exida® AI

AI & Safety in Automotive Applications

AI Safety Standards, Current Applications, and Future Work

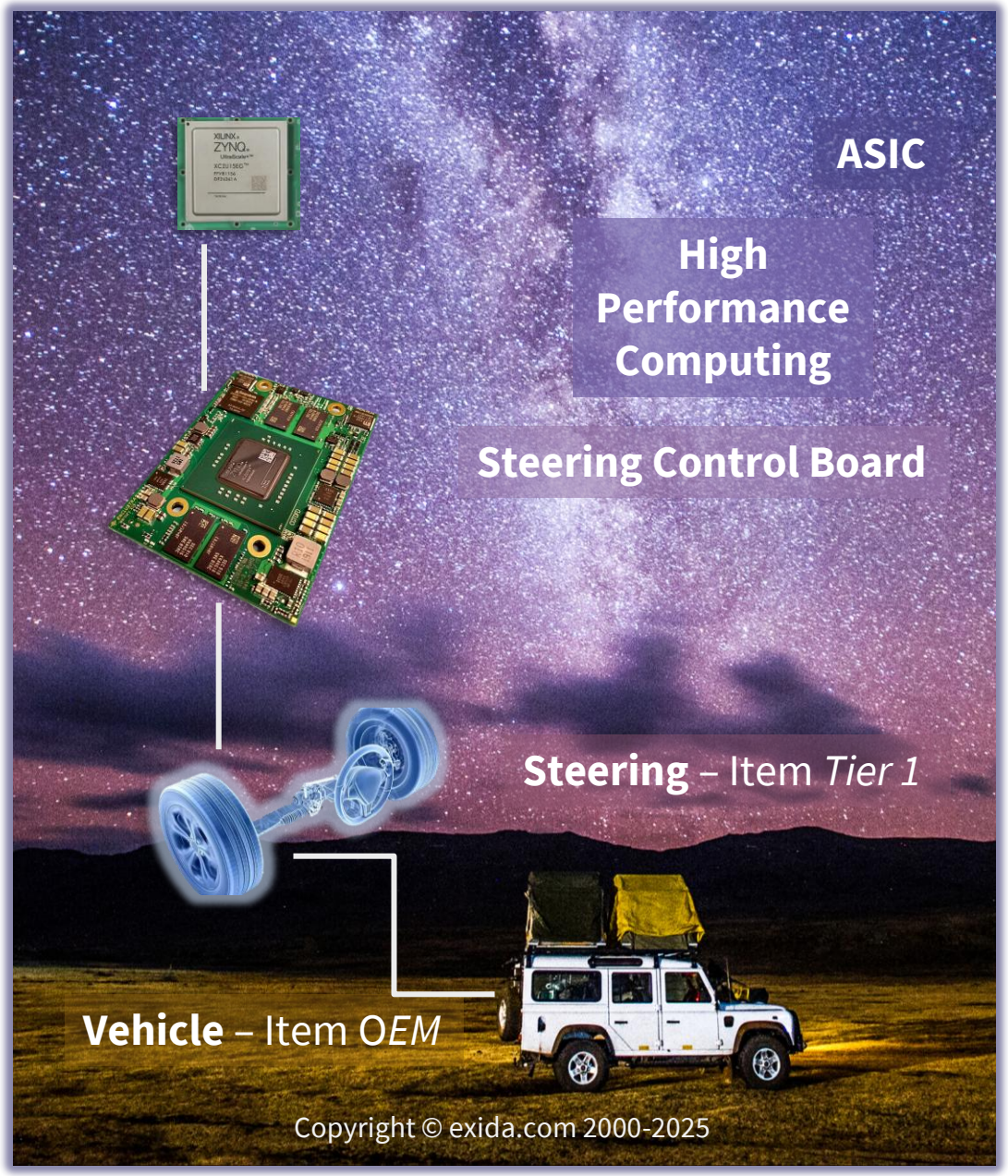
Dr. Molly O'Brien

September 24, 2025

Molly O’Brien, Ph.D.

- PhD in Computer Science, from The Johns Hopkins University
- Expertise in Computer Vision, Deep Learning, and Surgical Robotics
- Senior Safety Engineer at exida
- SME in safe AI research efforts and projects at exida
- Performs Hardware and Software Functional Safety Assessments, FMEAs, and FMEDAs
- Automotive, Process, and Semiconductor Industry

