Building Trustworthy Autonomy in Critical Scenarios

Katie Driggs-Campbell

Assistant Professor

Dept. of Electrical and Computer Engineering

University of Illinois at Urbana-Champaign

How can we design autonomous systems that can safely operate in the real-world with people?

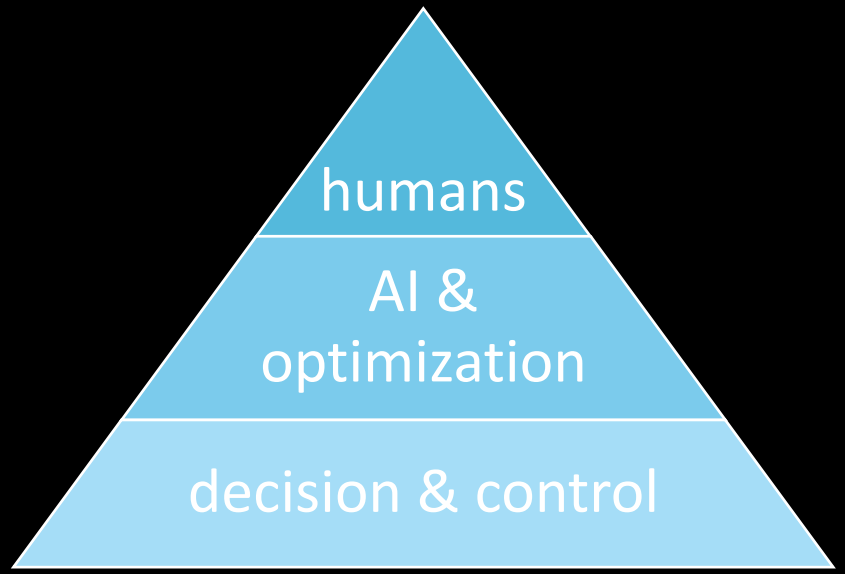
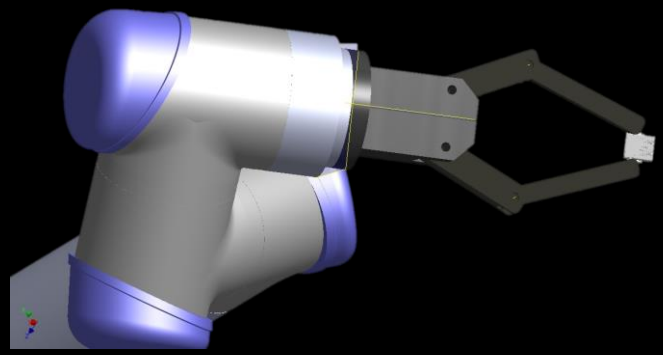
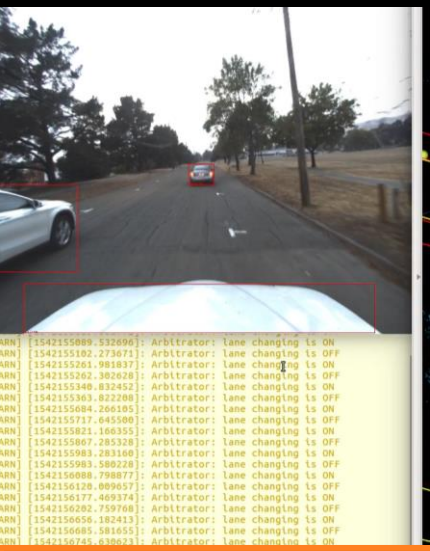
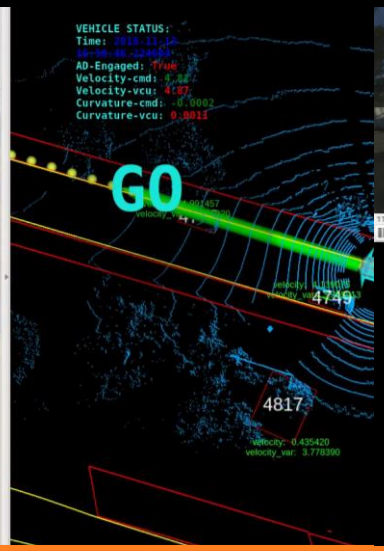


