Trustworthy AI

Bringing Together AI, Technology, and Ethics

Philipp Slusallek

German Research Center for Artificial Intelligence (DFKI)

Saarland University

Excellence Cluster Multimodal Computing and Interaction (MMCI)

European High-Level Expert Group on AI

DFKI: An Overview

German Research Center for Artificial Intelligence (DFKI)

- **Motto**
  - „Computer with Eyes, Ears, and Common Sense“
- **Overview**
  - Largest AI research center worldwide (founded in 1988)
  - Germany’s leading research center for innovative SW technologies
  - 6 sites in Germany
    - Saarbrücken, Bremen, Kaiserslautern; Berlin, Osnabrück, Oldenburg
  - 18 research areas, 10 competence centers, 7 living labs
  - More than 575 core research staff (>1050 total)
  - Revenues of ~50 M€ (2018)
  - More than 90 spin-offs

Germany Has a Head-Start

DFKI: The World´s Largest Center for Research & Application in AI

CLAIRE Confederation of Laboratories for Artificial Intelligence Research in Europe

Excellence across all of AI.
For all of Europe.
With a human-centred focus.
(more info at https://claire-ai.org)

CLAIRE Offices:

- Den Haag (HQ), **Saarbrücken**, Rome, Prag, Oslo, Paris, …

CLAIRE Core Team:

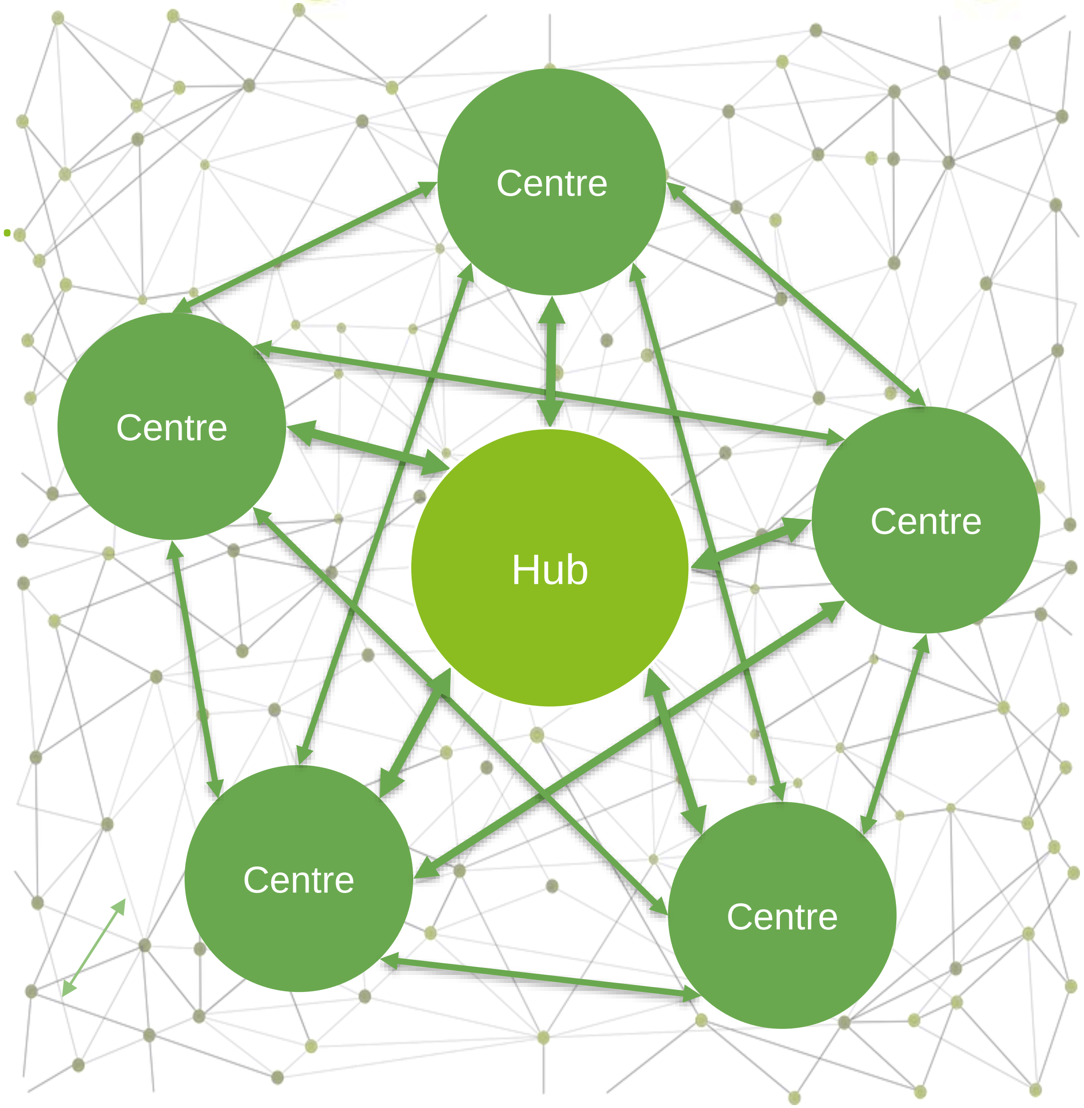
- Philipp Slusallek
  DFKI (DE)
- Holger Hoos
  Leiden University (NL)
- Morten Irgens
  Oslo Metropolitan University (NO)

CLAIRE Supporters:

- >3200 AI experts and stakeholders
- Research Orgs: DFKI, FBK, Inria, TNO, …
- AI-Orgs: EurAI, AAAI, ESA
- EU-Gov.: **BE, CZ, ES, FI, GR, IT, LU, NL, SK**
- WIP: Add Industry & innovation networks

CLAIRE Confederation of Laboratories for Artificial Intelligence Research in Europe

Excellence across all of AI.
For all of Europe.
With a human-centred focus.
(more info at https://claire-ai.org)



CLAIRE Vision for European AI:

1) Network of Research Labs (~330)