Automotive Applications

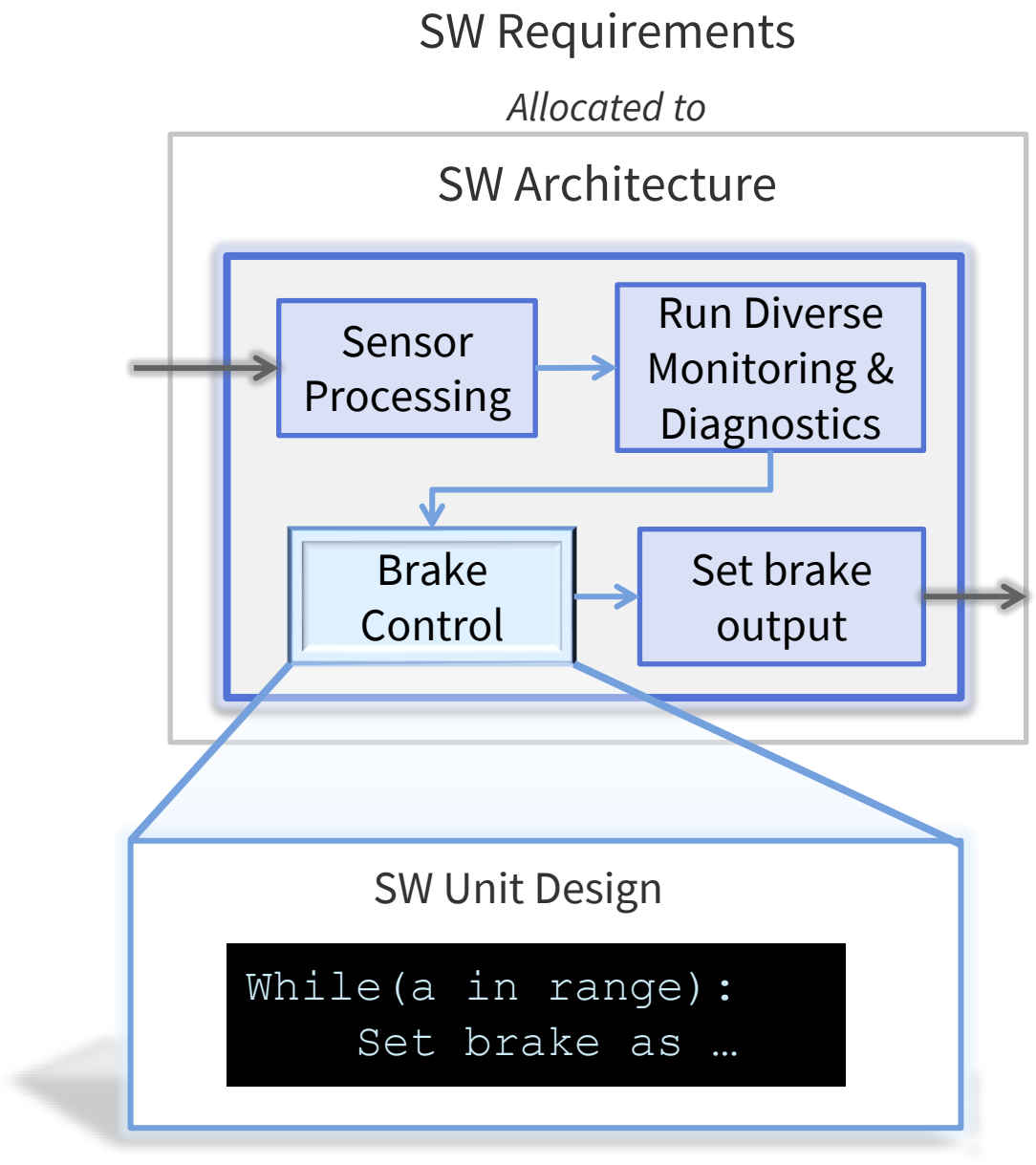


Space Applications

Automotive Functional Safety

- Independent safety analysis and certification body, impartial 3rd party technical reviewers
  → **exida's role**
- Assessment / certification are done to different Automotive Safety Integrity Levels (ASILs)
- Higher ASIL:
  - More stringent design and validation activities
  - Lower quantitative probability of hardware failure
- Consensus in the community is:
  - HW analyzed quantitatively
  - **SW analyzed qualitatively**

