Safety cage: an approach for safe machine learning systems

Sankar Raman Sathyamoorthy

QRTECH
INNOVATIVE ENGINEERING
an EMBRON Company

SMILE II

RISE QRTECH INNOVATIVE ENGINEERING semcon VOLVO Volvo Group

FFI Fordonsstrategisk Forskning och Innovation

1997

~125

Mölndal

QRTECH

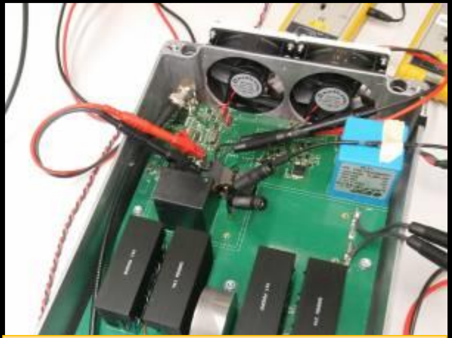
INNOVATIVE ENGINEERING

an EMBRON Company

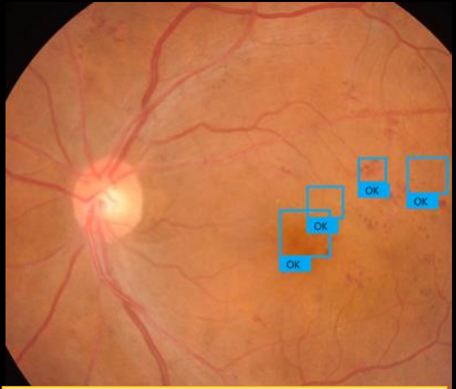
Research

Safe and Explainable AI

Electric Aircraft



Light Weight Power electronics



Automated Diagnosis

Vehicle Perception & Fault Tolerant ADAS

SMILE II

RISE QRTECH INNOVATIVE ENGINEERING Semcon VOLVO Volvo Group

FFI Fordonsstrategisk Forskning och Innovation

- Safety analysis and verification/validation of MachIne Learning based systems
- Vinnova FFI, EMK, Machine Learning
- 2017-2019
- 9 445 000 kr

Machine learning

"a large portion of real-world problems have the property that it is significantly easier to collect the data than to explicitly write the program"

Andrej Karpathy

Director of AI at Tesla

https://medium.com/@karpathy/software-2-0-a64152b37c35

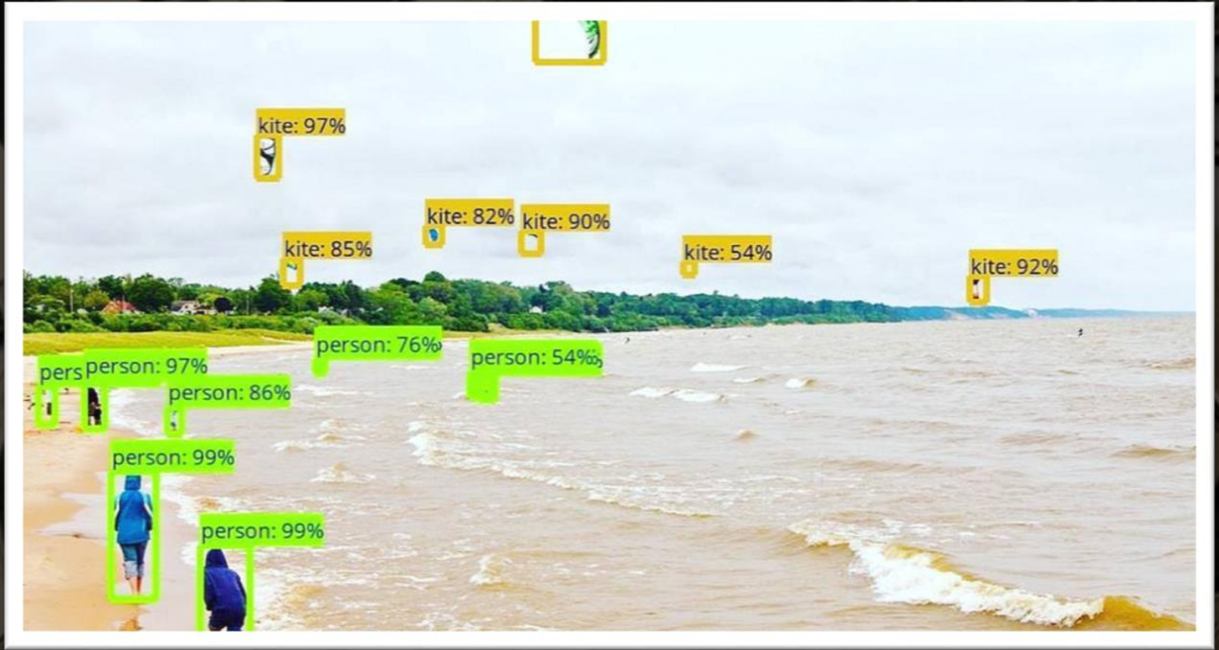
Software 2.0

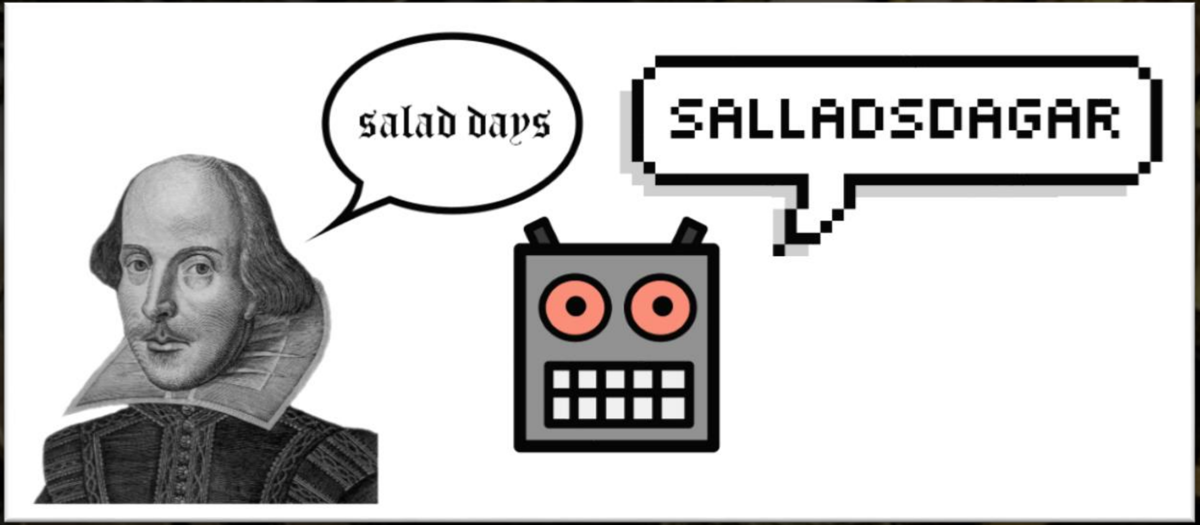
- Humans curate data and specify goals
- Backprop. and gradient descent produces millions of weights in neural network
- Humans cannot comprehend mapping from input to output

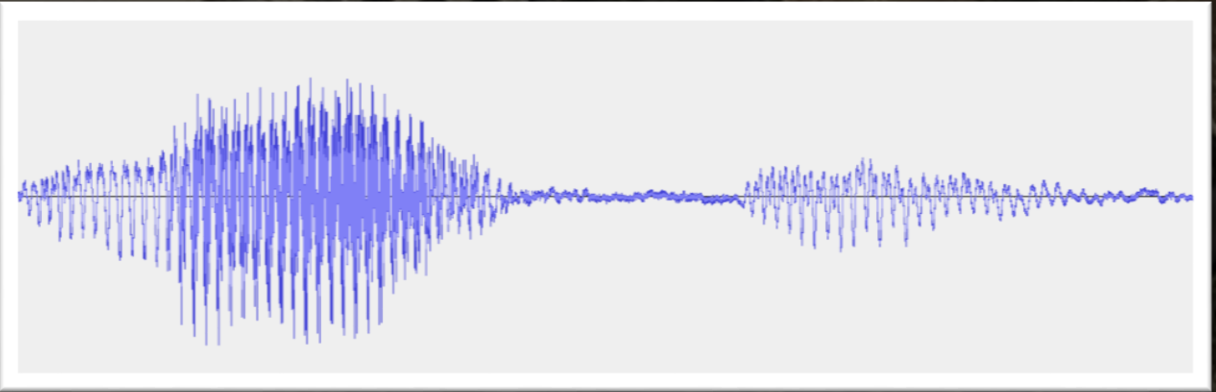
Software 1.0

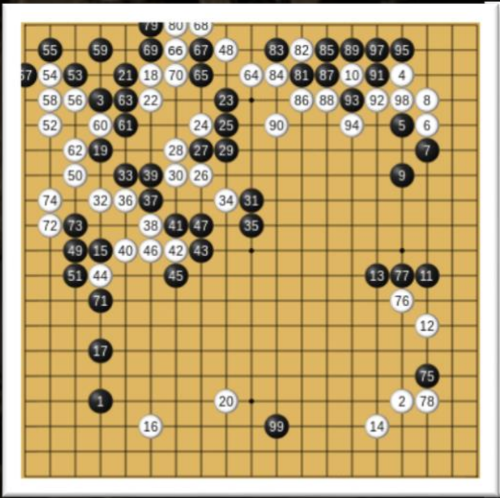
- Humans write source code
- Other humans comprehend the source code