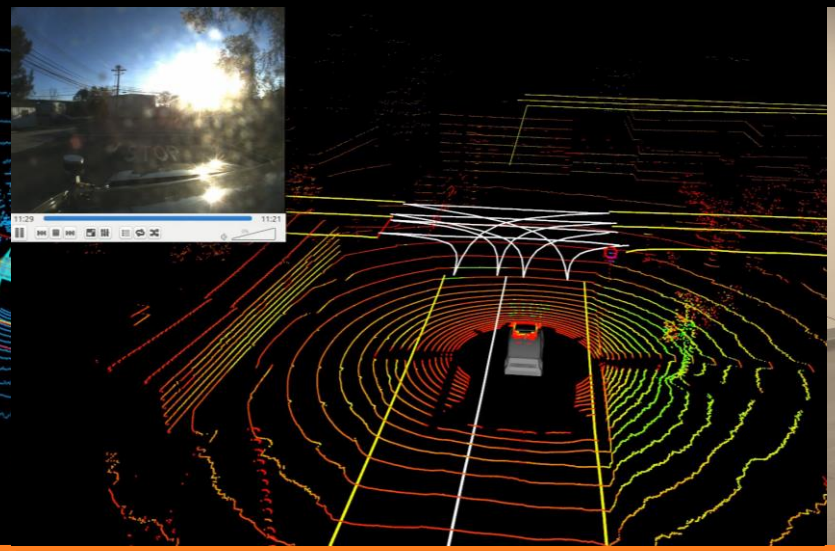

Image credit: Girish Chowdhary

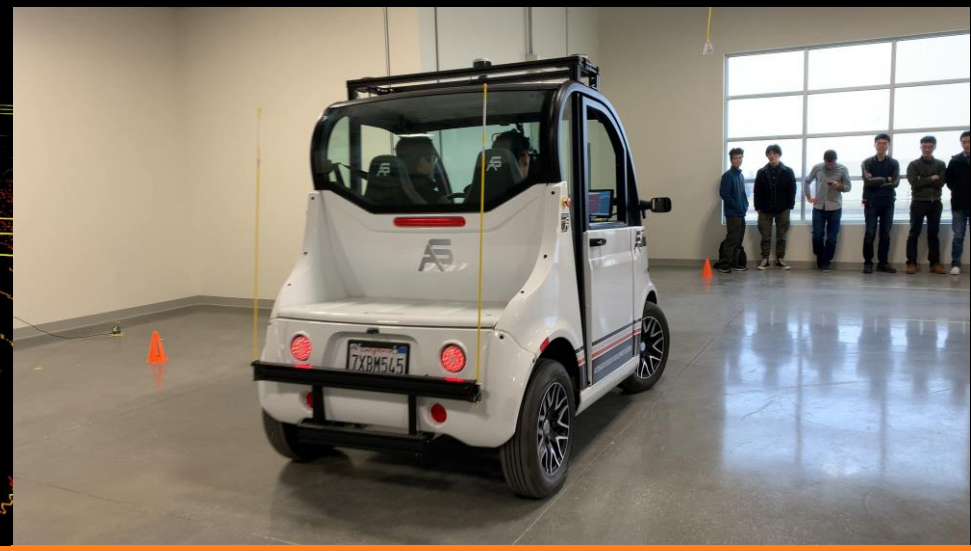


Intervention Scenarios









Cause: model mismatch	Cause: perception uncertainty	Cause: localization failure

Stakeholder Safety and Trust

Today's Topics

Critical and Timely Explanations (Ongoing work)