Software Assessment



Software Assessment

Verified with

SW Requirements

Allocated to

Bidirectional Traceability

SW Architecture

Sensor Processing

Run Diverse Monitoring & Diagnostics

Brake Control

Set brake output

System

SW Unit Design

```
While(a in range):
    Set brake as …
```

Black-Box Validation Testing

Module Testing

Testing at interfaces

Code Coverage Analysis

Static Code Analysis

Code Reviews

Branch Testing

Unit Testing

AI and Neural Network Assessment

AI Requirements

drive

Data Collection & Labelling

optimizes

Input → AI Model → Prediction

AI and Neural Network Assessment

AI Requirements

drive

Data Collection & Labelling

optimizes

Input → **AI Model** → **Prediction**

Bidirectional Traceability

Cannot trace requirements to implementation in NN

Code Coverage Analysis

Static Code Analysis

Code Reviews

Verified with

Module Testing

Testing at internal interfaces

Branch Testing

Unit Testing

Black-Box Validation Testing

AI Assessment Challenges

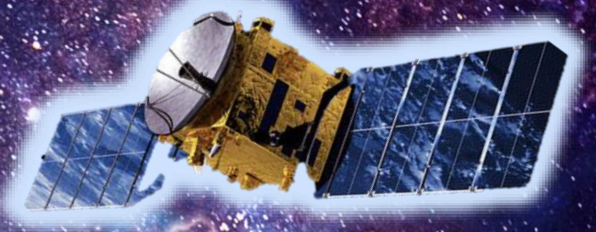
- The following traditional techniques are not directly applicable to typical AI Models:

Technique	Objective
Tracing requirements to implementation	Confirm all requirements are fulfilled by the SW
Code coverage analysis	Confirm only SW requested by requirements is created - Make sure there are no missing requirements - Make sure there is no unintended functionality
Static code analysis	Confirm SW was developed according to coding standard
Statement / branch testing	Confirm all lines of code have been exercised at least one in testing (so that there is not unexpected behavior in operation)
Unit Testing	Confirm building blocks of the SW behave as expected
Module Testing	Confirm higher level blocks of the SW behave as expected
Testing at internal interfaces	Confirm the correct flow of information between SW components