Machine learning



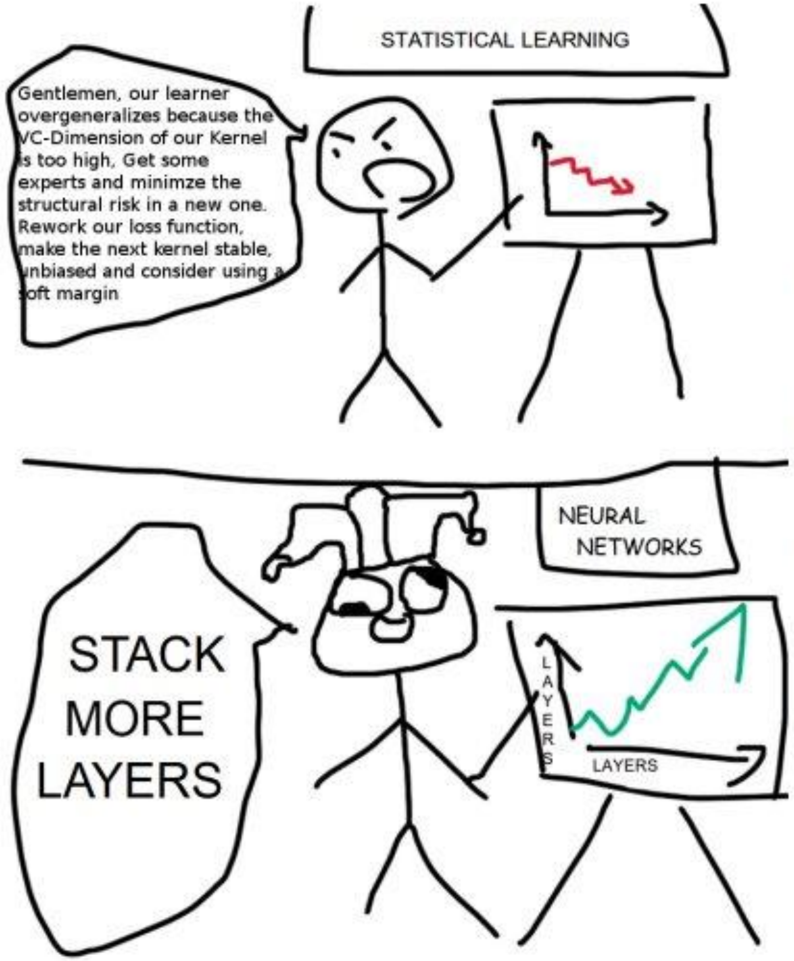


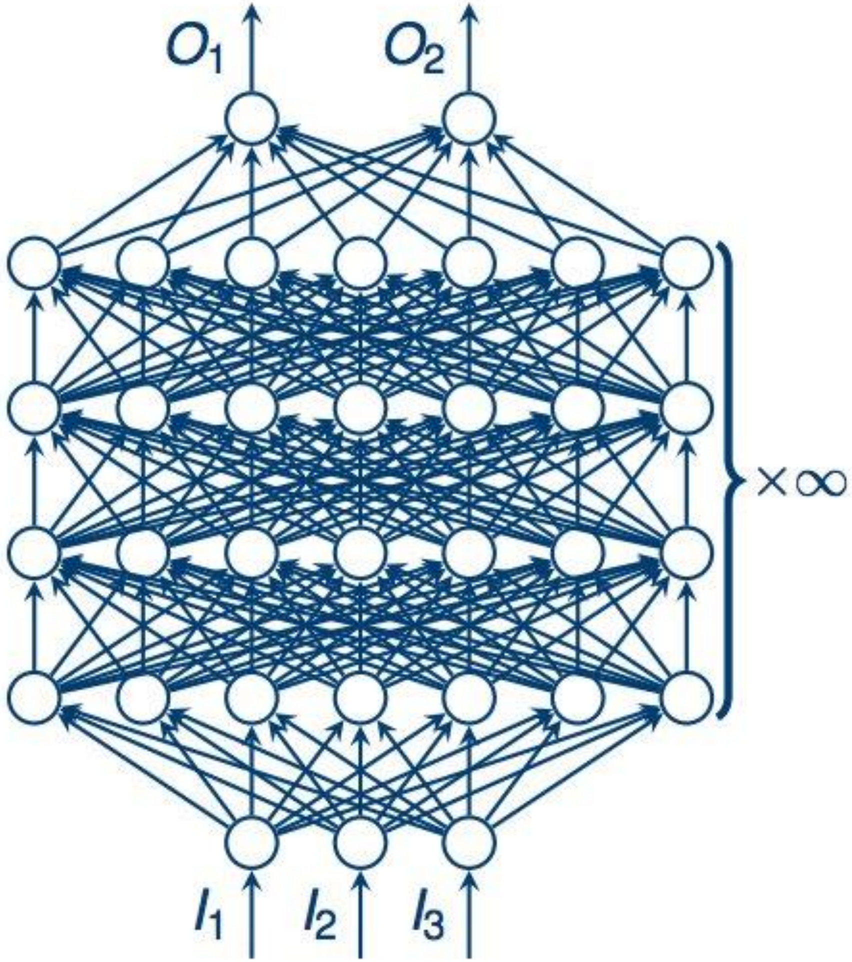




Slide from Markus Borg, RISE

Machine learning

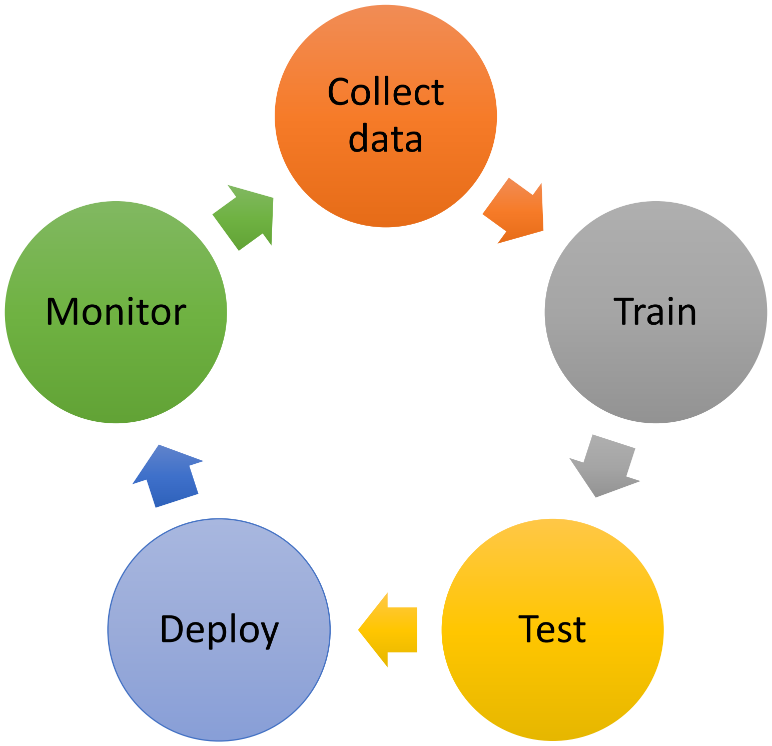


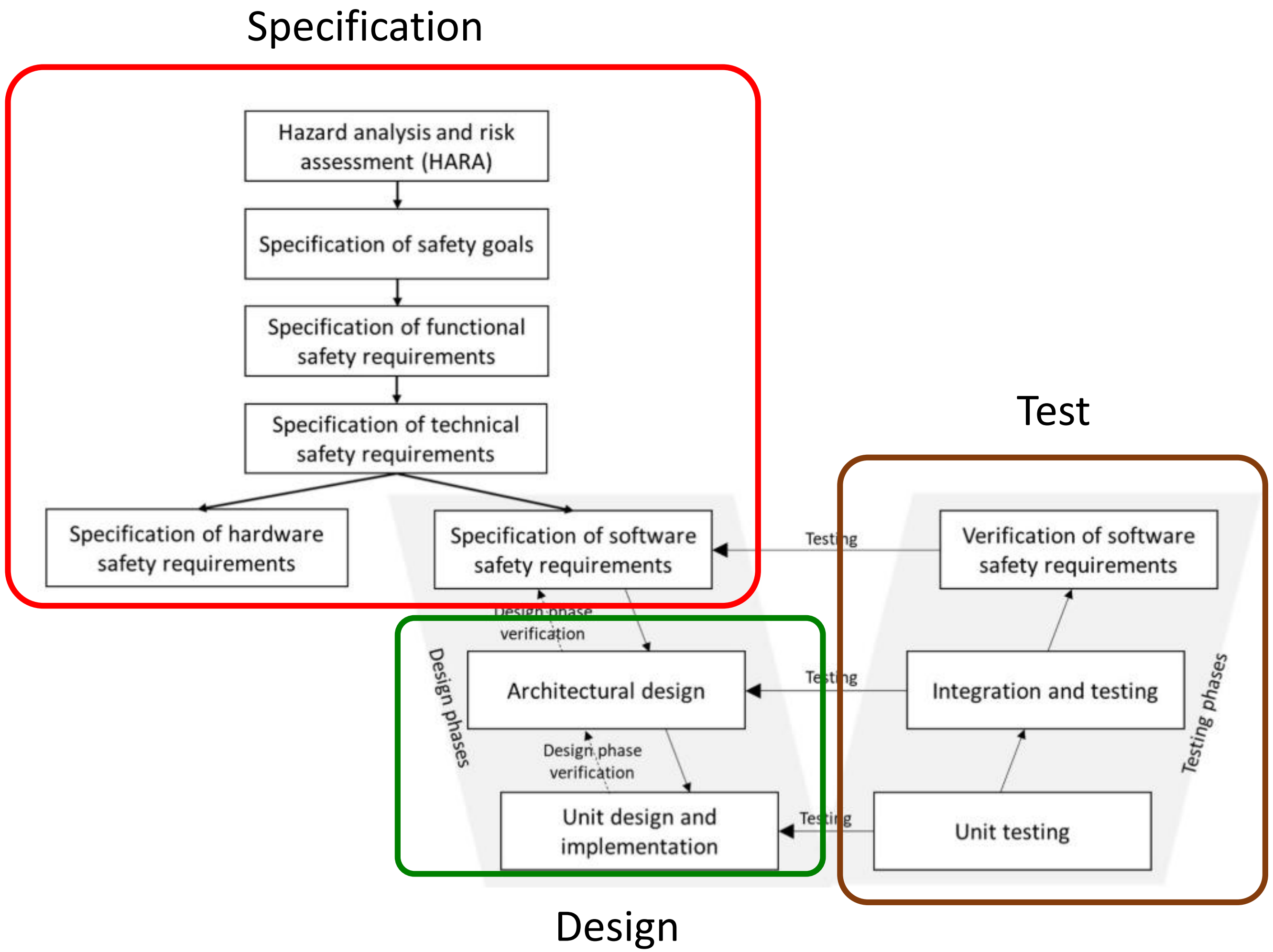


Petar Velickovic AI Group, University of Cambridge

Slide from Markus Borg, RISE

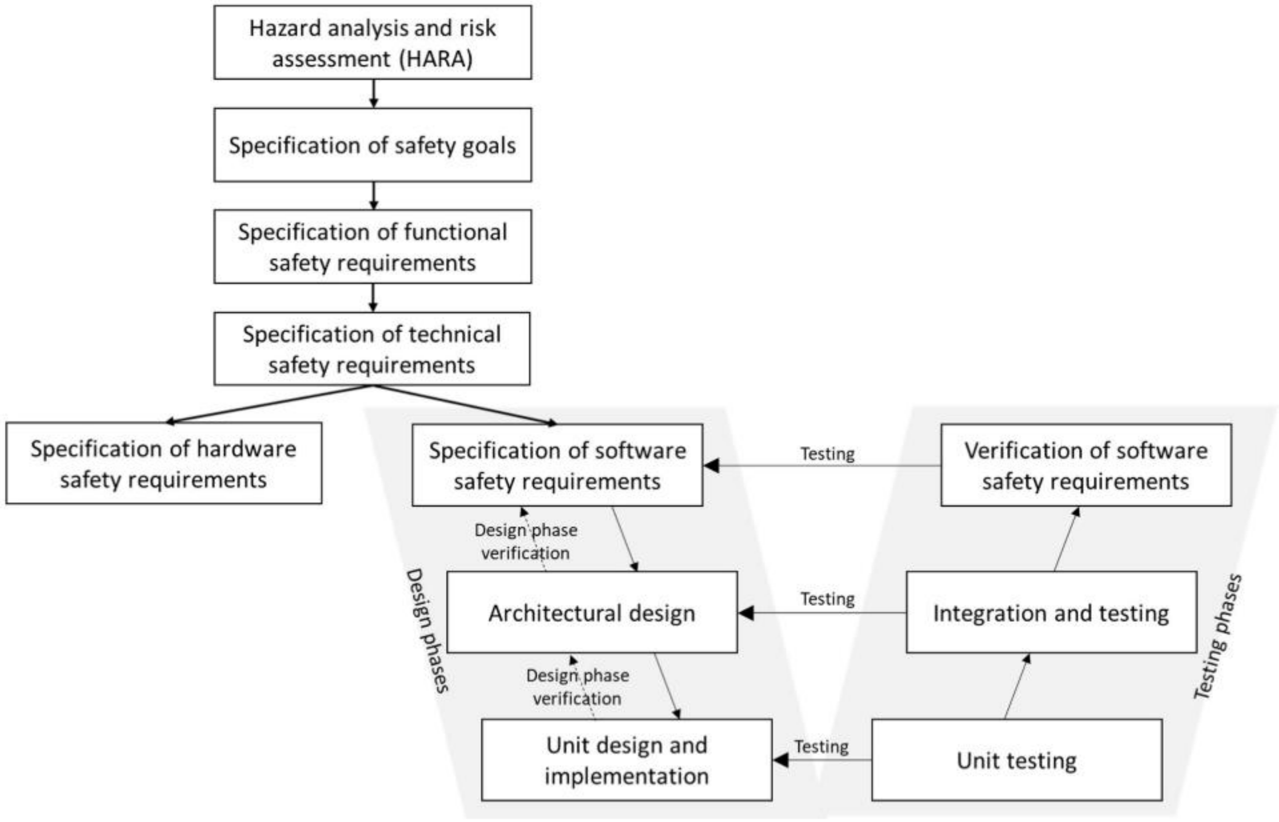
Verification and Validation – Challenges for Machine learning systems





Slide from Lars Tornberg, VCC

Verification and Validation – Challenges for Machine learning systems



<u>Testing in Machine Learning:</u>

- Estimate prediction/generalization performance
- Improve performance during model development.