1. AutoPreview
2. Explanation Necessity Dataset

Offline Safety Validation

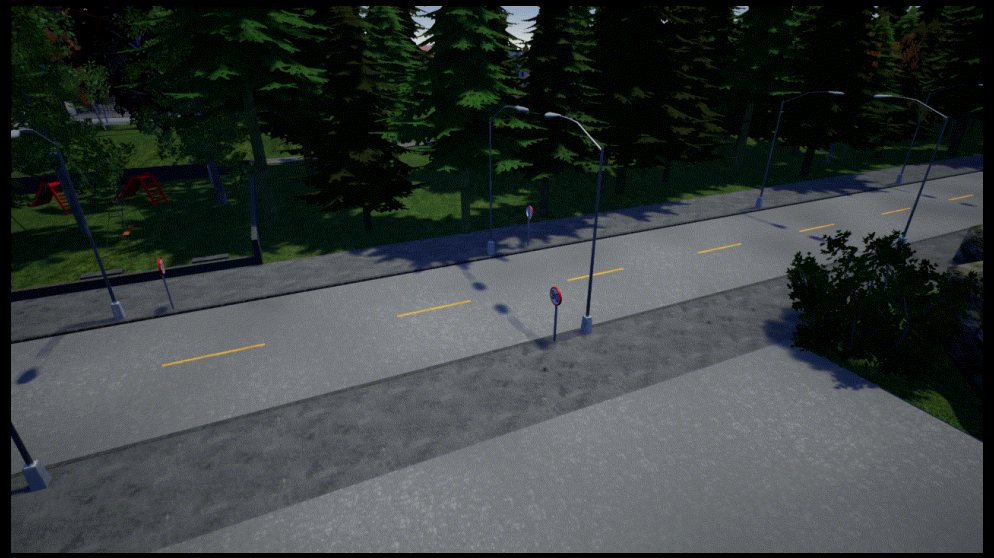
1. Adaptive testing frameworks
2. Integrating human insight into validation via critical states

Low-Probability, High-Risk Events

Hazardous Event Frequencies	
Disengagement Rate	0.12 per 1000 km
Collision Rate	12.5 per 100 million km
Fatality Rate	0.70 per 100 million km

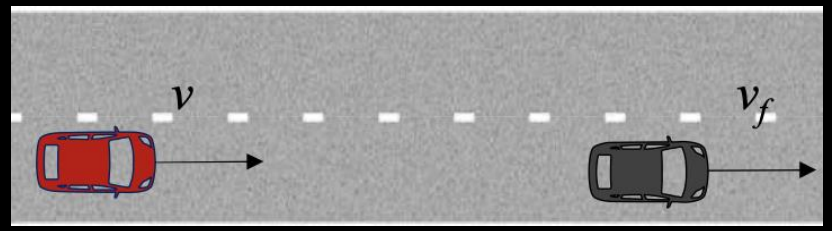
Low-Probability, High-Risk Events

Hazardous Event Frequencies	
Disengagement Rate	0.12 per 1000 km
Collision Rate	12.5 per 100 million km
Fatality Rate	0.70 per 100 million km



Low-Probability, High-Risk Events

Hazardous Event Frequencies	
Disengagement Rate	0.12 per 1000 km
Collision Rate	12.5 per 100 million km
Fatality Rate	0.70 per 100 million km