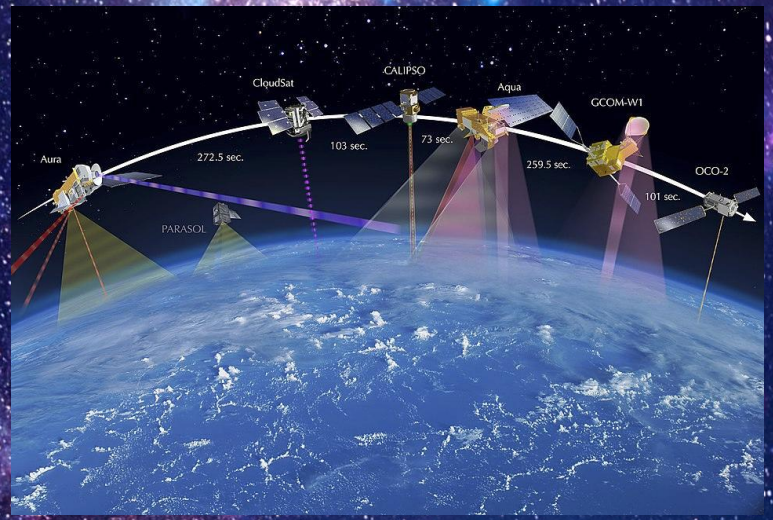
Automotive Applications

Control ← Perception

Space Applications

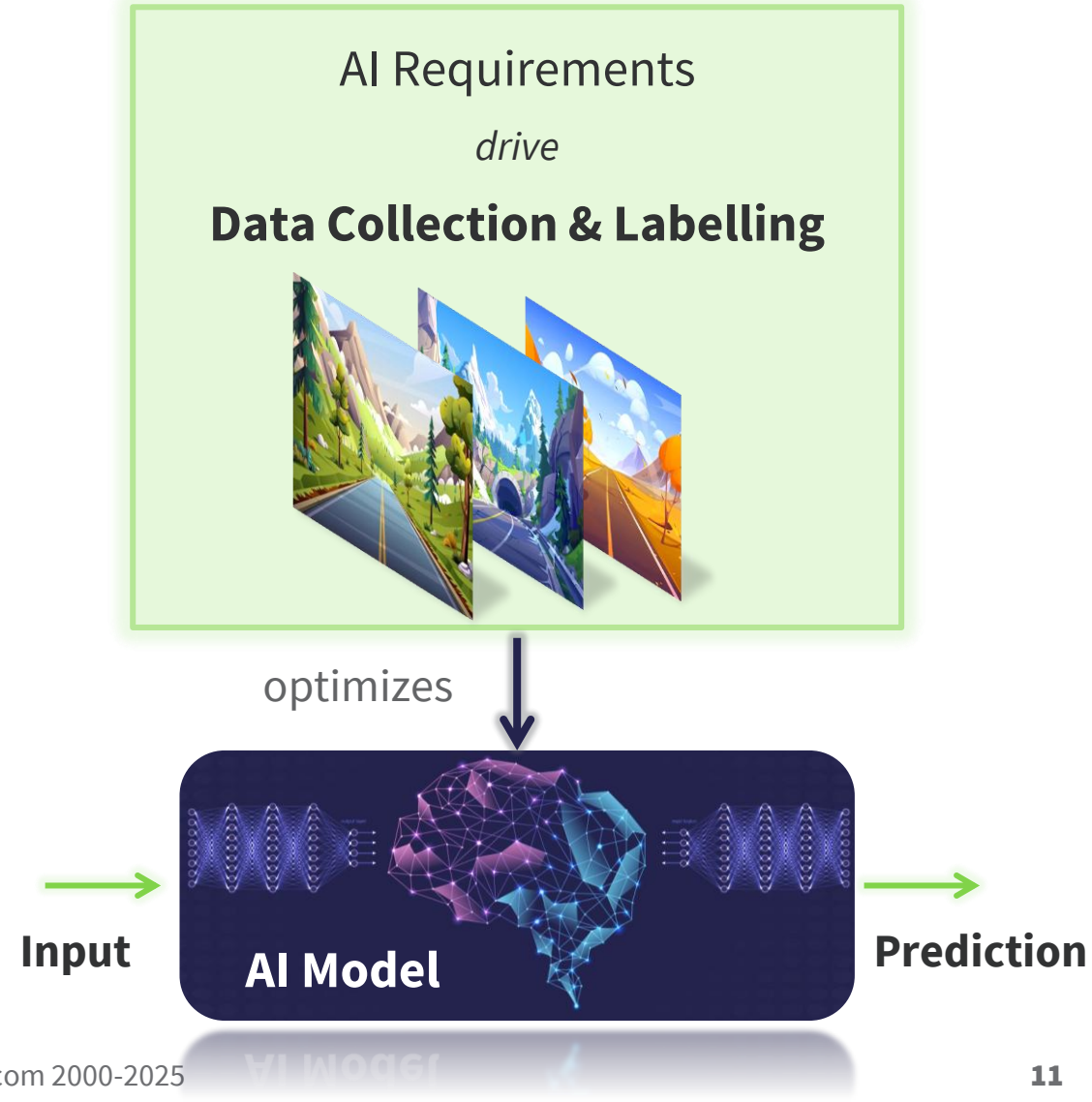


Control ← Perception



New Challenges for AI Assessment

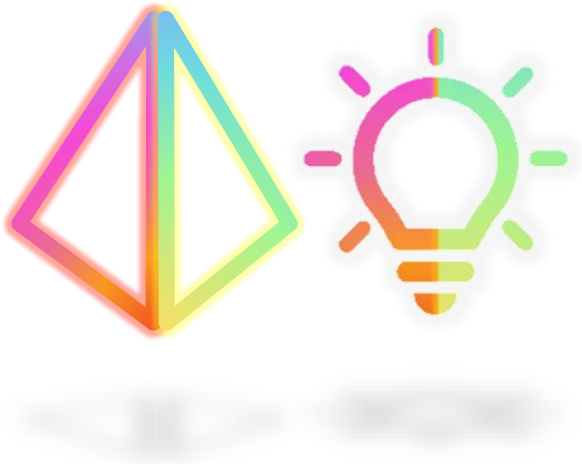
- AI Requirements must capture key performance
  - Acceptance Criteria
  - Legal Requirements
- Training data specifies the behavior of an AI model. This introduces new challenges for AI verification like:
  - Confirming “enough” training data is used
  - Confirming training data “covers” the entire Operating Domain
  - Confirming the AI Model is used in scenarios that “match” the training data



New design and verification techniques are needed to analyze:

-AI Model Behavior

- Data representing complex Environments



Functional Safety Standards

For ML / AI

First modern safety standard

Process Industry

IEC 61508

Derived

Automotive Industry

ISO 26262

ISO IEC TR 5469
Jan-2024

- Technical report that summarizes the state-of-the-art AI risk factors, techniques for verification / validation, and mitigating risks
- Does not include requirements.
- Will be integrated in requirement form for the next version of IEC 61508

ISO ISO/PAS 8800
Dec-2024

Requirements for AI in automotive

ISO/PAS 8800 Structure and Key Areas

7. AI **safety management**

8. Assurance arguments for AI systems

10. Selection of AI **technologies, architectural** and **development measures**

11. Data-related considerations

9. Derivation of AI **safety requirements**

15. Confidence in use of AI **development frameworks** and **software tools**

14. Measures during **operation**

12. Verification and **validation** of the AI system

13. Safety analysis of AI systems

ISO/PAS 8800 Structure and Key Areas

7. AI **safety management**

8. Assurance arguments for AI systems

Data for Complex Environments

10. Selection of AI **technologies, architectural** and **development measures**

11. Data-related considerations

9. Derivation of AI **safety requirements**

15. Confidence in use of AI **development frameworks** and **software tools**

14. Measures during **operation**

12. Verification and **validation** of the AI system

13. Safety analysis of AI systems

Confirming AI Model Behavior

11. Data-Related Considerations

- ISO/PAS 8800 defines an example data lifecycle
- ISO/PAS 8800 requires:
  - Identifying data insufficiencies
  - Dataset requirements
  - Dataset Safety Analysis
  - Dataset Verification
- Can be unclear **how** to meet the requirements of ISO/PAS 8800

exida Recommendations: 