2) Network of Centers of Excellence

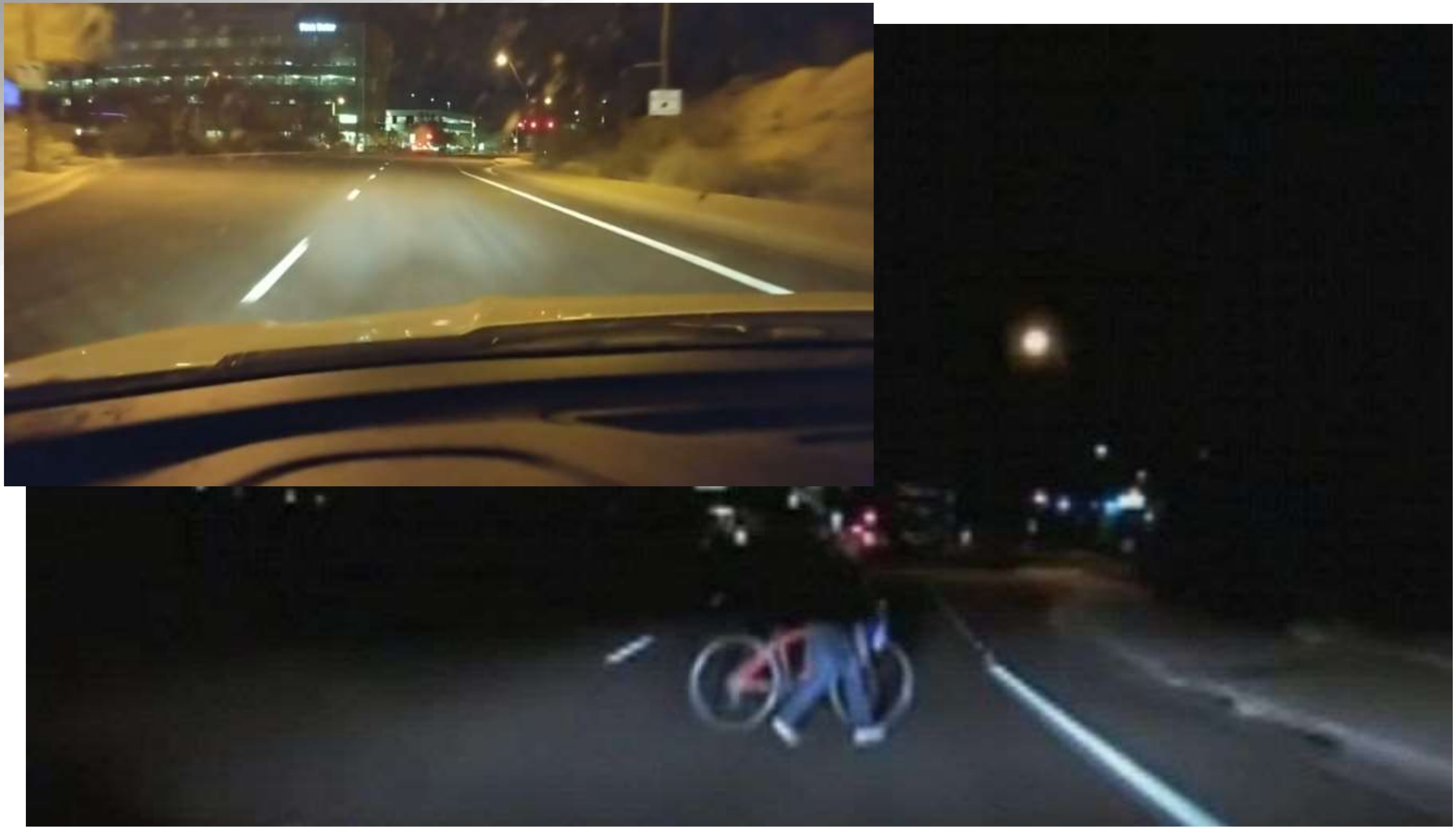
3) European AI Hub ("CERN for AI")

- Focal point for exchange and interaction
- World-leading infrastructure & support
- Global attractor for AI talent
- **Symbol for European excellence & ambition in AI**

Digital Reality for Trustworthy AI: Using Synthetic Data to Train & Validate Autonomous Systems

(using autonomous driving as an example)

Why Do We Need Training and Validation via Synthetic Data?



Autonomous Systems:
The Problem

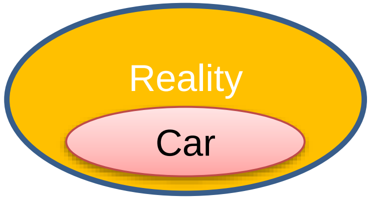
- **Our World is extremely complex**
  - Geometry, Appearance, Motion, Weather, Environment, …
- **Systems must make accurate and reliable decisions**
  - Especially in ***Critical Situations***
  - Increasingly making use of (deep) machine learning
- **Learning of critical situations is essentially impossible**
  - Often little (good) data even for “normal” situations
  - Critical situations rarely happen in reality – per definition!
  - Extremely high-dimensional models

➔Goal: Scalable Learning from *synthetic* input data

- Continuous benchmarking & validation (“Virtual Crash-Test“)