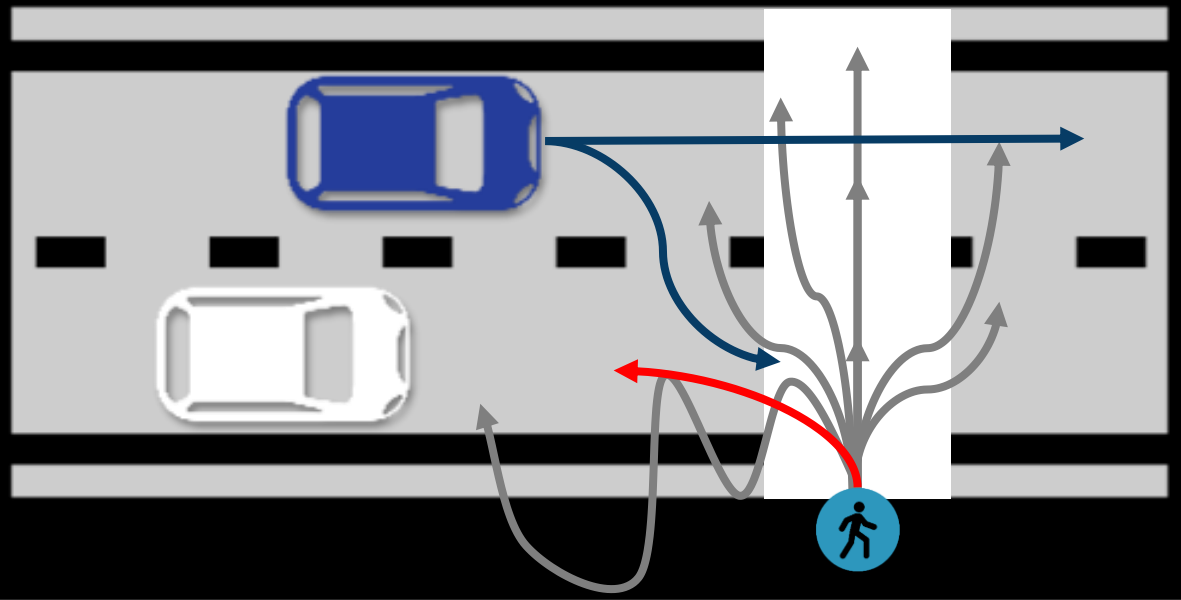


Name	Description	Min	Max	Offset
V_f	Speed of FV (m/s)	0	30	0.5
Dif_v	Speed of FV minus speed of AV (m/s)	-10	10	0.5
Dis	Distance between two vehicles (m)	5	50	1