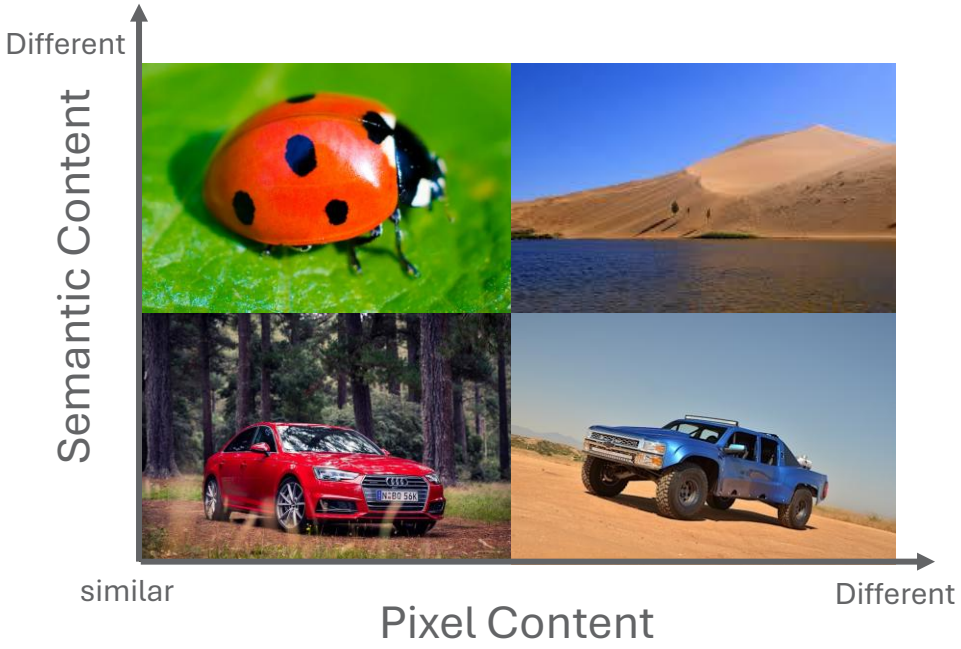
Identify key factors from the data that impact the AI model behavior. Then you can check that:

- The training and test sets **cover** the full range of these factors to avoid **insufficiencies**
- The AI Model is **used in scenarios** where these factors are within the range of the training data

Data for Complex Problems

- ML models operate on the data, e.g., pixel space, but for performance, we are interested in the semantic content
  - Semantic content: the key meaning, relationship, understanding we want the model to extract from data
- Identifying semantic content that is important
  - Enables identifying data insufficiencies
  - Drive dataset analysis & verification that is meaningful for model performance

Reference Photo



ISO/PAS 8800 Structure and Key Areas

7. AI **safety management**

8. Assurance arguments for AI systems

Data for Complex Environments

10. Selection of AI **technologies, architectural** and **development measures**

11. Data-related considerations ✓

9. Derivation of AI **safety requirements**

15. Confidence in use of AI **development frameworks** and **software tools**

14. Measures during **operation**

12. Verification and **validation** of the AI system

13. Safety analysis of AI systems

Confirming AI Model Behavior

12. Verification and validation of the AI system

- Testing of AI model alone and integrated into the system
- Want to confirm AI requirements are fulfilled
- 12.3.3 Note 1: “For ML, analysis of requirements relies on statistical tests to analyze whether the safety relevant performance requirements are met.”
- Use good practice from ISO 26262 for verification
  - Methods to specify test cases, pass / fail criteria, etc., hierarchical integration
- Safety aware metrics, safety-relevant examples