Reality

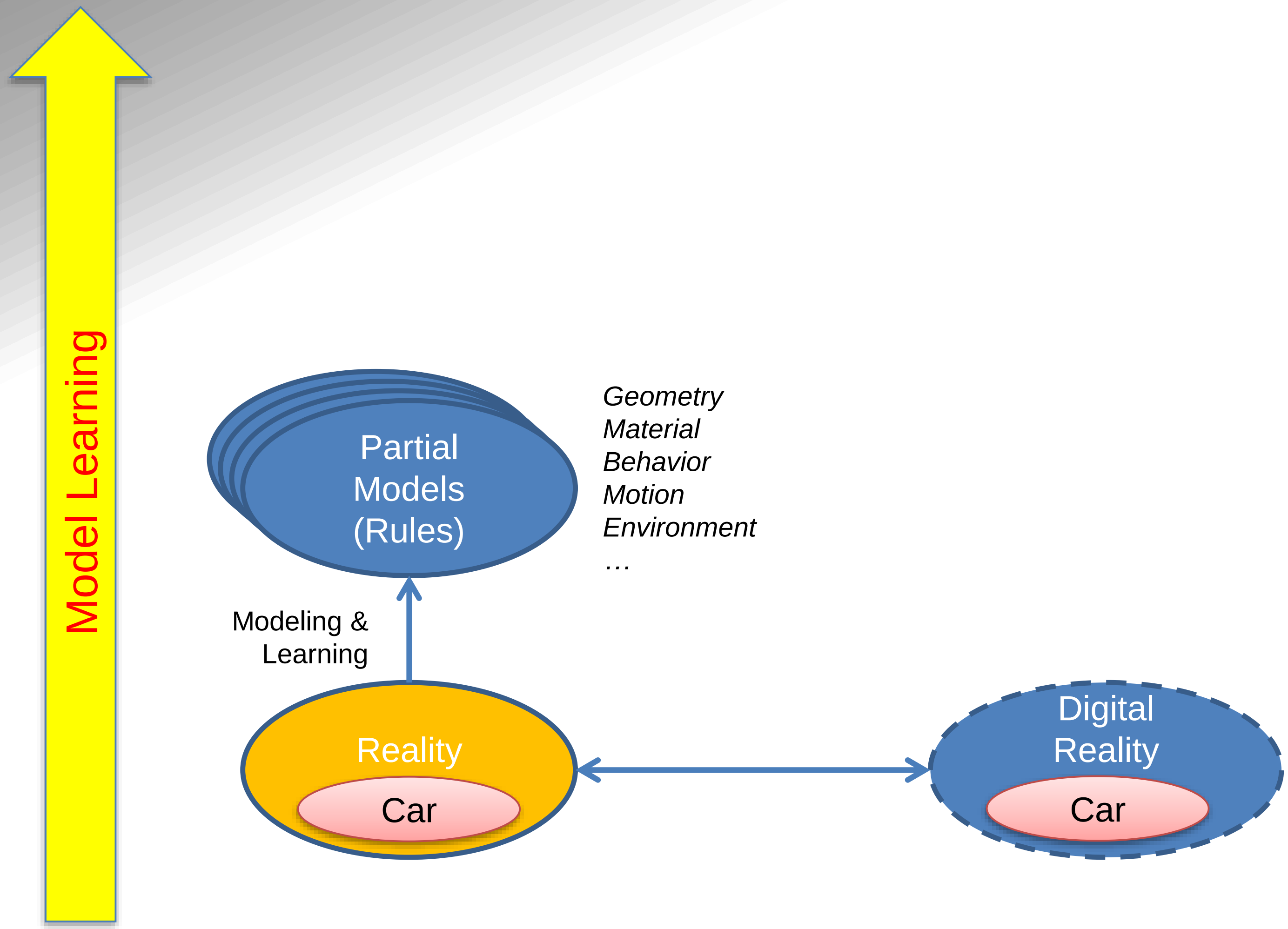
- **Training and Validation in Reality**
  - E.g. driving millions of miles to gather data
  - Difficult, costly, and non-scalable



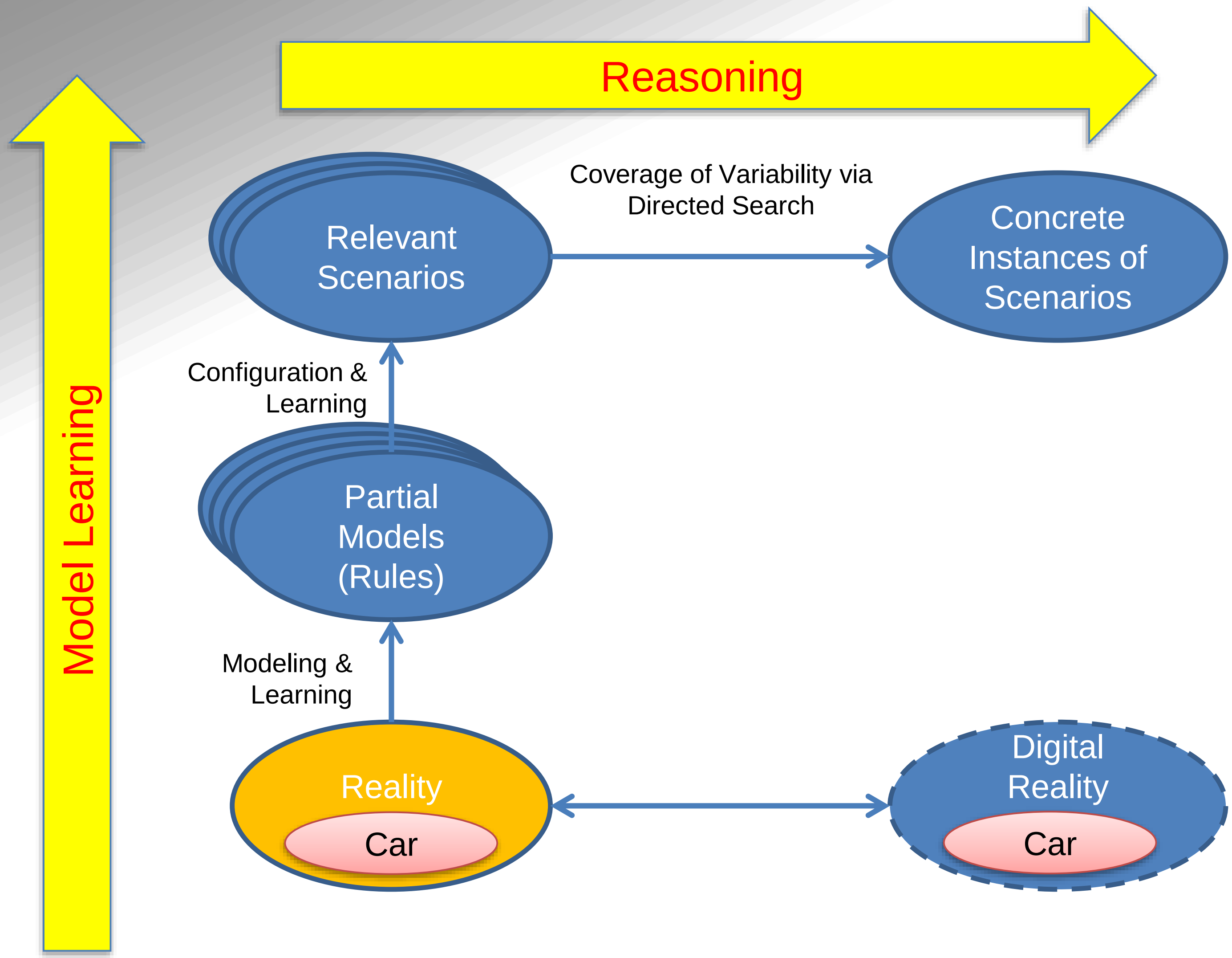
Digital Reality

- **Training and Validation in the *Digital Reality***
  - Arbitrarily scalable (given the right platform)
  - But: Where to get the models and the training data from?