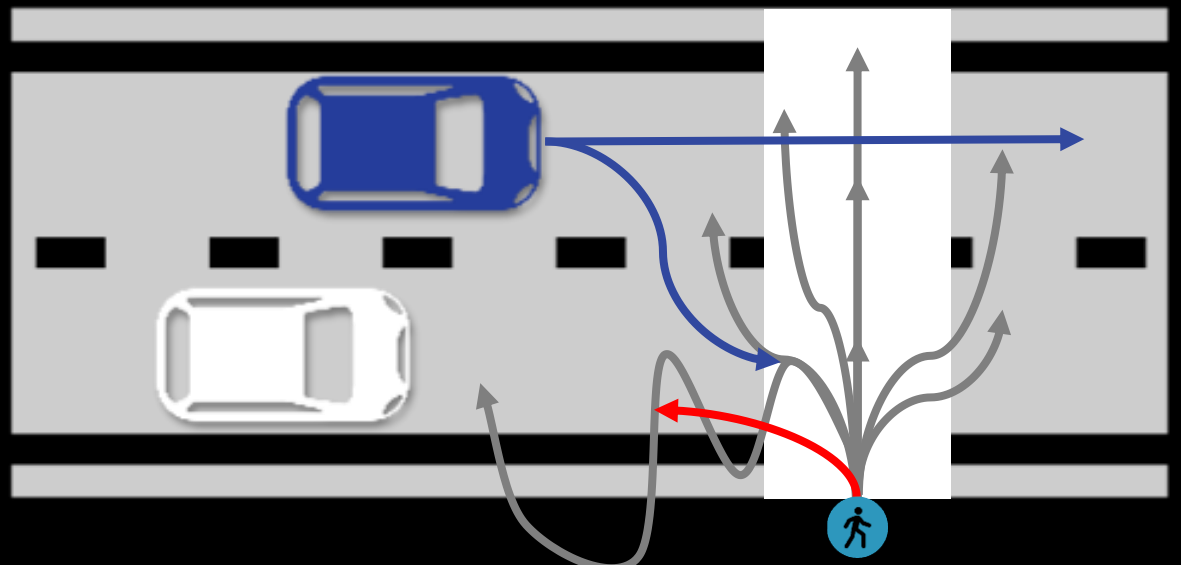
~100,000 scenarios in this test matrix

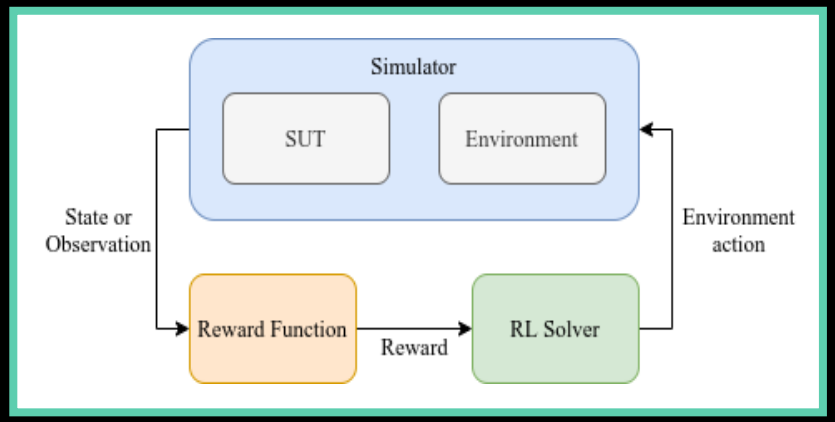
Low-Probability, High-Risk Events

Hazardous Event Frequencies	
Disengagement Rate	0.12 per 1000 km
Collision Rate	12.5 per 100 million km
Fatality Rate	0.70 per 100 million km