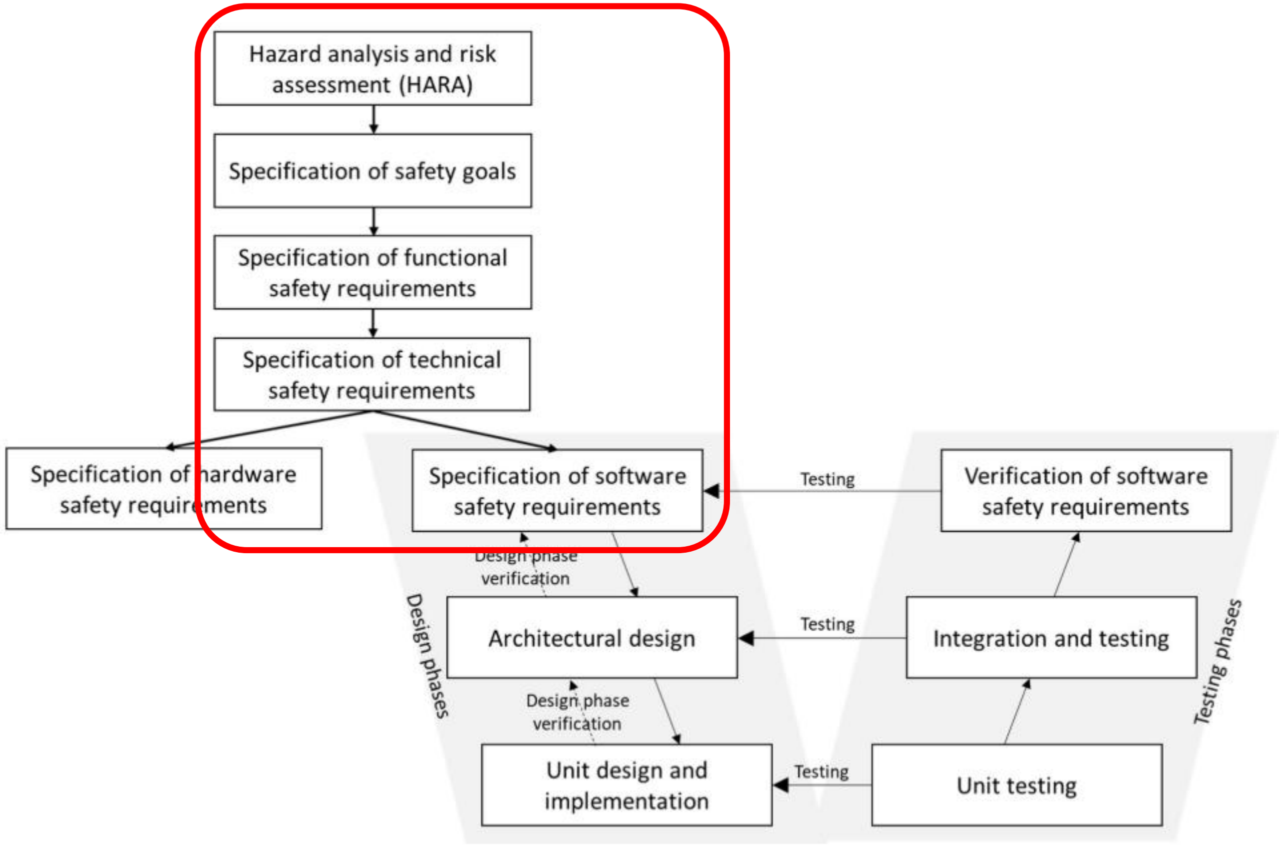
<u>Testing in Software Testing:</u>

- Other attributes e.g.,
    - Correctness,
    - Robustness,
    - Reliability,
    - Safety
    - Interpretability
    - ...
- Interaction with other system components

Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262

Slide from Lars Tornberg, VCC

Verification and Validation – Challenges for Machine learning systems

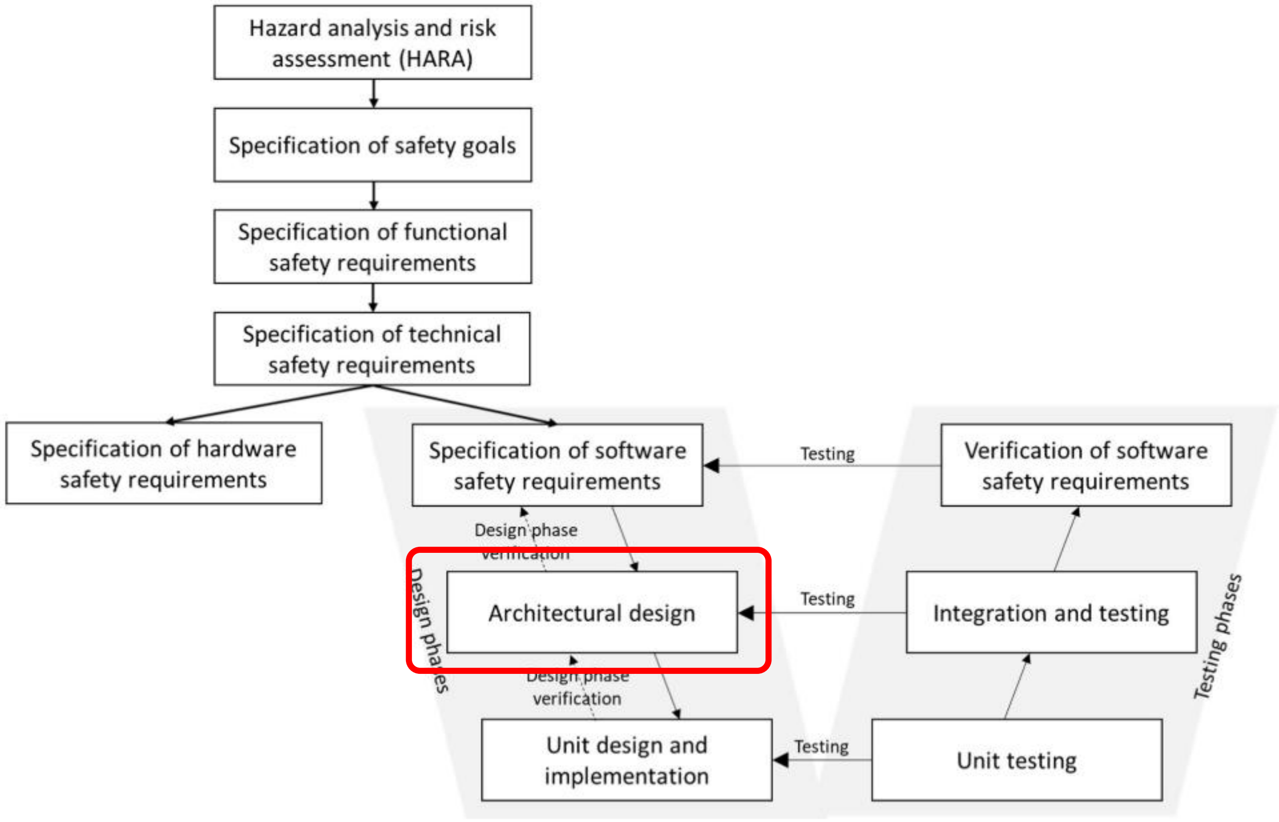


- Lack of specifications
  - Models are not rule based – learning from examples
- Training set is not a substitute for specifications
  - Specification is general
  - Training data is a sample
  - Control distributional shift
  - Data is imbalanced w.r.t. to safety critical cases.
- Specification break down is difficult
  - Important for the *safety case*, which traces the model behavior to design and specification.

Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262

Slide from Lars Tornberg, VCC

Verification and Validation – Challenges for Machine learning systems

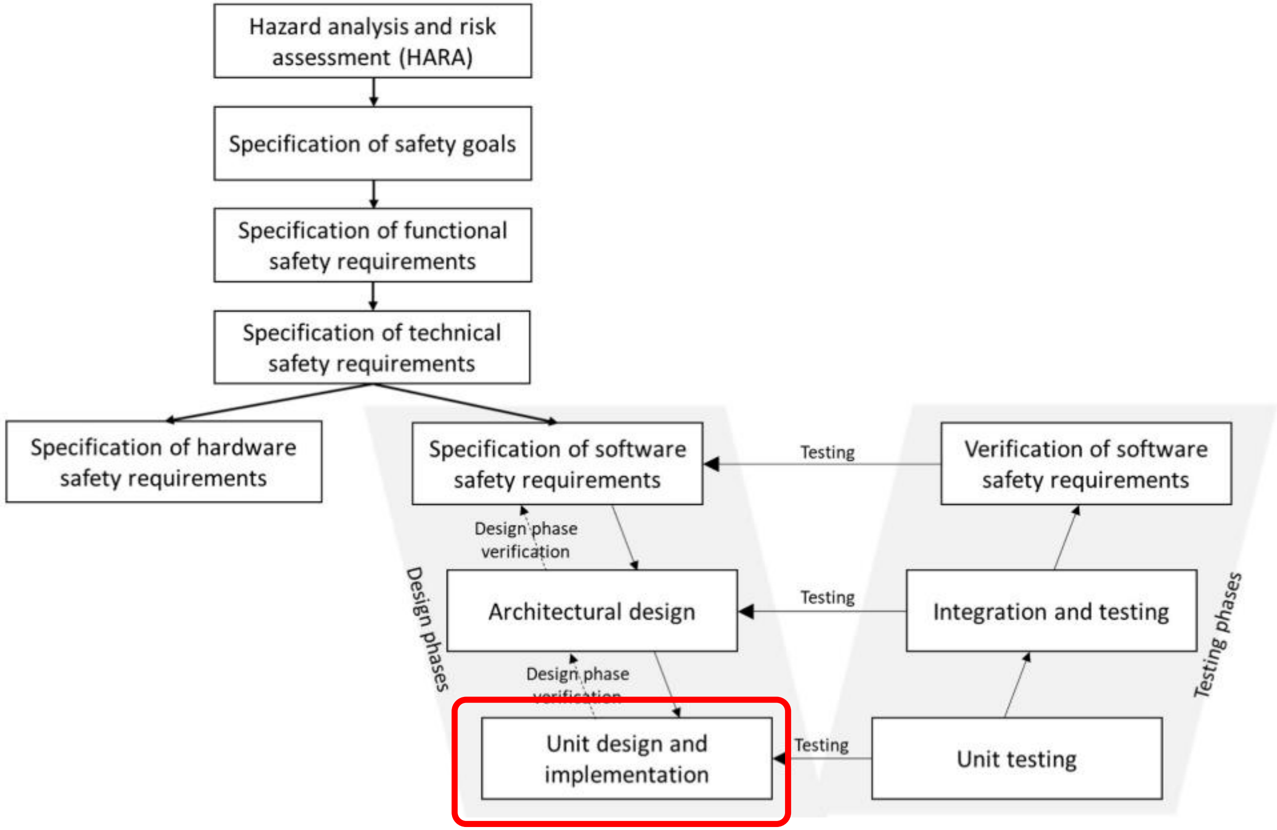


- How to control dependencies between models.
- Quality assurance of predictions/outputs.
  - Data quality
  - Where should predictions be done?
  - Trade off between execution speed and e.g., model accuracy
- Explicit vs Implicit dependencies?

Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262

Slide from Lars Tornberg, VCC

Verification and Validation – Challenges for Machine learning systems



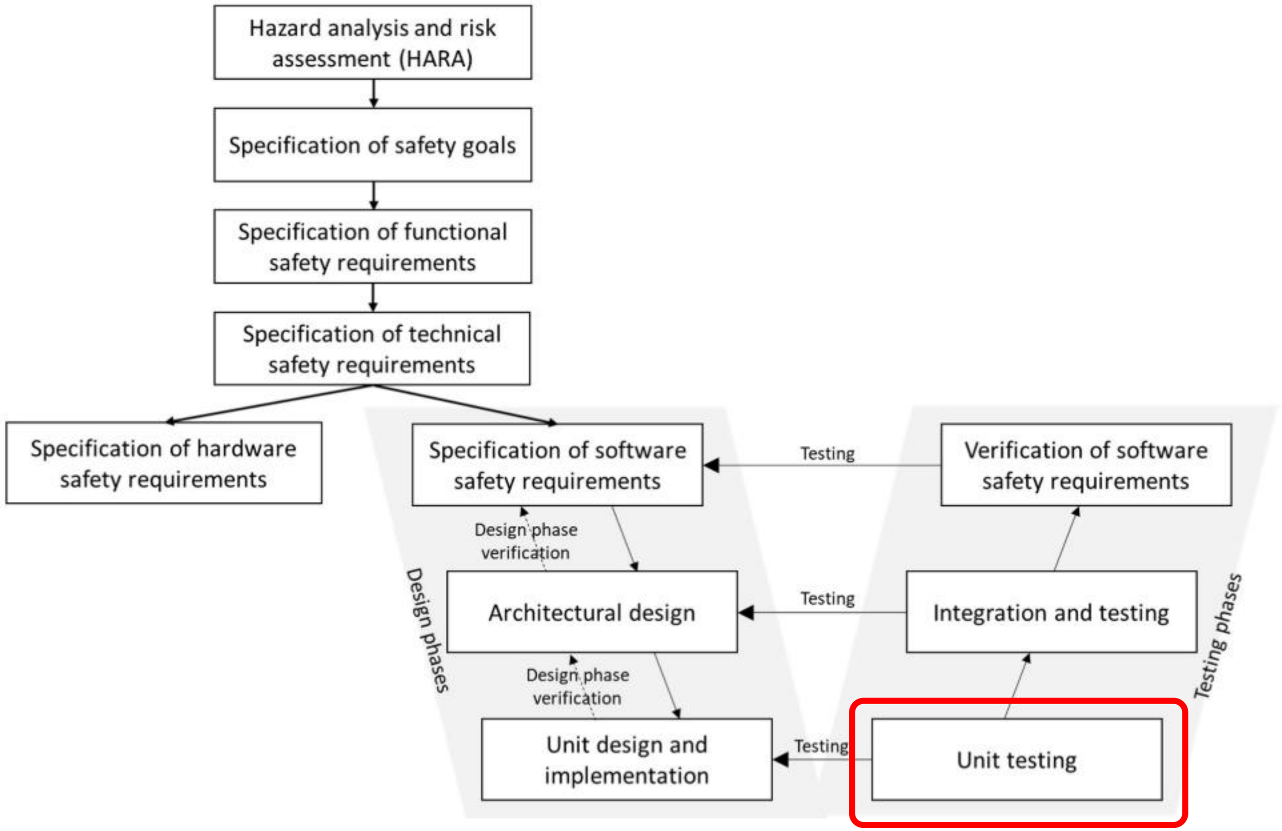
- How do we design more principled and general objective functions to include e.g.,
  - Safety aspects,
  - Fairness,
  - Interpretability,
  - Safe exploration

- Mismatch between ideal specification (what we want the model to do) and model behavior.

Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262

Slide from Lars Tornberg, VCC

Verification and Validation – Challenges for Machine learning systems

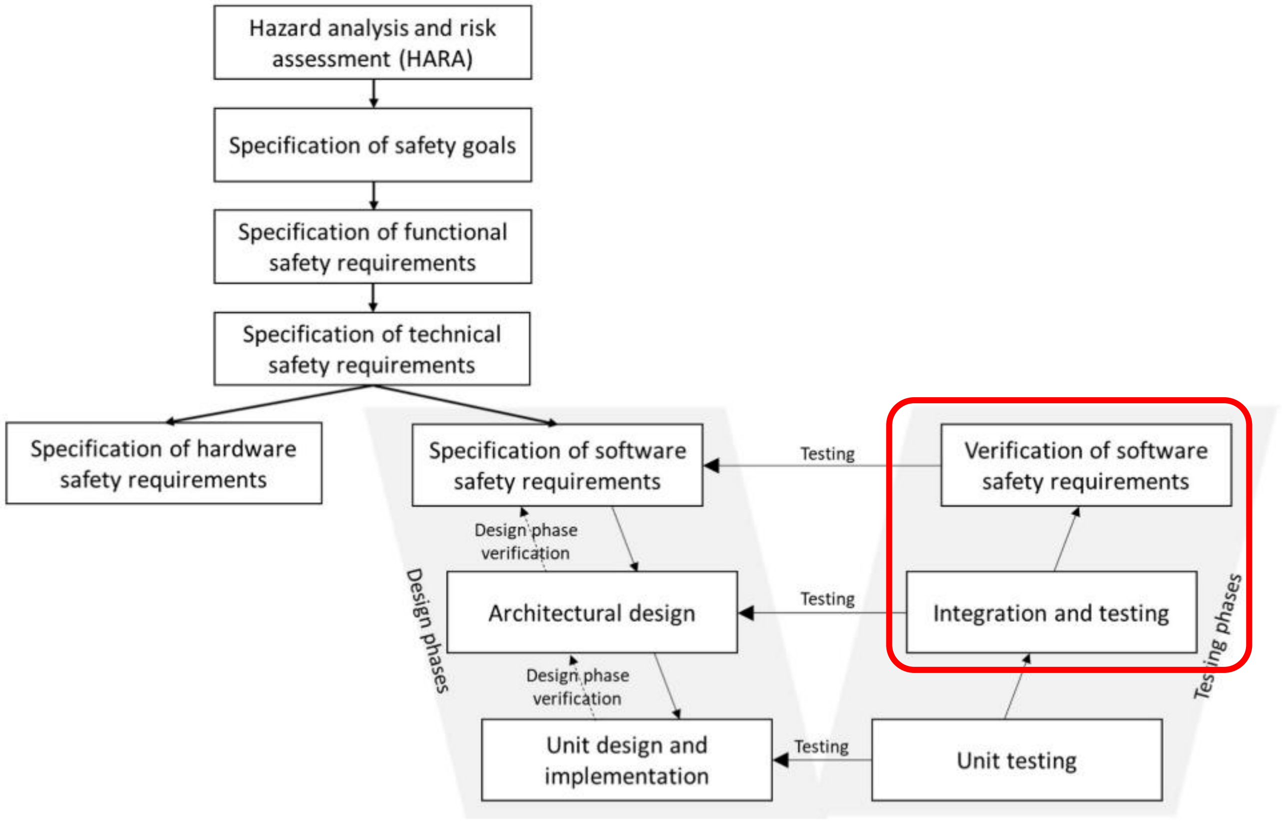


- Model is stochastic
  - Lack of test oracle
- Large input space
  - Unfeasable to cover all scenarios
  - Lack of robustness makes this even more demanding
  - Models are shown to not be robust to small perturbations
  - Feature extraction makes it hard to monitor input data.
- How to identify safety critical cases.
- Interpretability/ Traceability
  - Is wrong prediction = bug?
  - Where is bug?
  - How to correct the bug?
  - Prevents the use of inspection and walkthroughs

Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262

Slide from Lars Tornberg, VCC

Verification and Validation – Challenges for Machine learning systems



- System level
  - Quality assurance of signals from individual models
  - Hard to get error bounds on predictions for many models
- Test under increasing complexity
  - Interpretability
  - Data dependencies
- Future data – distributional shift

Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262

Slide from Lars Tornberg, VCC

Distributional shift

Concrete Problems in AI Safety

Dario Amodei*
Google Brain

Chris Olah*
Google Brain

Jacob Steinhardt
Stanford University

Paul Christiano
UC Berkeley

John Schulman
OpenAI

Dan Mané
Google Brain

- **Robustness to Distributional Shift:** How do we ensure that the cleaning robot recognizes, and behaves robustly, when in an environment different from its training environment? For example, strategies it learned for cleaning an office might be dangerous on a factory workfloor.

arXiv:1606.06565

Distributional shift

Training example

Car, score – 0.998

Example anomalies

Bike, score – 0.958

Person, score – 0.93

Confidence from a deep learning model is not a good proxy for true confidence!

Distributional shift

Step 1: pick starting image ("sloth")

Step 2: pick target class ("race car")

Step 3: create adversarial image by adding carefully chosen imperceptible noise

"sloth"
>99% confidence

"race car"
>99% confidence

Confidence from a deep learning model is not a good proxy for true confidence!