AI Statistical Analysis

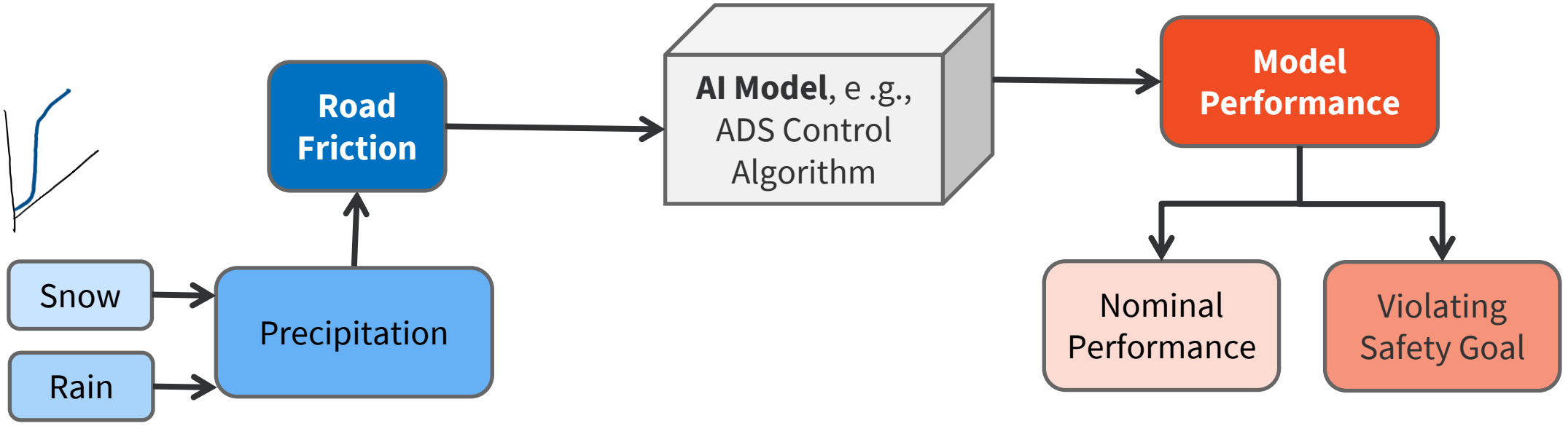
- AI Statistical Analysis provides evidence that the learned model performs the intended function and meets the KPIs
- There are many possible statistical analysis techniques, for example:
  - Bayesian Network driven simulation and analysis
  - Network Generalization Prediction
- **Exida recommends analysis techniques that:** 
  - Provide insight into what impacts the model's performance
  - Identify targeted data of interest
  - Leverage application knowledge in the analysis

Bayesian Network Driven Simulation and Analysis

Input: Data-supported knowledge →Bayesian Network

Output: data of system's behavior

Probabilistic data analysis



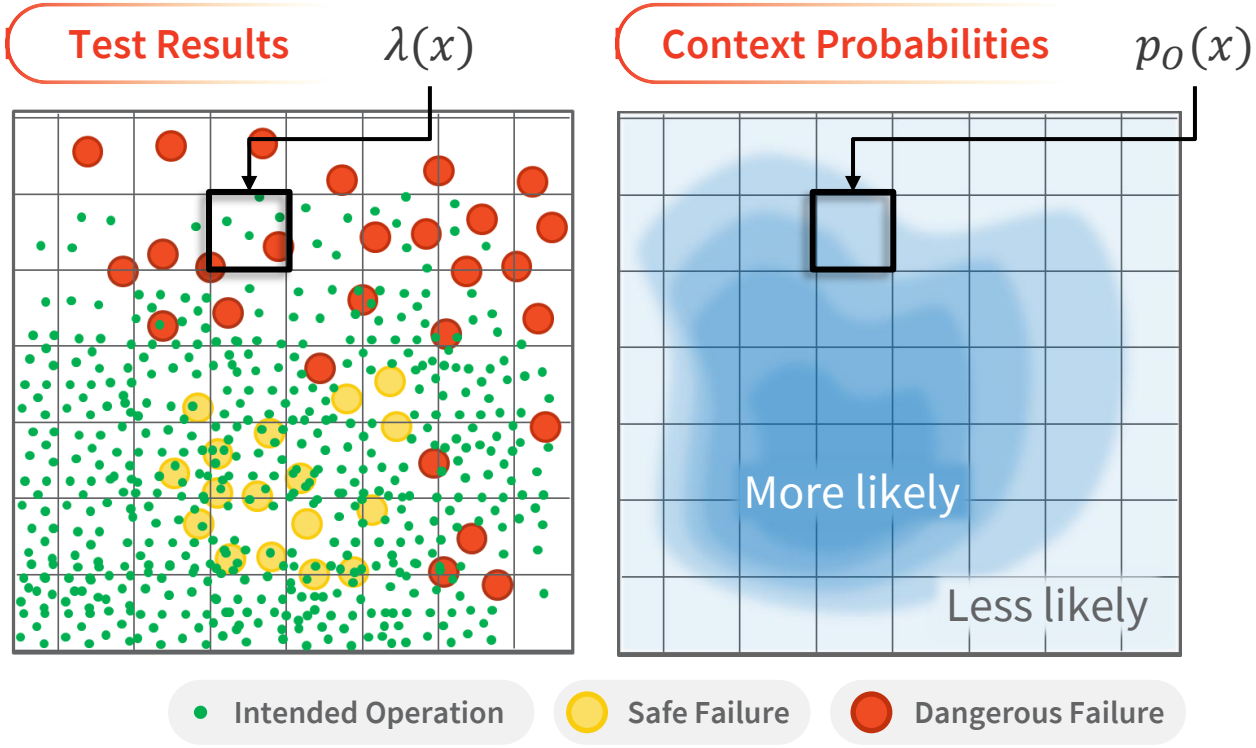
[1] Faller, Rainer. "Explainable Statistical Evaluation and Enhancement of Automated Driving System Safety Architectures." Fuzzy Systems and Data Mining X. IOS Press, 2024. 562-570.