Adaptive Stress Testing

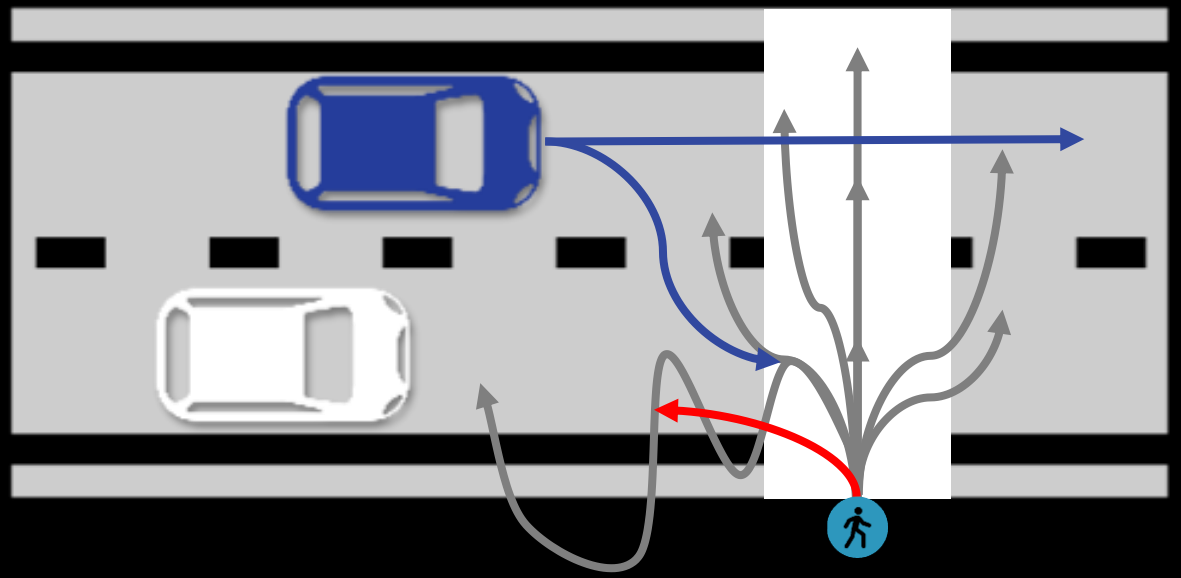


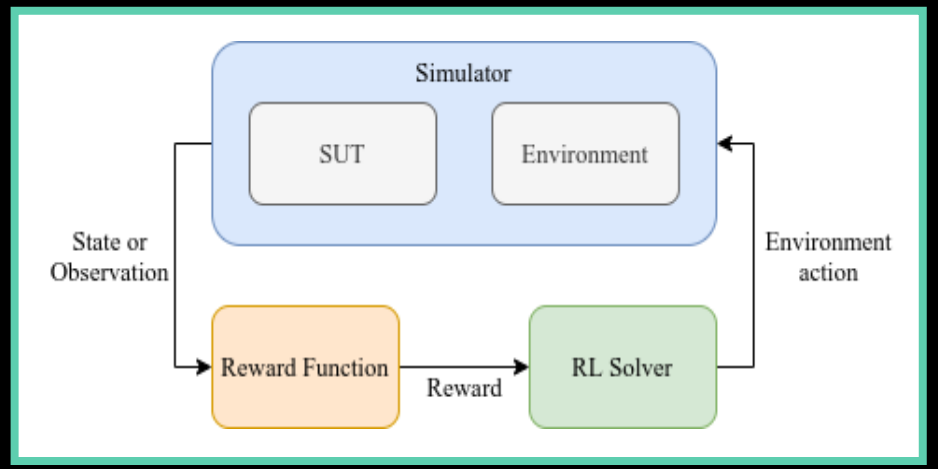


- A simulator which contains the system under test (SUT) and environment
- A reward function to guide (adversarial) policy training to find failure cases
- An RL solver to generate environment actions that updates the simulator through time

P. Du,* A. Corso,* K. Driggs-Campbell, and M.J. Kochenderfer, *Adaptive Stress Testing with Reward Augmentation for Autonomous Vehicle Validation,* ITSC 2019.
R. Lee, et al. *Adaptive Stress Testing of Airborne Collision Avoidance Systems*, in DASC 2015.