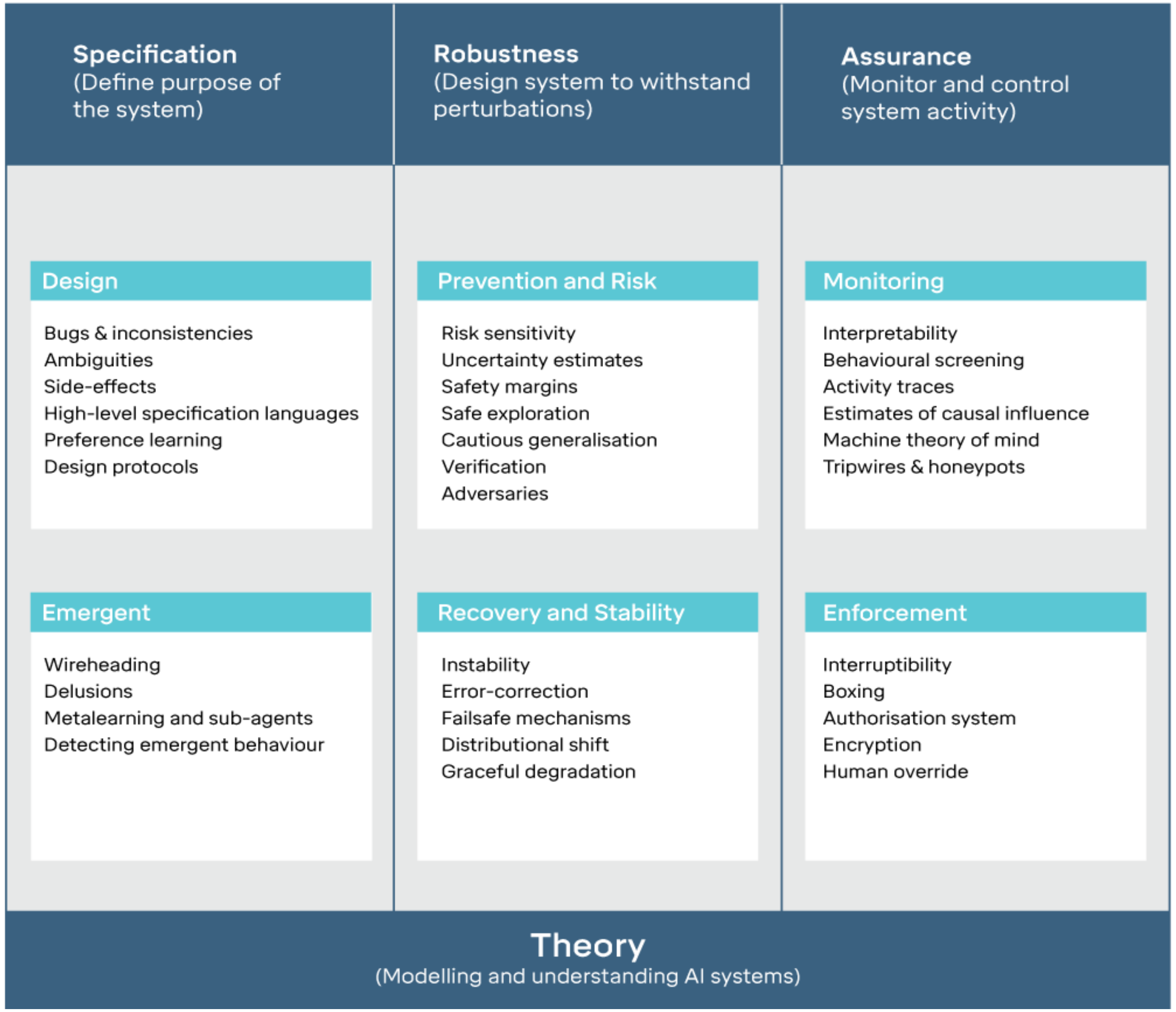
https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1

Verification and Validation – Challenges for Machine learning systems

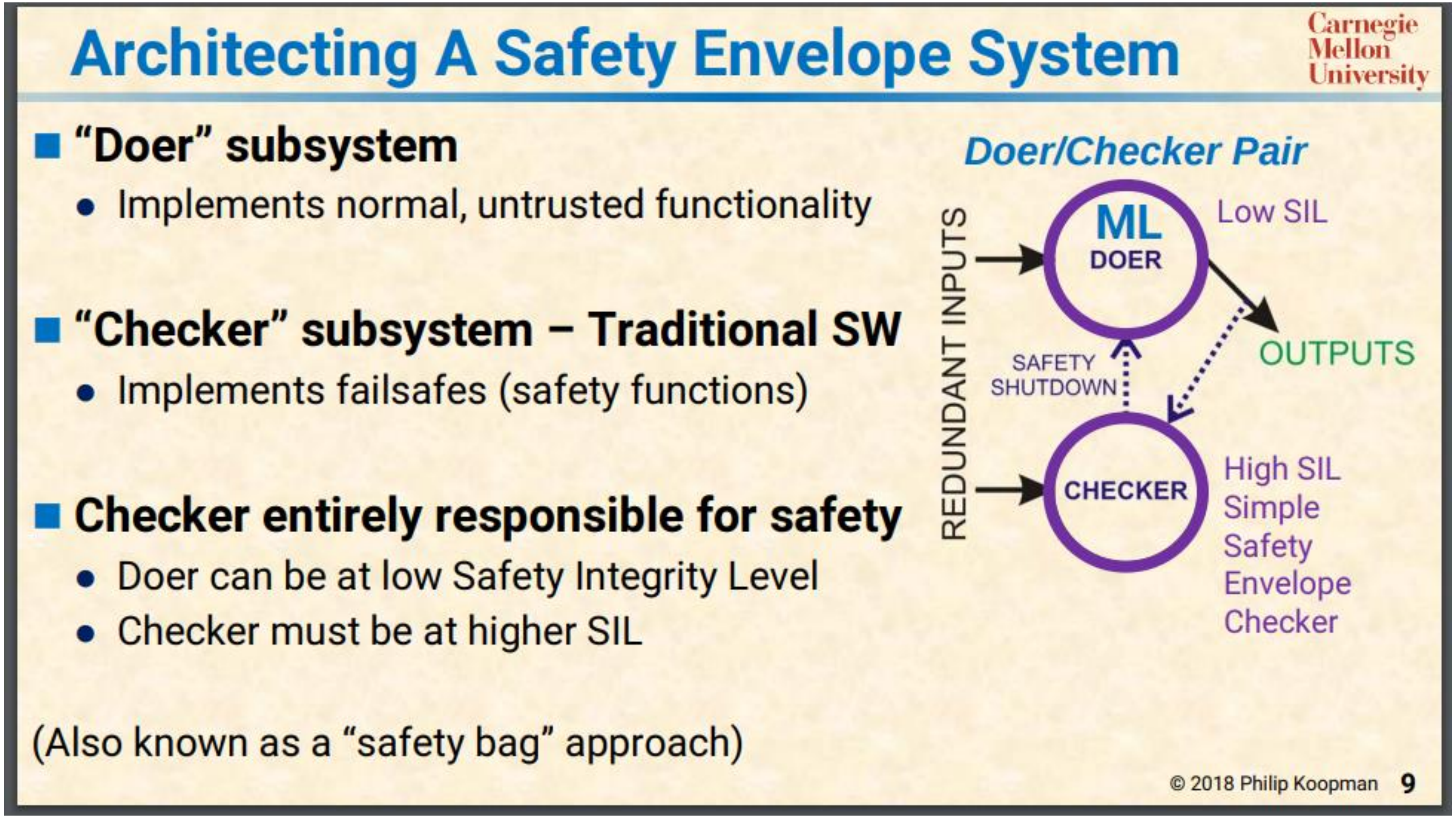
Building safe artificial intelligence: specification, robustness, and assurance

