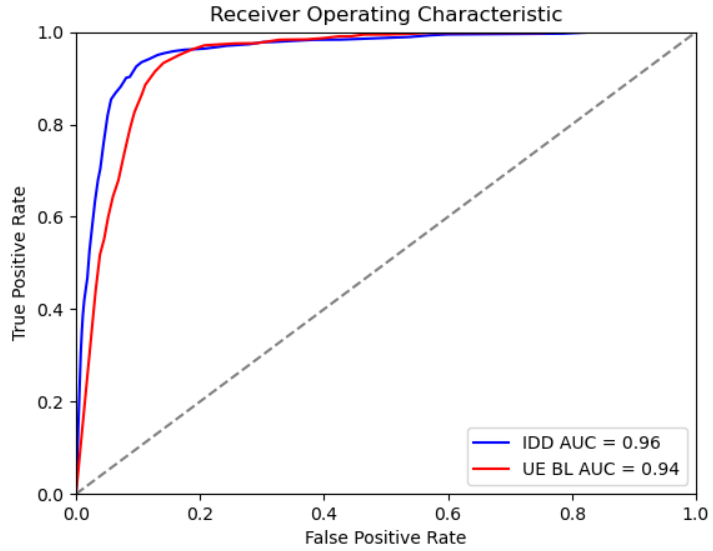


Safety Assurance Argumentation

- In this paper: exemplary show-case for the safety argumentation on a simplified toy use-case
- Iterative process to derive design-time and operation-time measures to reduce residual safety risk
- Uncertainty estimation as operation-time measure to mitigate impact of residual errors in the ML component
- Goal Structuring Notation (GSN) to identify potential causes of insufficiencies and measures to reduce impact
- Iterate through the safety assurance process when dealing with ontological uncertainties



Iterate if a convincing argument cannot be achieved or changes in the environment are detected



Takeaways

1. **Ensemble-based uncertainty estimation can be utilized in safety critical applications**
2. **Delta-to-ID (IDD) can iliminate the effect of alternative valid actions within the ensemble variance**
3. **Iterative safety assurance argumentation can be utilized for systems with ontological uncertainties**

Contact Information:

Dirk Eilers
Fraunhofer Institute for Cognitive Systems (IKS)
dirk.eilers@iks.fraunhofer.de

Your are welcome to discuss with us at the poster booth.