Network Generalization Prediction

1. Test throughout the intended Operating Domains
2. Identify patterns in AI performance during testing
3. Leverage patterns seen in testing to predict AI performance in new Operating Domain

Let $p_O(x)$ describe the likelihood of context x in Operating Domain O. The failure rate in Operating Domain O, λ_O, can be calculated:

$$\lambda_O = \sum_{x \in O} \lambda(x) p_O(x)$$



[2] O'Brien, Molly, et al. "Network generalization prediction for safety critical tasks in novel operating domains. " Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

ISO/PAS 8800 Structure and Key Areas

7. AI **safety management**

8. Assurance arguments for AI systems

Data for Complex Environments

10. Selection of AI **technologies, architectural** and **development measures**

11. Data-related considerations ✓

9. Derivation of AI **safety requirements**

15. Confidence in use of AI **development frameworks** and **software tools**

14. Measures during **operation**

12. Verification and **validation** of the AI system ✓

13. Safety analysis of AI systems

Confirming AI Model Behavior

13. Safety analysis of AI systems

- Goal of safety analysis is to identify safety-related AI errors or functional insufficiencies
- Based on AI errors identified, can identify prevention or mitigation measures
- Proposed Techniques:
  - Failure Modes and Effects Analysis or HAZOP
  - Fault Tree Analysis or Event Tree Analysis
  - Bayesian Network Analysis

ISO/PAS 8800 Structure and Key Areas

7. AI **safety management**

8. Assurance arguments for AI systems

Data for Complex Environments

10. Selection of AI **technologies, architectural** and **development measures**

11. Data-related considerations ✓

9. Derivation of AI **safety requirements**

15. Confidence in use of AI **development frameworks** and **software tools**

14. Measures during **operation**

12. Verification and **validation** of the AI system ✓

13. Safety analysis of AI systems ✓

Confirming AI Model Behavior

14. Measures During Operation

- ISO/PAS 8800 requires that:
  - There is a process “to assure the AI safety” during operation.
  - Safety-related field events are identified, evaluated, and if necessary, mitigations are taken
  - Periodic maintenance is done to monitor AI risk and keep at an acceptable level

Practical monitoring during operation

- Applicable areas of research for monitoring behavior during operation:
  - E.g., confirm the data is mapping to the expected regions in the NN embedding space; see [3]
  - Error prediction and Out-of-Domain detection
  - Monitor for dataset drift
- Identifying key factors from the data that impact the AI model behavior can make it practical to monitor for distribution changes during operation

[3] O'Brien, Molly, et al. "Mapping DNN Embedding Manifolds for Network Generalization Prediction. " Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