DeepMind Safety Research
Sep 27, 2018 · 9 min read

https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1

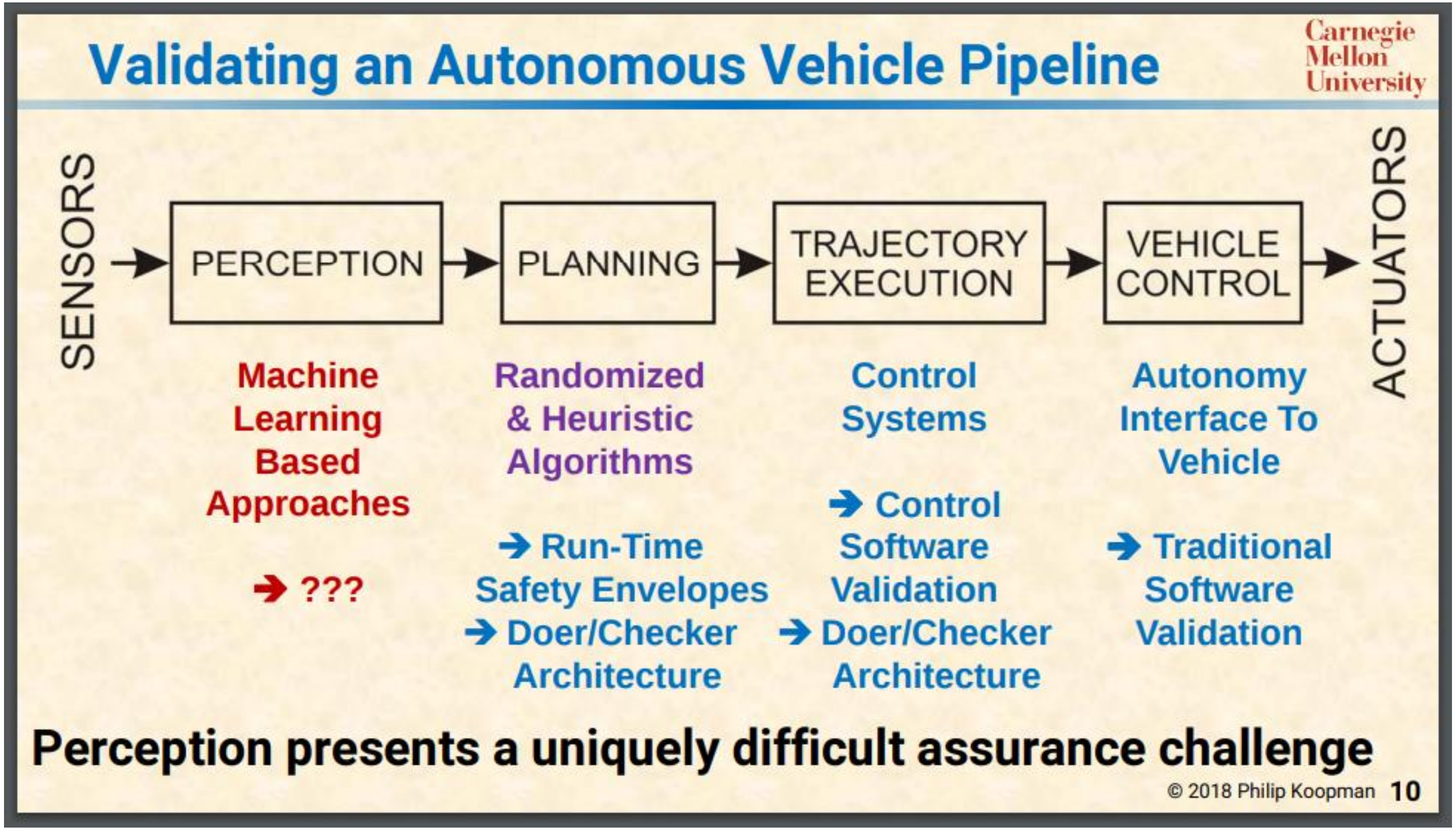


Safety cage



https://users.ece.cmu.edu/~koopman/pubs/koopman18_waise_keynote_slides.pdf

Safety cage

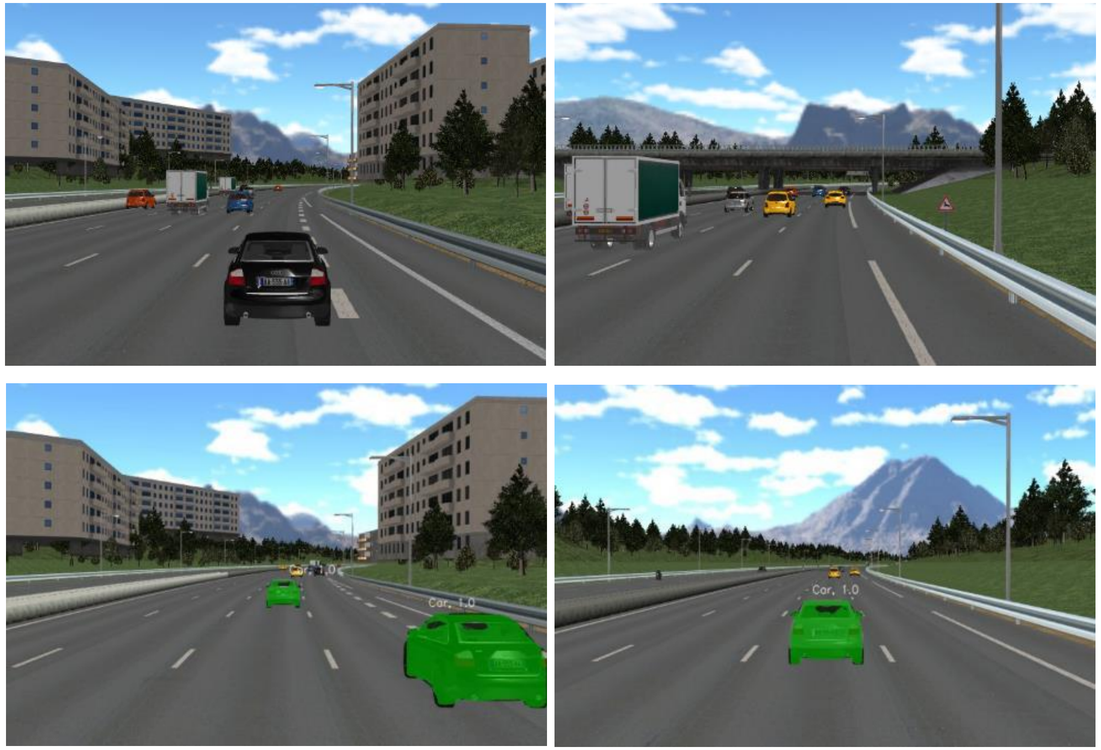


https://users.ece.cmu.edu/~koopman/pubs/koopman18_waise_keynote_slides.pdf

SMILE II – Use cases

Safety cage with Semantic segmentation

- Mask R-CNN trained to detect cars, motorcycles and trucks driving in a highway on a sunny day.
  - Pretrained on COCO dataset
- Data generated from simulation platform: Pro-SiVIC™ from ESI.
  - Training set contains around 3000 of each car, truck and motorcycles
- Safety cage applied by analyzing the neuronal activations of the last fully connected layer of the classifier inside Mask R-CNN
  - The safety cage is not trained (like a neural network).
  - Inputs rejected by the safety cage can be stored and used in further training to improve the AI



Semantic segmentation – outlier data

- Example outlier scenario: Driving in an urban environment

Green mask: accepted by Safety cage
Red mask: rejected by Safety cage

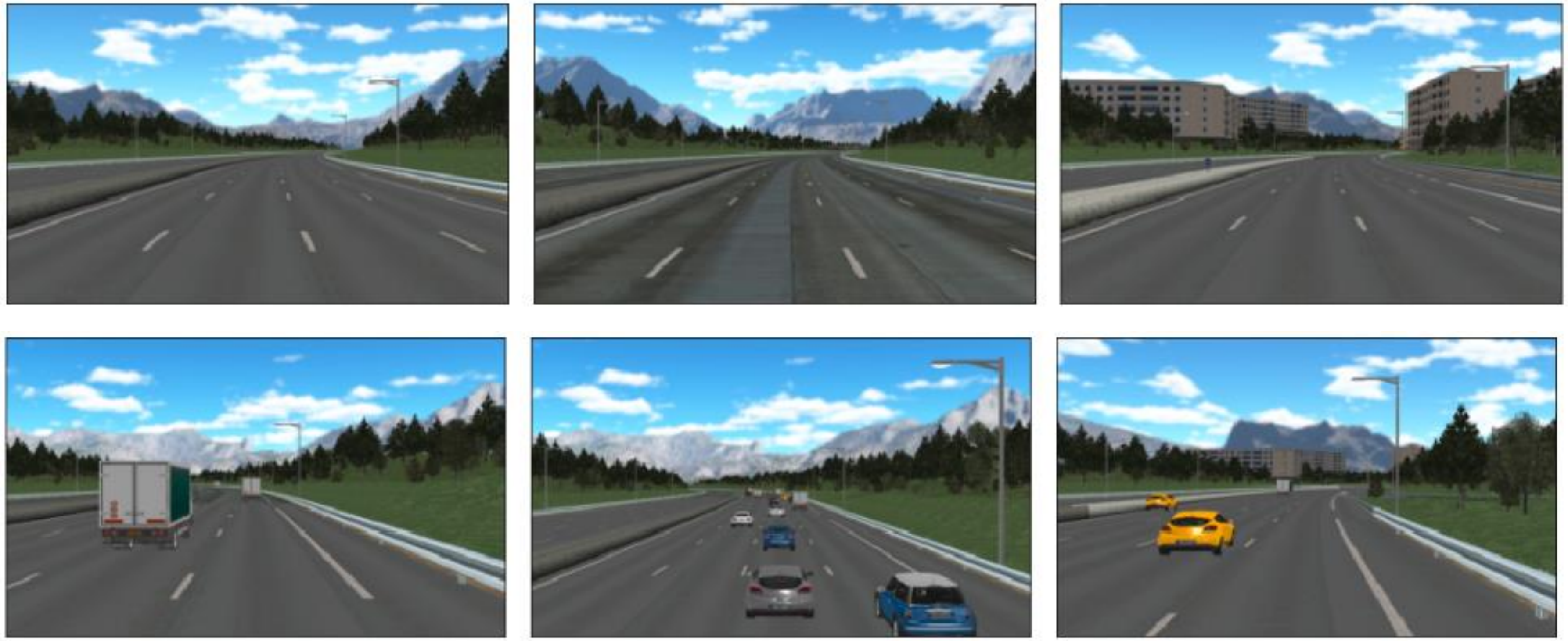
Semantic segmentation with safety cage – demo video

Green mask: accepted by Safety cage
Red mask: rejected by Safety cage

https://youtu.be/M_1gD69-DTQ

Live version shown at VECS 2019 had DDS communication between simulator and the python code (NN + Safety cage)

Safety Cage for perception layer

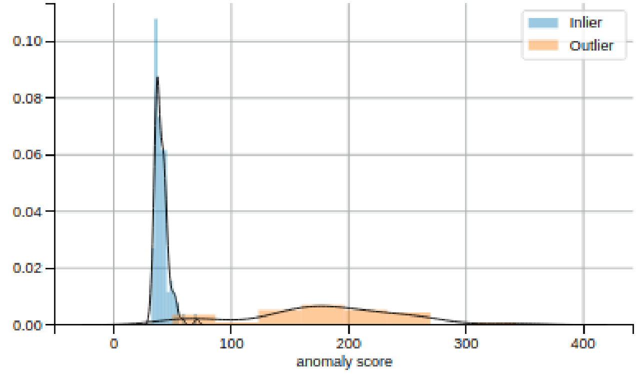


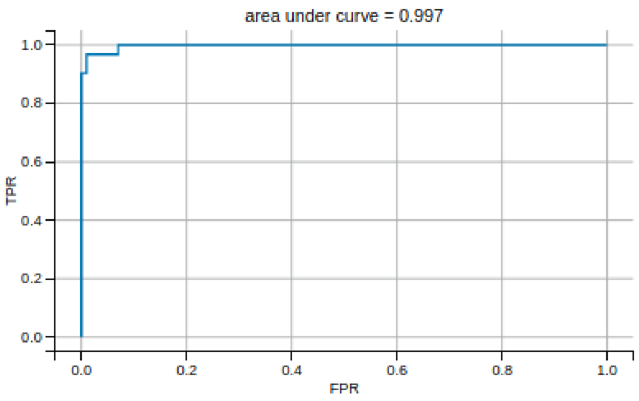
Inlier data

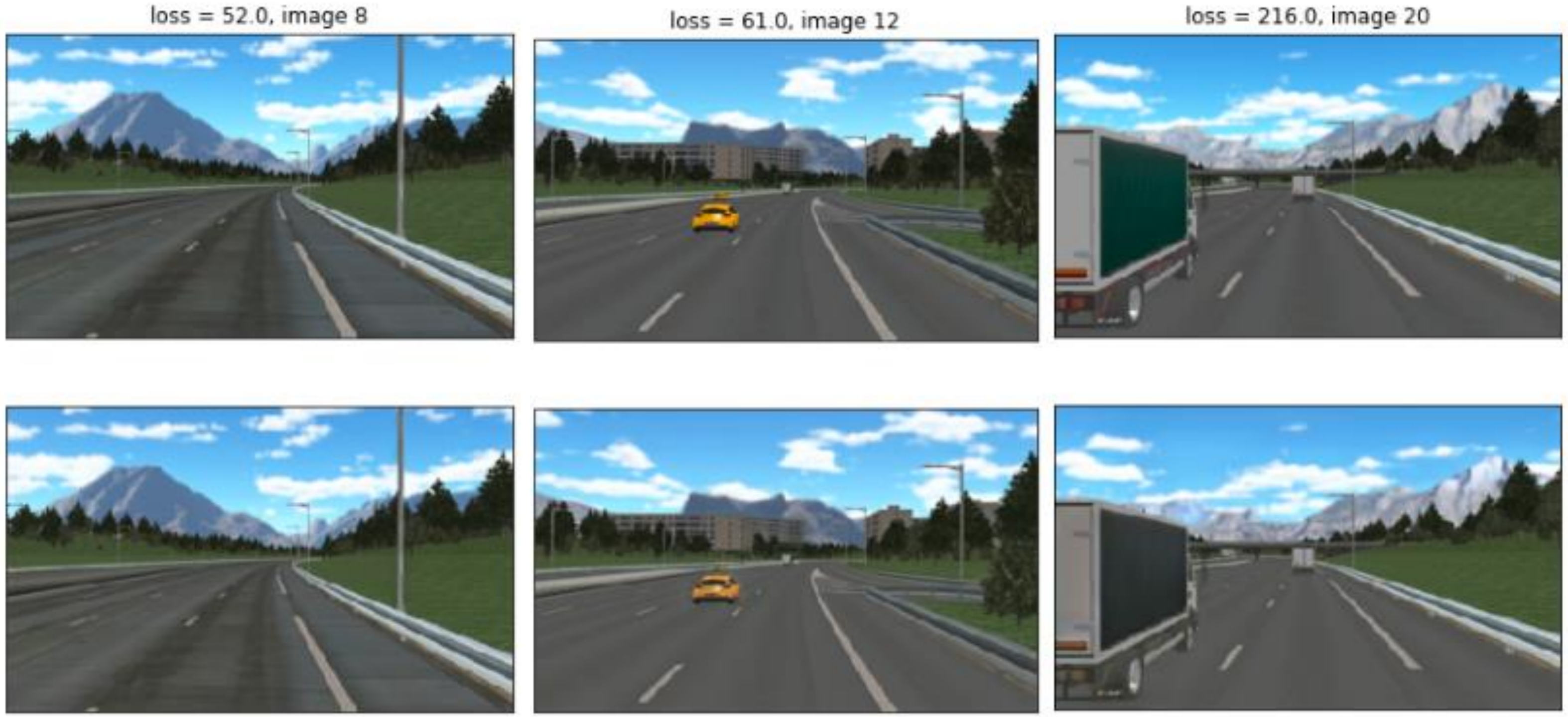
Outlier data

Slide from Lars Tornberg, VCC

Safety Cage for perception layer







Slide from Lars Tornberg, VCC

Evaluation of safety cages

Towards Structured Evaluation of Deep Neural Network Supervisors

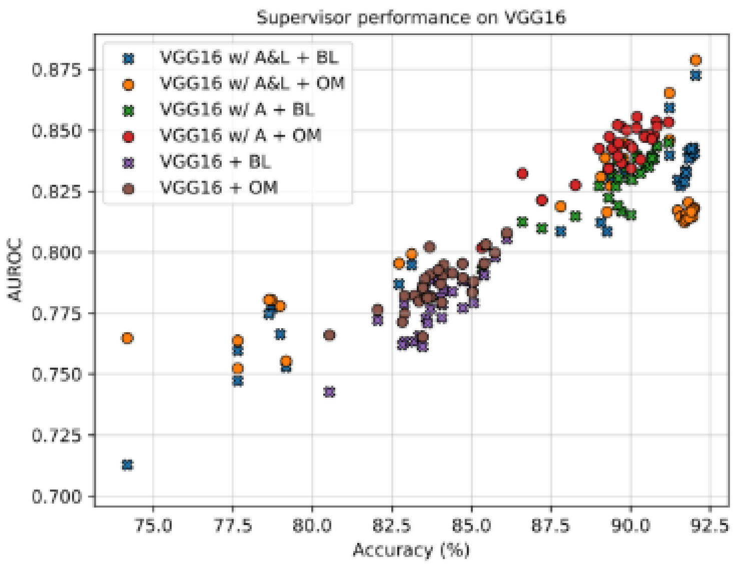
Jens Henriksson*, Christian Berger†, Markus Borg‡, Lars Tornberg§, Cristofer Englund‡, Sankar Raman Sathyamoorthy¶, Stig Ursing*