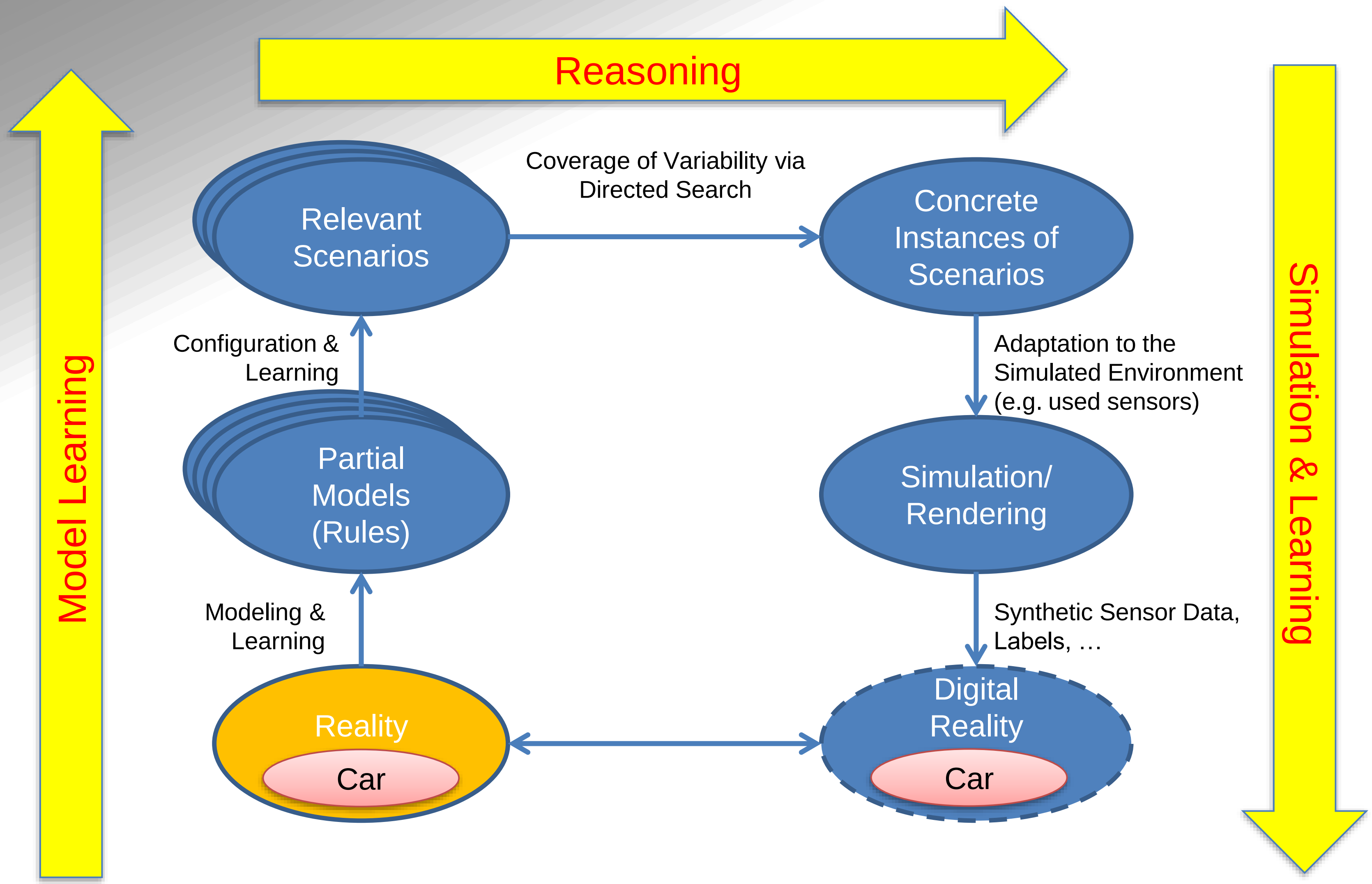
Digital Reality: Learning



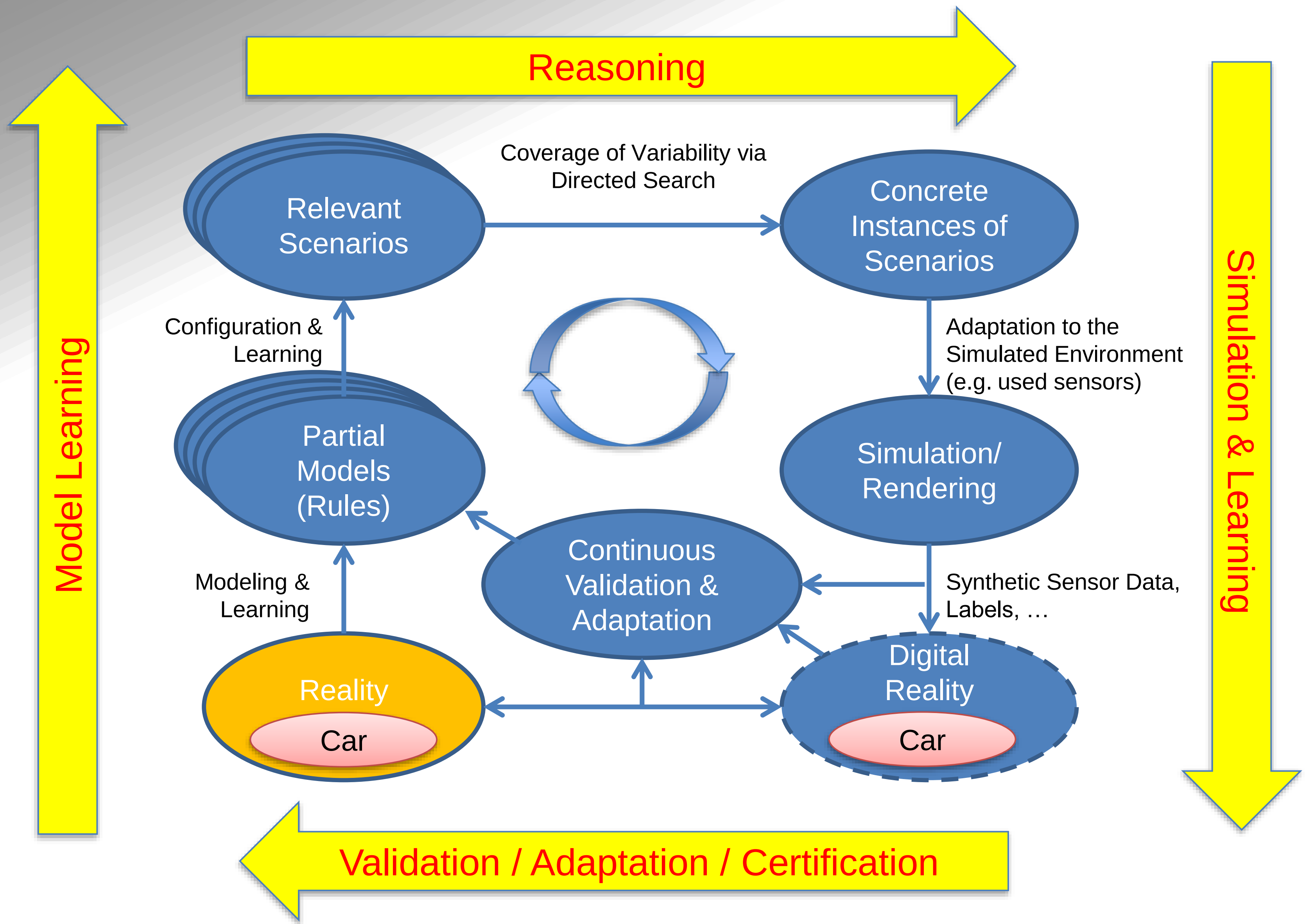
Digital Reality: Reasoning



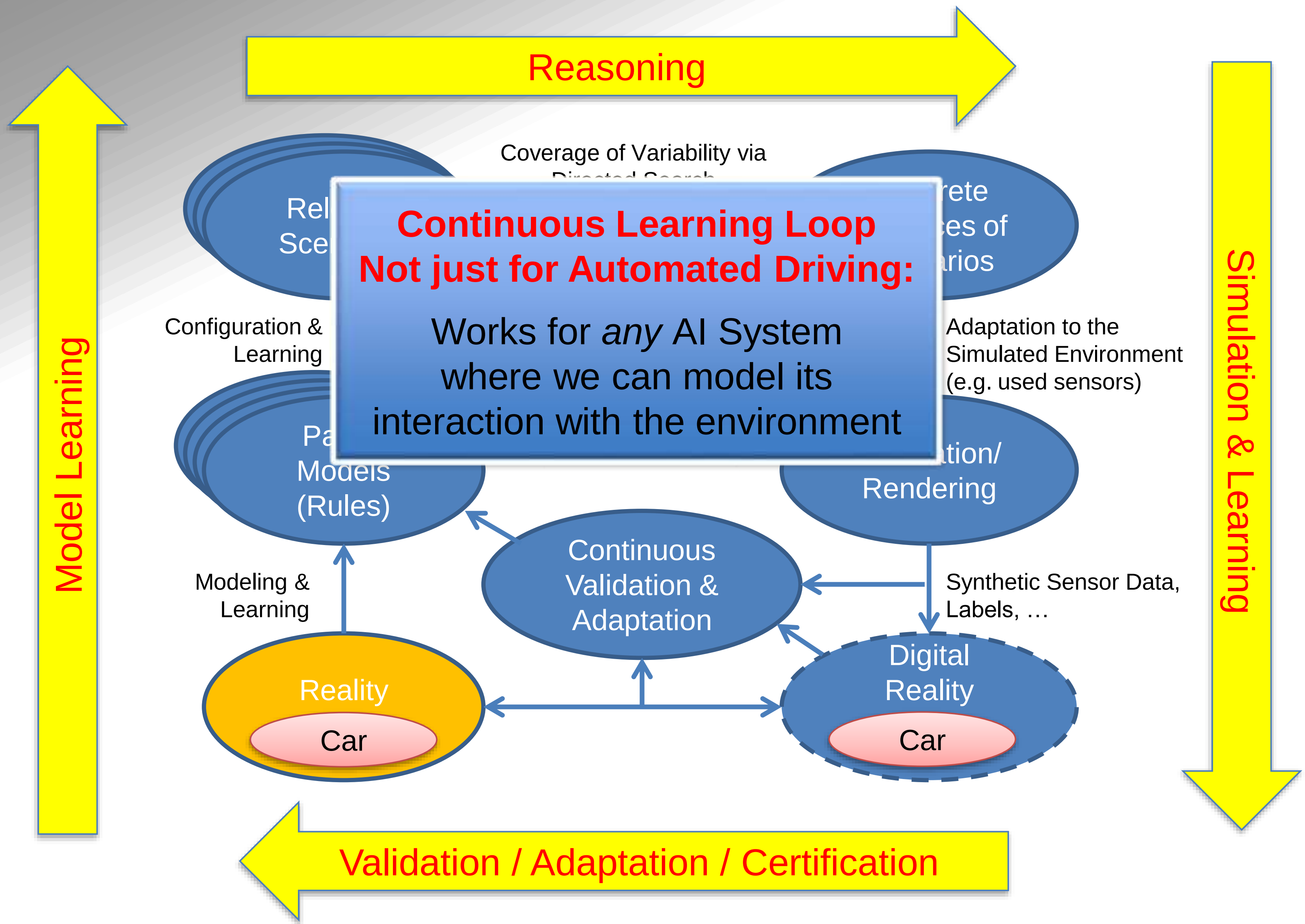
Digital Reality: Simulation



Digital Reality: Validation/Adaptation

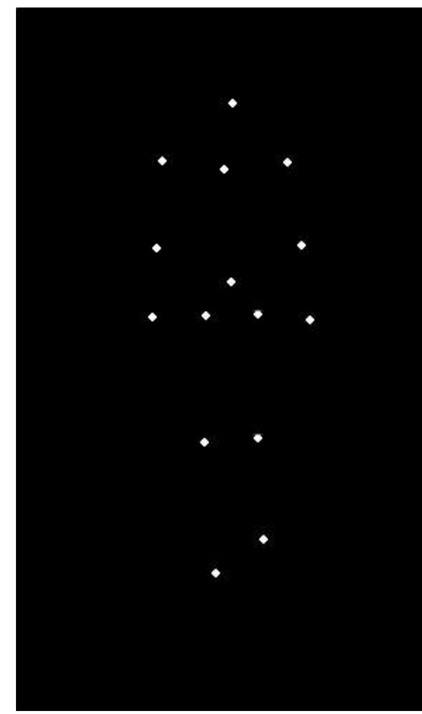


Digital Reality: Continuous Learning



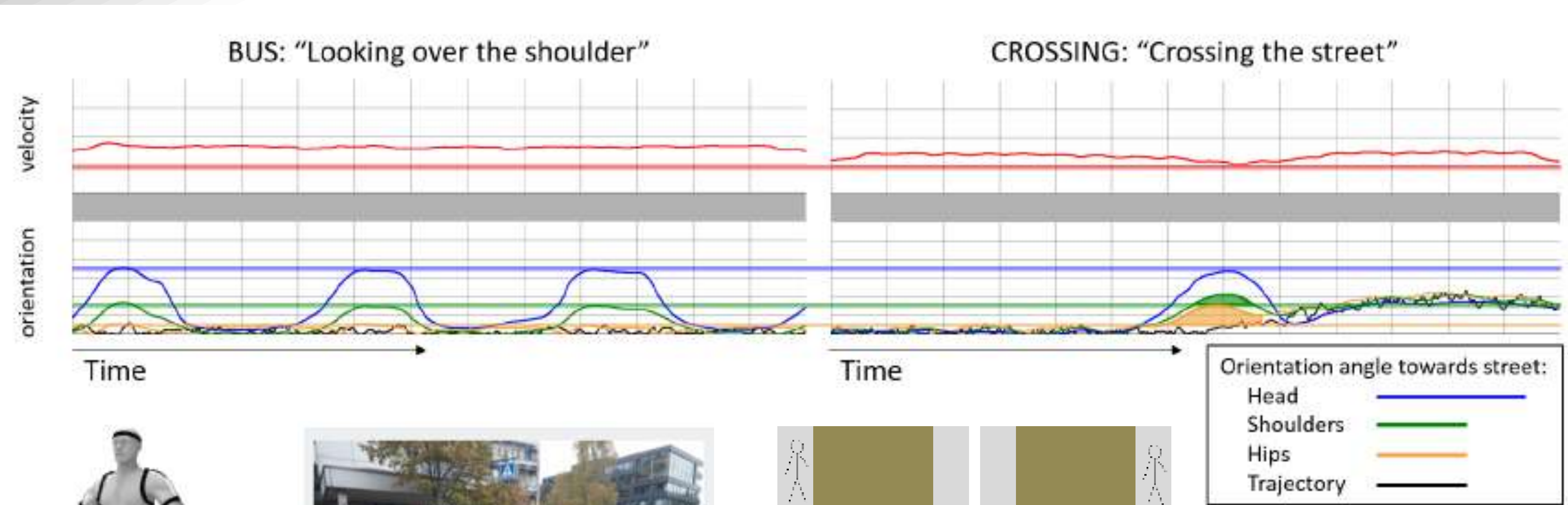
Challenge: Better Models of the World (e.g. Pedestrians)

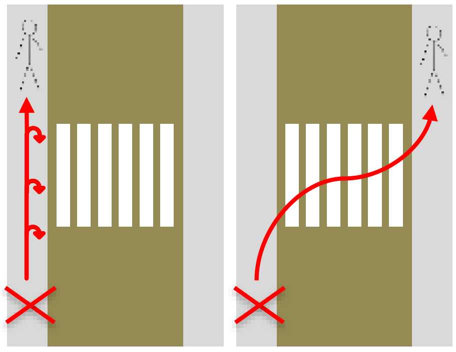