Adaptive Stress Testing



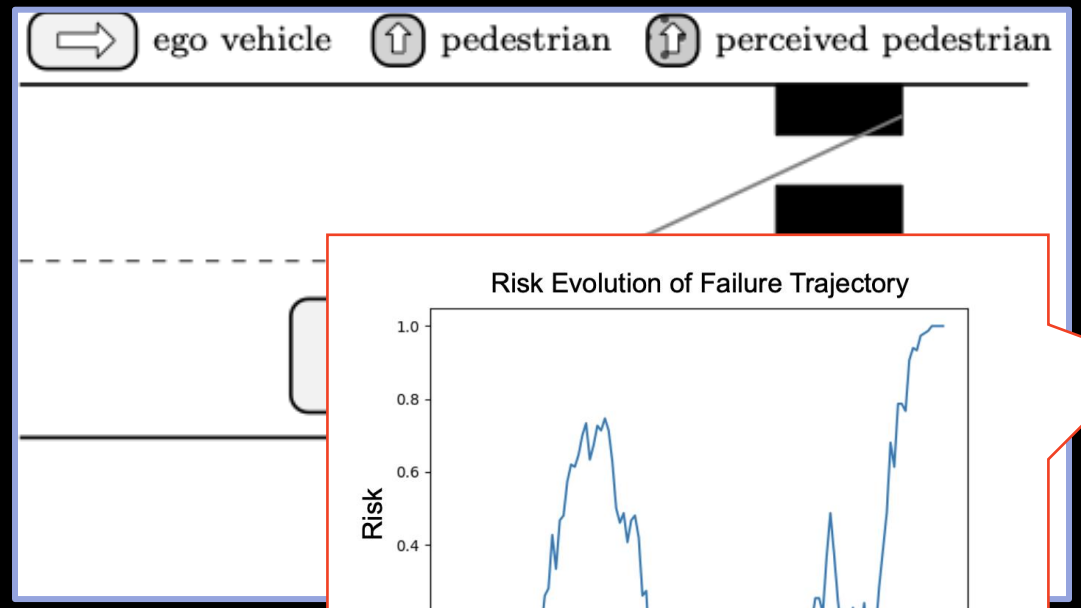


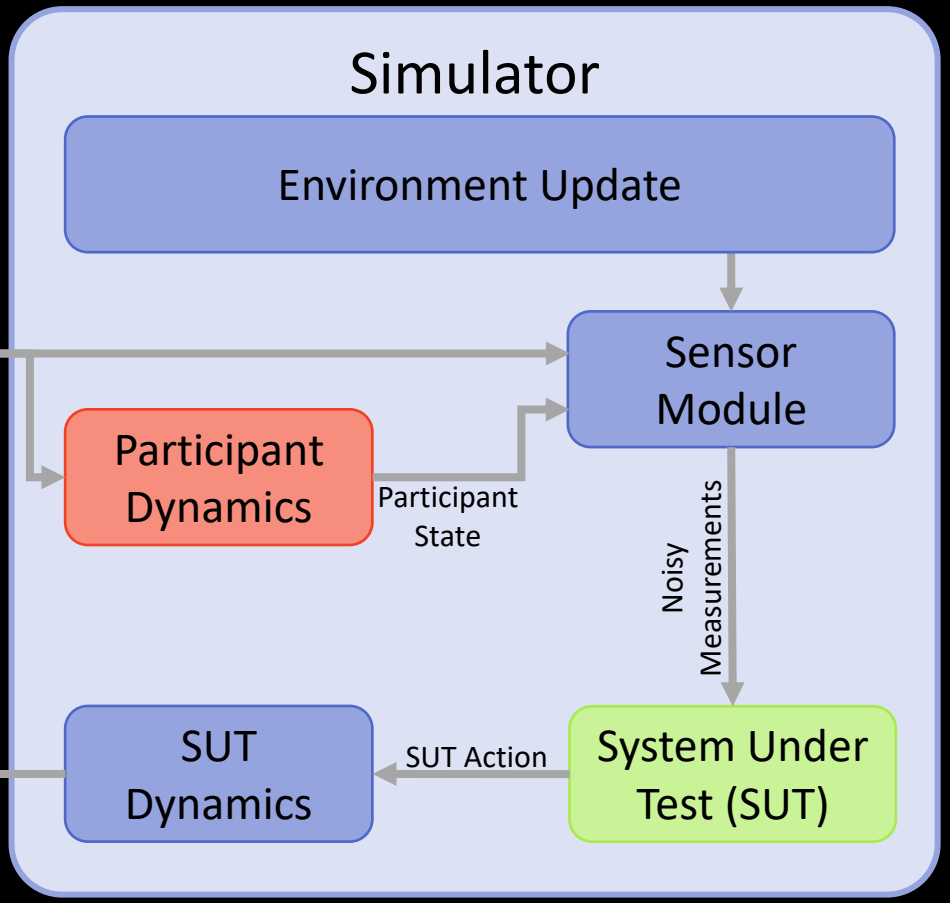
AST phrases the validation as an RL problem to actively search and find likely failures.

$$R(s,a) = \begin{cases} 0, & \text{if failure} \\ -\alpha - \beta f(s), & \text{if no failure and } t = T \\ -\log P(a) - \eta h(s), & \text{if no failure and } t < T \end{cases}$$

P. Du,* A. Corso,* K. Driggs-Campbell, and M.J. Kochenderfer, *Adaptive Stress Testing with Reward Augmentation for Autonomous Vehicle Validation,* ITSC 2019.
R. Lee, et al. *Adaptive Stress Testing of Airborne Collision Avoidance Systems*, in DASC 2015.