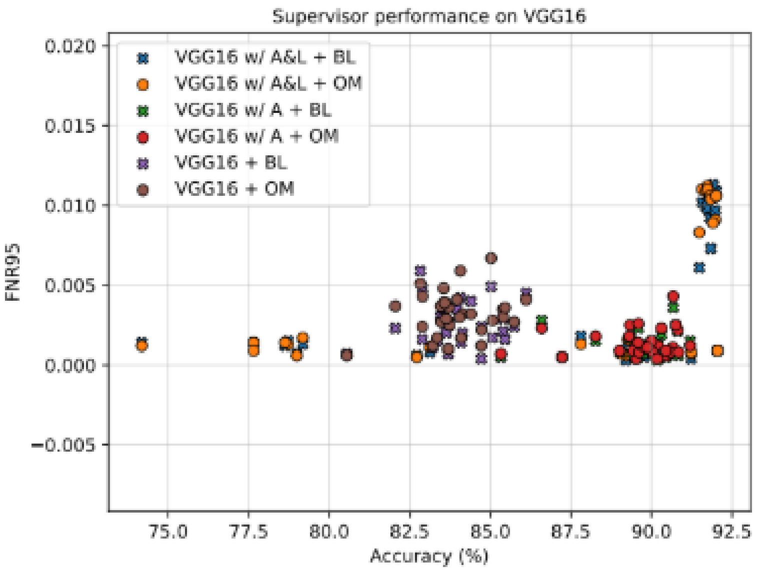
Performance Analysis of Out-of-Distribution Detection on Variedly Trained Neural Networks

Jens Henriksson*, Christian Berger†, Markus Borg‡, Lars Tornberg§, Sankar Raman Sathyamoorthy¶, Cristofer Englund‡

Inlier – CIFAR10

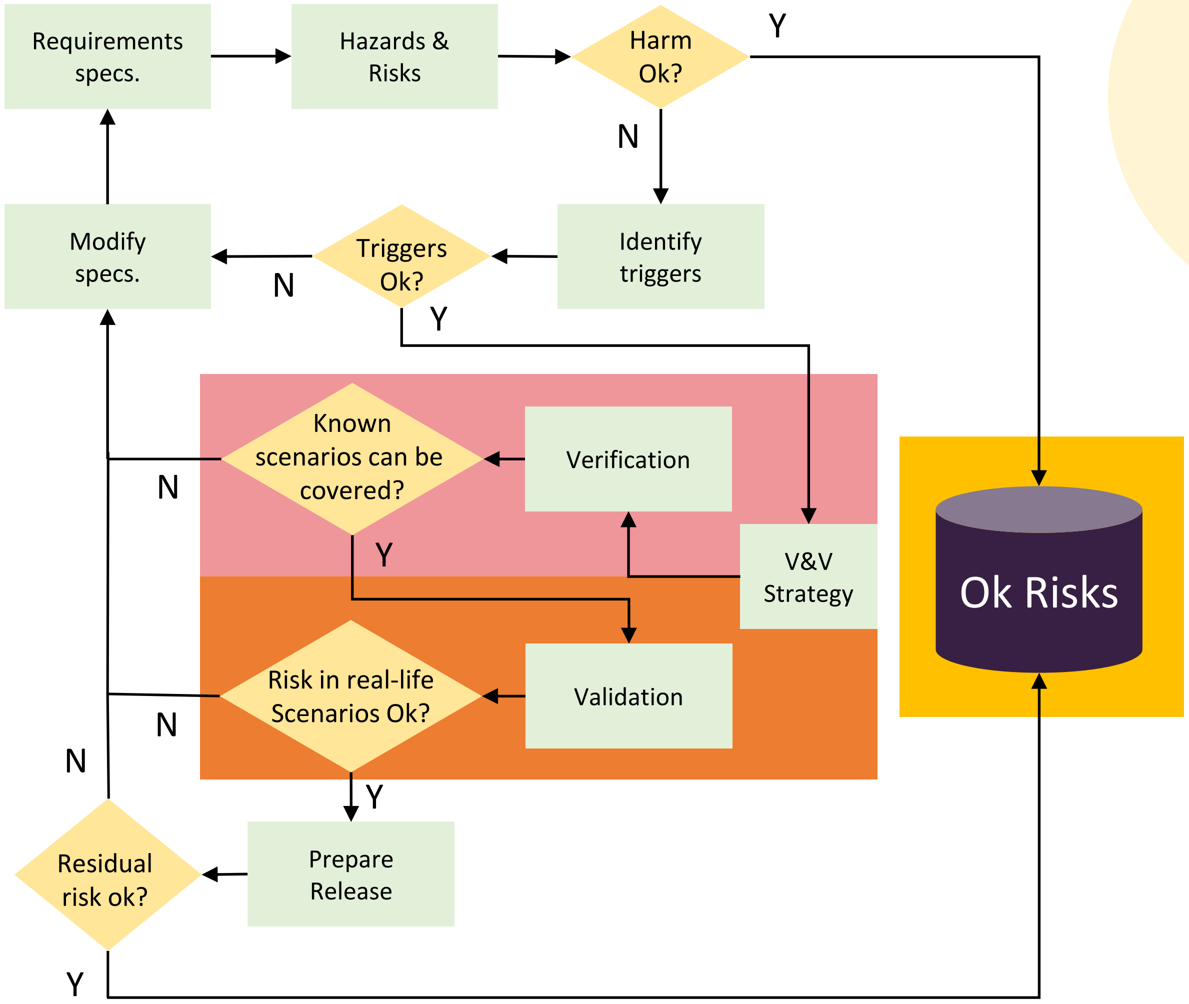
Outlier – TinyImageNet

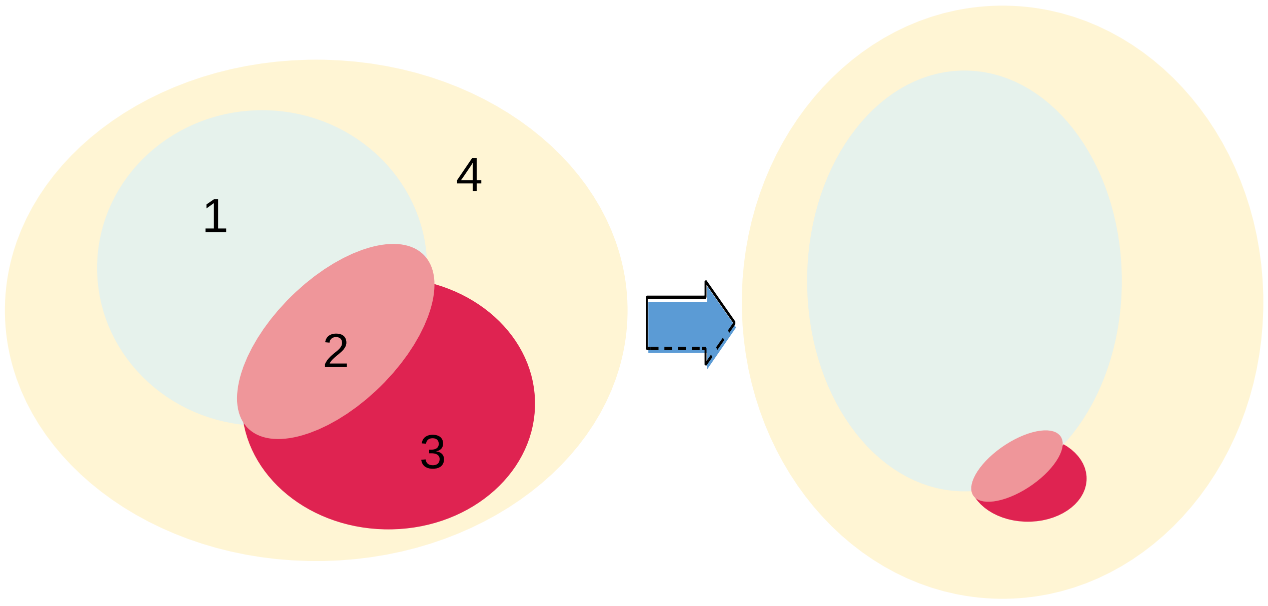


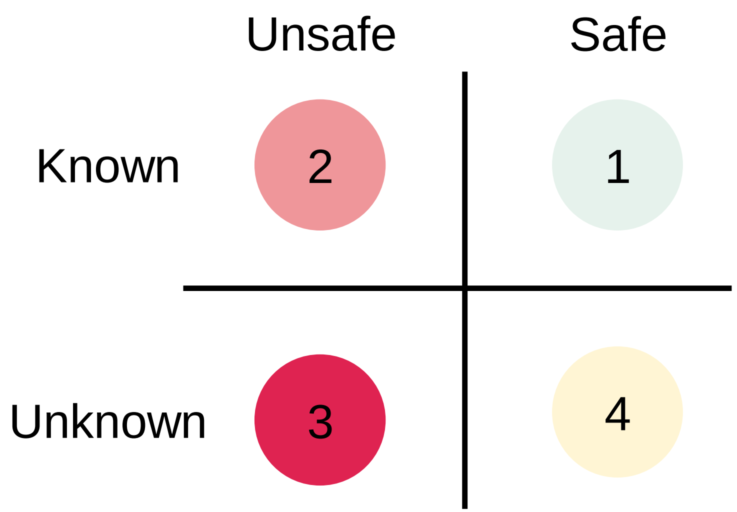


BL = Base Line, OM = OpenMax, A = Data Augmentation, L = Learning Rate
Bendale, A. & Boult, T., Towards open set deep networks, CVPR, 2016