- **Long history in motion research (>40 years)**
  - E.g. Gunnar Johansson's Point Light Walkers (1974)
  - Significant interdisciplinary research (e.g. psychology)
- **Humans can easily discriminate different styles**
  - E.g. gender, age, weight, mood, ...
  - Based on minimal information
- **Can we teach machines the same?**
  - Detect if pedestrian will cross the street
  - Parameterized motion model & style transfer
  - Predictive models & physical limits



Challenge: Pedestrian Motion

- **Characterizing Pedestrian Motion**
  - Clear motion differences when crossing the street





Bus Crossing

Challenge:
Verification ~~versus~~ and Validation

- **Verification (Top-Down)**
  - Strict formal models and exact mathematical proofs
  - But: Limited expressiveness and complexity

Find path between both worlds, e.g.

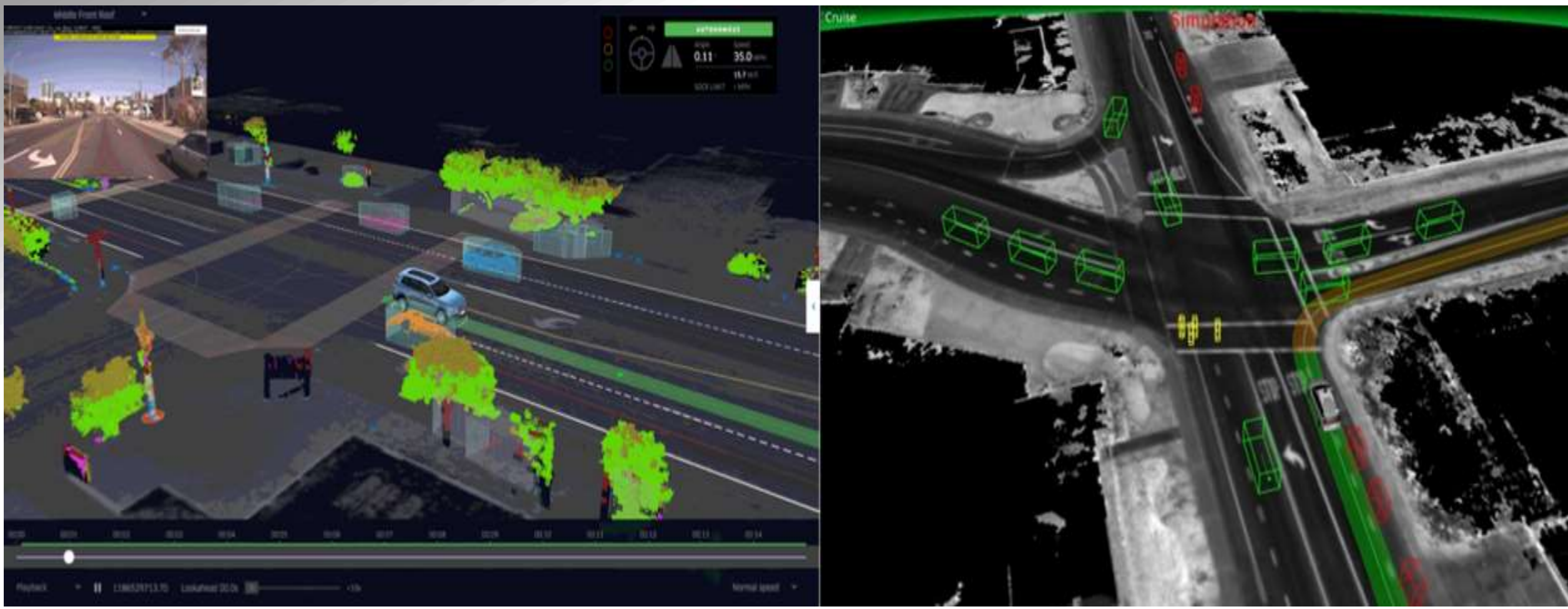
- Identifying potential critical situations
- Limiting the search space for testing