ISO/PAS 8800 Structure and Key Areas

7. AI **safety management**

8. Assurance arguments for AI systems

Data for Complex Environments

10. Selection of AI **technologies, architectural** and **development measures**

11. Data-related considerations ✓

9. Derivation of AI **safety requirements**

15. Confidence in use of AI **development frameworks** and **software tools**

14. Measures during **operation** ✓

12. Verification and **validation** of the AI system ✓

13. Safety analysis of AI systems ✓

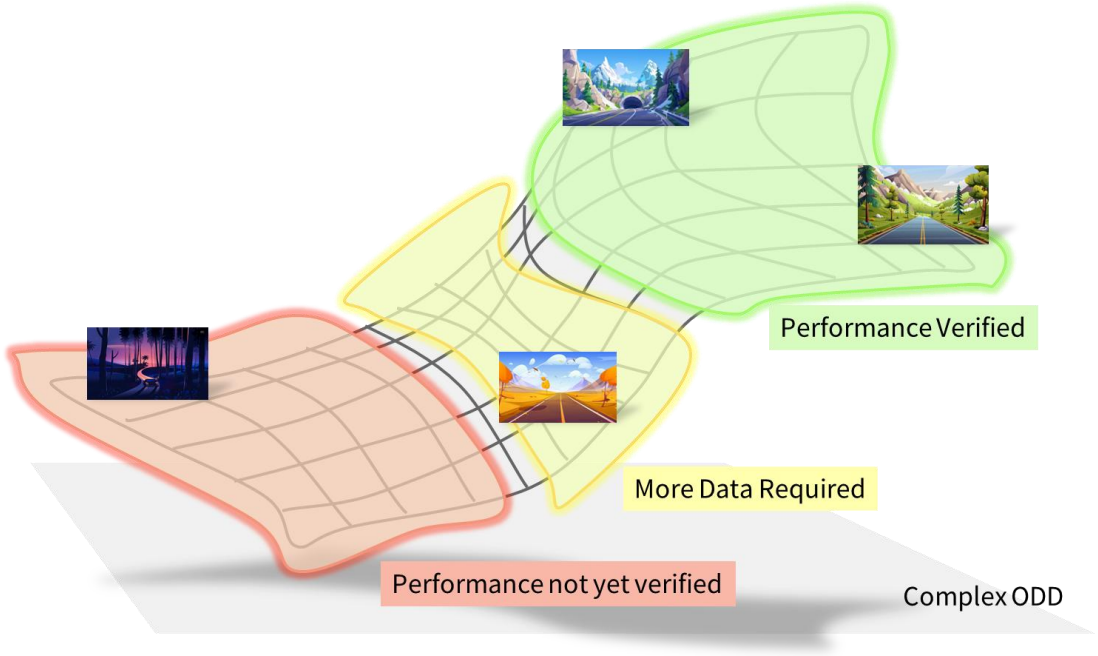
Confirming AI Model Behavior

Current State of AI in Automotive

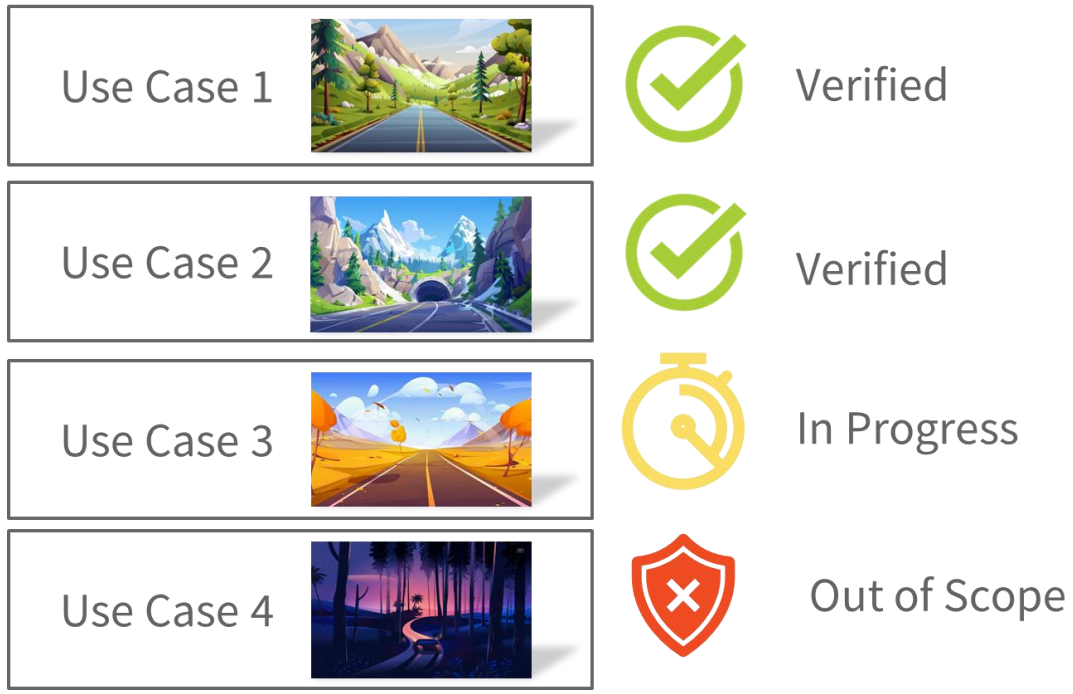
- ISO/PAS 8800 is open ended, so it leaves interpretation for how to achieve the requirements
  - ☑ Open to many different solution techniques, so it allows innovation and for companies to develop their own techniques
  - ☒ Not yet agreed upon **how** to meet the requirements, e.g., how do you demonstrate you have reduced AI risk to an acceptable level?
- exida is working with automotive OEMs, Tier Ones, and Sensor companies incorporating AI into safety functions
  - Autonomous Systems, Lane Keeping, Automatic Emergency Braking
  - Open-Source integrator companies on safety qualification and assessment
- Increasing interest in analysis, validation, and certification as more companies incorporate AI

exida Analysis and Assessment Services

Failure and Statistical Analysis



Assessment Services



New Course: ML 301
ML and AI in Safety Critical Tasks

Theoretical Sessions

Coding Sessions

Discussions

February 17th – 20th Online-Live Streaming Half-Days

Thank you!