SOTIF



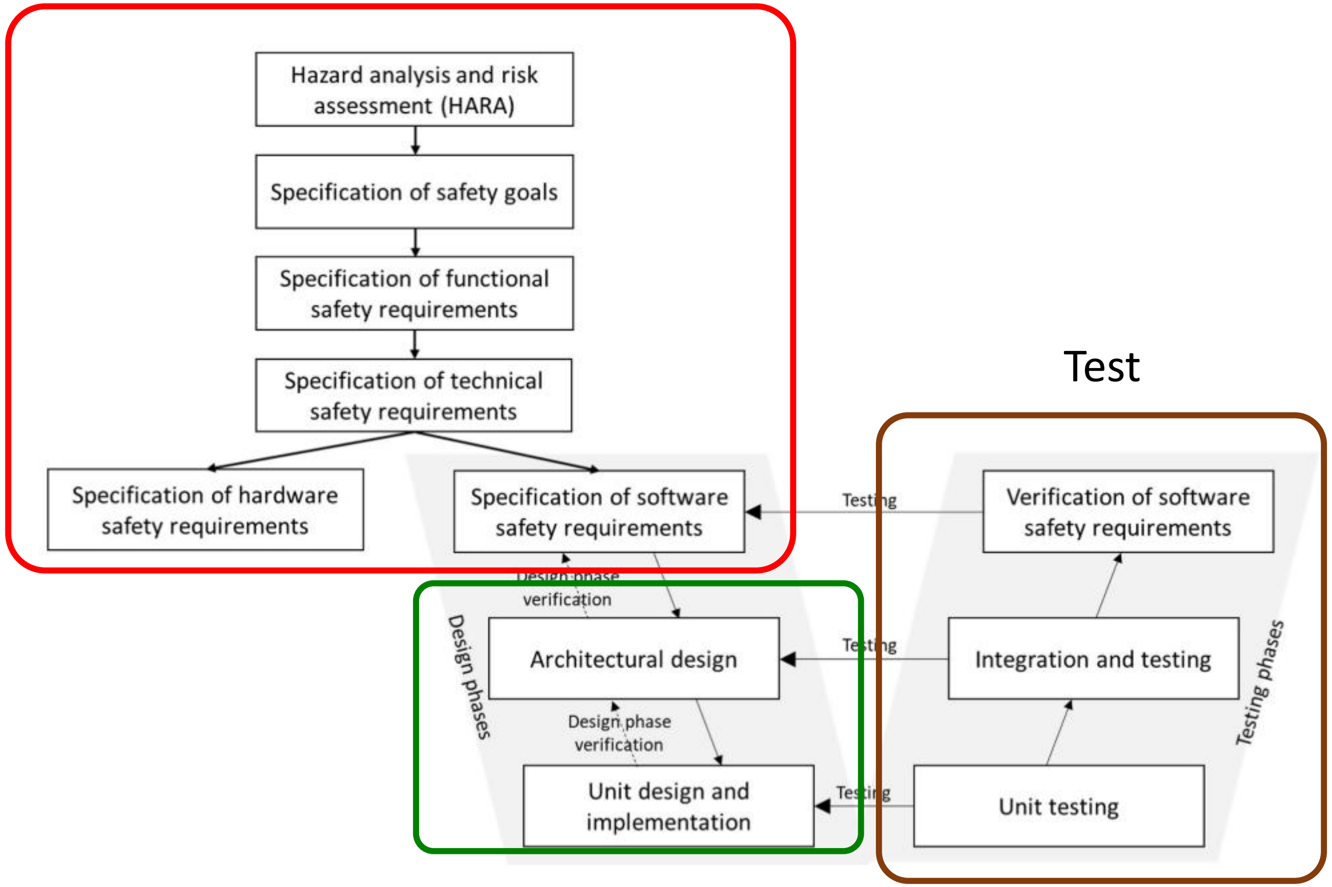




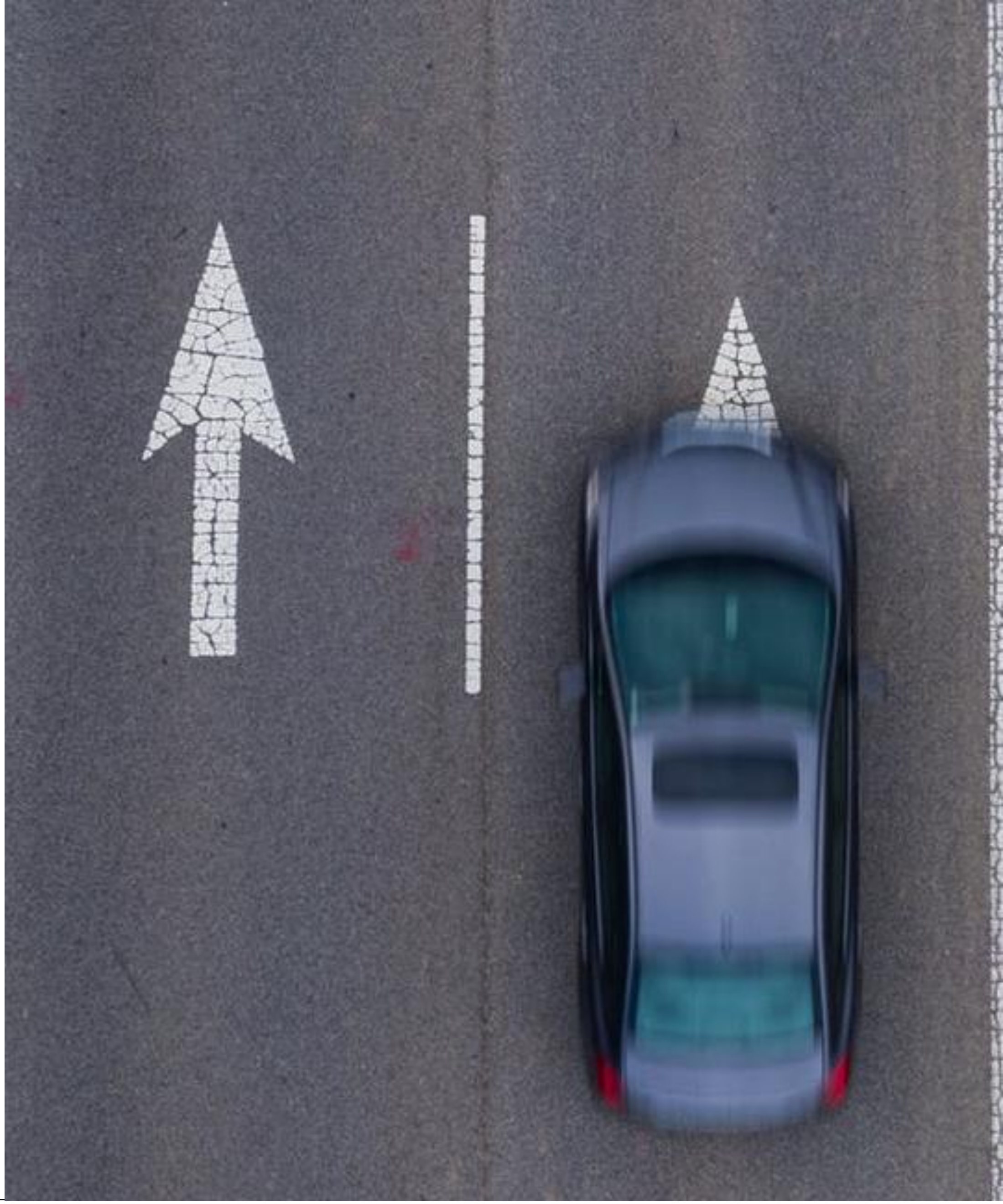
Slide from Markus Borg, RISE

SMILE III

Specification



Design



SMILE III

- **WP2: Architectural design**
  - What components should be encapsulated?
  - Sensor fusion (e.g., lidar, radar, and time series data from the engine)
- **WP3: Safety strategy**
  - Safety cages in the light of the emerging standards ISO/PAS 21448 SOTIF and UL 4600
  - How to act when the safety cage rejects input? (e.g., mitigation strategies, handover to driver, and graceful degradation)
- **WP4: Safety-cage design and optimization**
  - Explore approaches to improve safety cage performance (e.g., Bayesian networks)
  - Strategies to utilize data that was rejected by the safety cage. (e.g., collecting the data for retraining/model updates)
- **WP5: Verification & Validation of the safety cage**
  - Component level testing (e.g., building on the metrics developed in SMILE II)
  - System level testing both using simulators and real applications
    - Demonstrator using Pro-SiVIC (Qrtech)
    - Demonstrator implemented in car on public roads (VCC)
    - Demonstrator implemented in truck in closed setting (AB Volvo)
- **WP6: Novel test methods**
  - Evaluate feasibility of metamorphic testing, search-based testing, mutation testing, DNN coverage testing etc.
  - Meta testing (i.e., testing the testing) using demonstrator implemented using Pro-SiVIC (RISE)

References

- M. Borg. Explainability First! Cousteauing the Depths of Neural Networks to Argue Safety. In *Explainable Software for Cyber-Physical Systems (ES4CPS), Report from the GI Dagstuhl Seminar 19023*, pp. 26-27, 2019.
- M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist. Safely Entering the Deep: A Review of Verification and Validation for Machine Learning and a Challenge Elicitation in the Automotive Industry, *Journal of Automotive Software Engineering*, 1(1), pp. 1-19, 2019. (Borg et al., 2019a)
- M. Borg, S. Gerasimou, N. Hochgeschwender, and N. Khakpour. Explainability for Safety and Security. In *Explainable Software for Cyber-Physical Systems (ES4CPS), Report from the GI Dagstuhl Seminar 19023*, pp. 15-18, 2019. (Borg et al., 2019b)
- J. Henriksson, C. Berger, M. Borg, L. Tornberg, C. Englund, S. Sathyamoorthy, and S. Ursing. Towards Structured Evaluation of Deep Neural Network Supervisors, In *Proc. of the 1st IEEE International Conference on Artificial Intelligence Testing (AITest)*, pp. 27-34, 2019. (Henriksson et al., 2019a)
- J. Henriksson, C. Berger, M. Borg, L. Tornberg, S. Sathyamoorthy, and C. Englund. Performance Analysis of Out-of-Distribution Detection on Various Trained Neural Networks, To appear in *Proc. of the 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2019. (Henriksson et al., 2019b) ***BEST PAPER AWARD***
- J. Henriksson, M. Borg, and C. Englund. Automotive Safety and Machine Learning: Initial Results from a Study on How to Adapt the ISO 26262 Safety Standard, In *Proc. of the 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS)*, 2018.
- E. Kratz, B. Duran, C. Englund. Novel Scenario Detection in Road Traffic Images. Prepared for submission. (E. Kratz 2019a)
- A. Vogelsang and M. Borg. Requirements Engineering for Machine Learning: Perspectives from Data Scientists. *To appear in Proc. of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, 2019.
- Patent Application No. 19196450.1 - Automatic Detection of Outlier Objects in Images for AD/ADAS
- Abdallah Alabdallah: Thesis Report, *Human Understandable Interpretation of Deep Neural Networks Decisions Using Generative Models*, 2019.
- Erik Kratz. *Novel scenario detection in road traffic images*. Examensarbete - Institutionen för elektroteknik, Chalmers tekniska högskola. 2019. https://hdl.handle.net/20.500.12380/256655 (E. Kratz 2019b)
- S. Gao and Y. Tan. Paving the Way for Self-driving Cars - Software Testing for Safety-critical Systems Based on Machine Learning: A Systematic Mapping Study and a Survey, MSc thesis, Blekinge Institute of Technology, 2017. http://urn.kb.se/resolve?urn=urn:nbn:se:bth-15681