- **Validation (Bottom-Up)**
  - Rich and flexible models close to physical reality
  - But: No completeness and only statistical results

Towards Explainable & Trustworthy AI

Integrating reasoning and learning

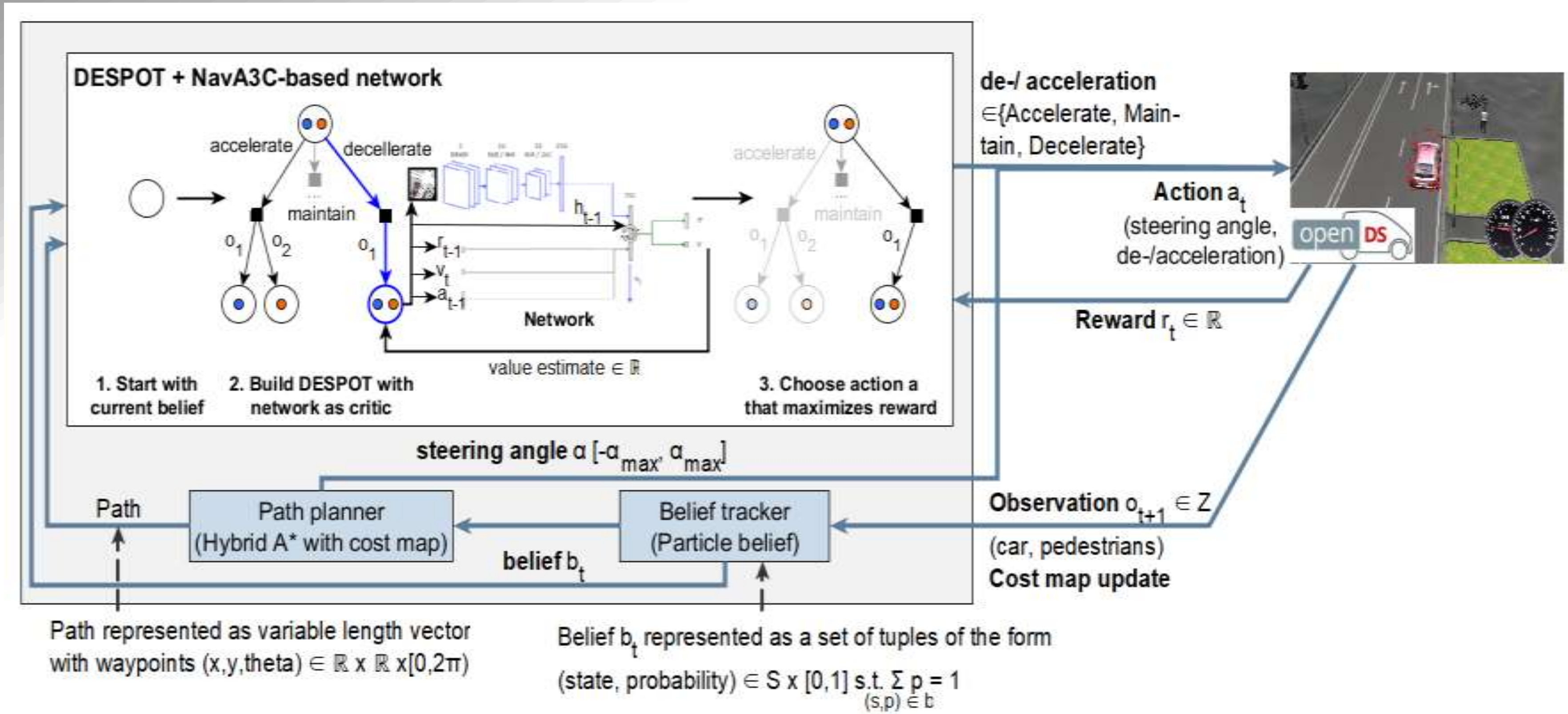
Domain:
Highly Automated Driving



While decision-making and maneuver planning is likely to be rule-based (more general: symbolic), the **perception part** in autonomous driving is expected to be **dependent on Deep Neural Nets**. Also **Interactive machine learning** approaches will contribute to the solution.

At the same time, the ISO 262626 extension **SOTIF (Safety of the Intended Functionality)** is likely to be the standard use in **validation and homologation** of self-driving cars. It requires driving automation functions of self driving cars to be **diagnosable**.

HyLEAP: System Architecture



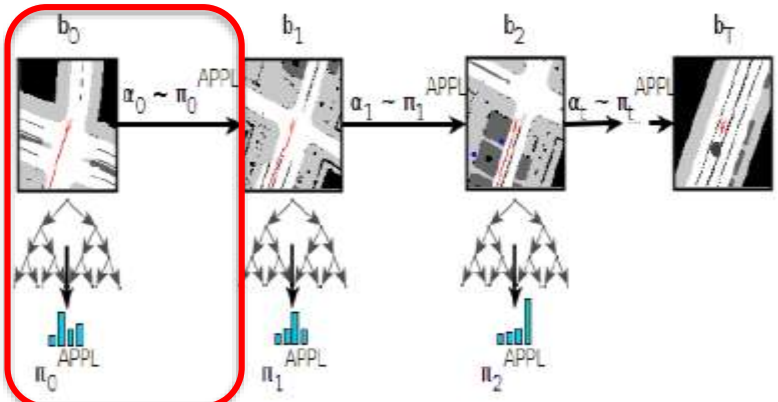
- Integrates **online approximated POMDP planner** DESPOT with **deep reinforcement learning network** NavA3C
- Trains **network to evaluate action policy of planner** (hybrid actor-critic system)

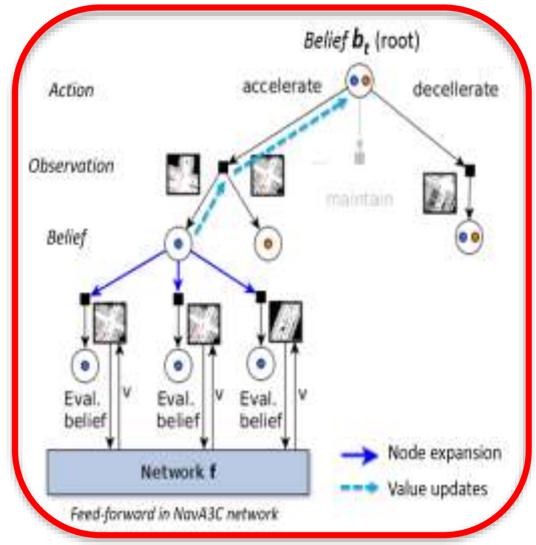
➔ ***Experience-based online navigation action planning*** ➔ ***Better XAI***

HyLEAP Training

For each traffic scene:

(1) For each simulation step t = 0..T: **Planning and execution of action by DESPOT** with its **belief tree construction guided by NavA3C network**
(Estimated NavA3C policy value used as heuristic upper bound on reward for belief node exp.)





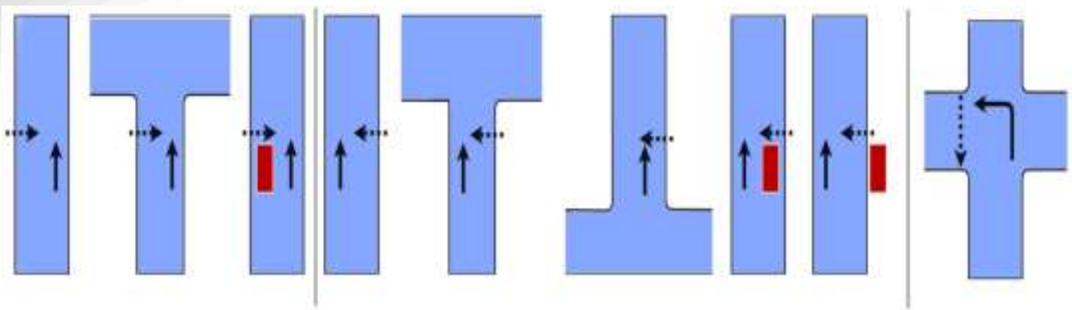
(2) **Train / Update NavA3C network on traffic scene**

→ Minimize mean squared error of own action policy *and* cross-entropy loss between both action policies. Update of network weights via SGD over accumulated gradients of loss L

F. Pusse, M. Klusch (2019): Hybrid Online POMDP Planning and Deep Reinforcement Learning for Safer Self-Driving Cars. Proc. 30th IEEE International Intelligent Vehicles Symposium

Experimental Evaluation

- **Benchmark OpenDS-CTS**: More than 37.000 scenes of 9 types of 3.200 real car-pedestrian accident scenarios based on **German In-Depth Accident Study (GIDAS)** simulated in OpenDS; **HyLEAP available@github**

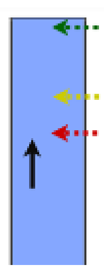


Legend:
Pedestrian
Car
Obstacle

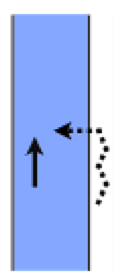
Drive 100 – 112m,
Start speed: 0, max 50km/h

- DRL network **trained over all 9 scenario types** on NVIDIA DGX-1 super computer

- **Variety of other scenarios for testing:**

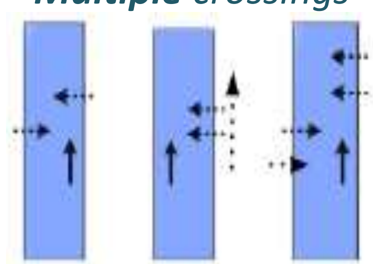


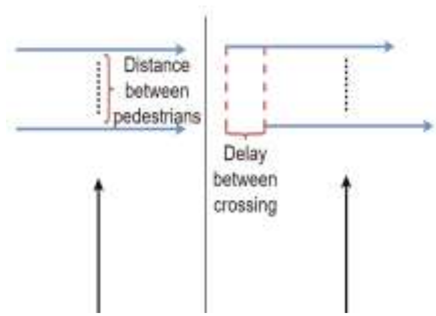
Distance of crossing *pedestrian to car*



Pedestrian ***zig-zag running***

Multiple *crossings*





Results: Summary

(1) **HyLEAP safer than both DESPOT and NavA3C** in most types of GIDAS accident scenarios

	GIDAS safe	Crashes (%)	Impact speed	Near-misses (%)	# De-/Acc.	TTG (s)	Exec. (s)
HyLEAP	**5**	3.01	14.27	8.19	22.91	13.23	0.28
IS-DESPOT-p	3	2.95	16.44	6.22	24.90	13.66	0.27
NavA3C-p	1	3.88	19.18	8.27	22.59	13.56	0.01
React. contr.	0	3.29	5.35	6.86	27.59	15.17	0.001

(2) Averaged over all test scenes

- **DESPOT and HyLEAP are more safe than NavA3C**
- All methods **equally competitive** regarding **comfort of driving** and **time to goal**

(3) **Behavior** of methods **varies over accident scenarios ...**

Ethics for AI ←→ AI for Ethics

Ethics and AI

- **Novel Research Collaborations on Saarland Campus**
  - ICE: Joint work of Philosophy/Ethics, Psychology, Law, and AI
- **DFKI: Projects/proposals focused on Ethics**
  - With Prof. Dabrock & with Prof. Nida-Rümelin

- **Central Research Questions:**
  - How can we embed ethical aspects deeply into AI systems?
  - How can we agree, specify, and ground proper ethics rules?
    - Or can/should we learn them?
  - How can we evaluate how such a system will behave?
    - Need to avoid negative emerging behavior (likely via simulation!)

Take-Aways

- **Digital Reality as a fundamental tool in AI**
  - Modeling and simulation even in complex environments
  - Learning and reasoning via feedback loop (e.g. RL)
  - **Key element for future AI systems**
- **Continuous Learning Loop using Synthetic and Real Data**
  - Loop of model learning, simulation, training, and validation
  - **Validation required to establish trust in AI systems**
  - **Needs significant HPC for simulations and AI**
- **Big Challenges Ahead**
  - Many promising partial results already – but largely islands
  - Requires closer collaboration of industry & academia
  - **CLAIRE: Towards large-scale European initiative**

→ AI: A Central Component for Many Years to come