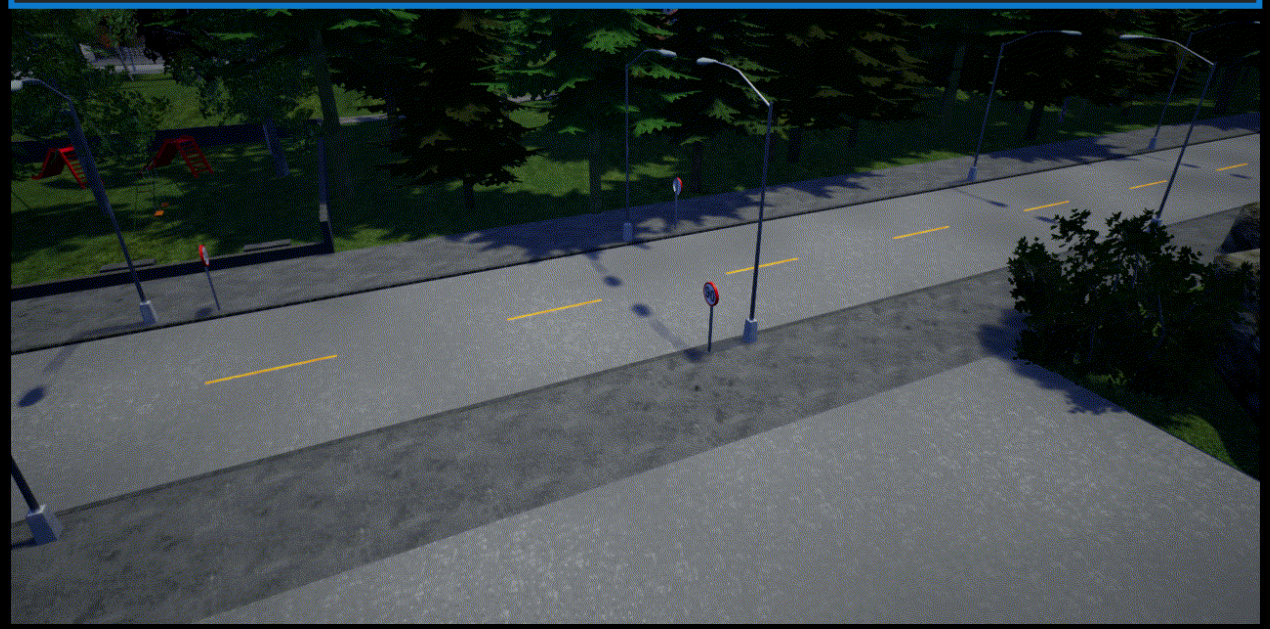
Multi-Agent Crosswalk Example





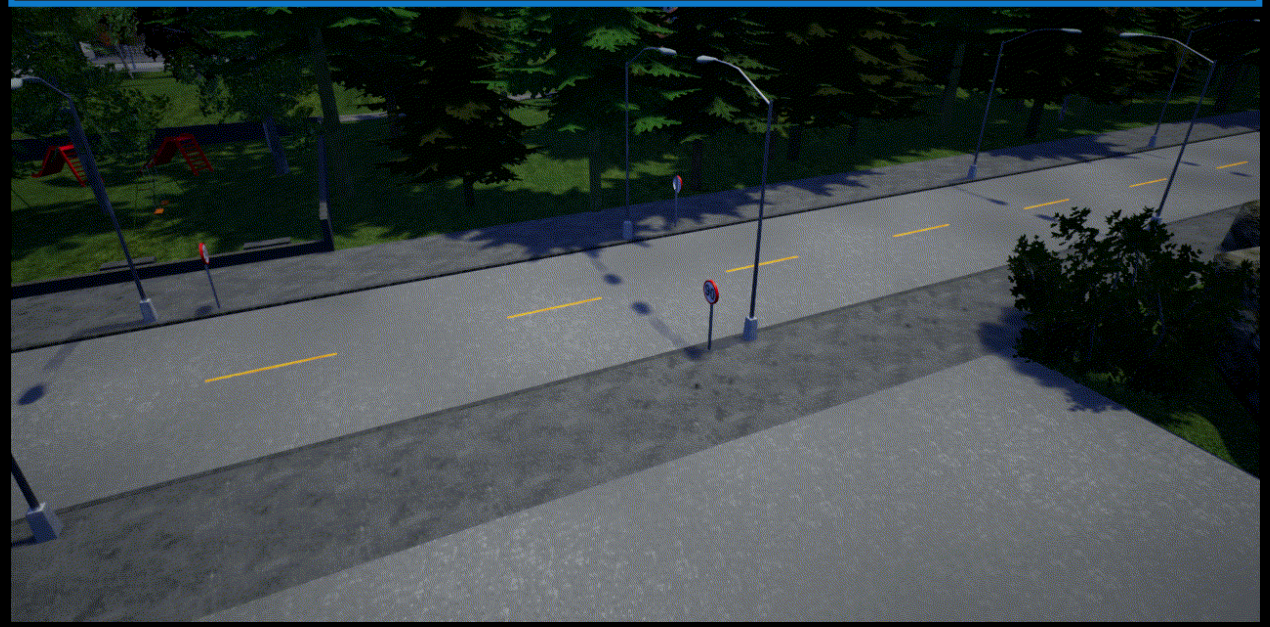
P. Du and K. Driggs-Campbell, *Finding Diverse Failure Scenarios in Autonomous Systems Using Adaptive Stress Testing,* SAE Journal of Connected and Automated Vehicles, 2019.
M. Koren, et al. *Adaptive stress testing for autonomous vehicles*, IV 2018.

Typical Simulation

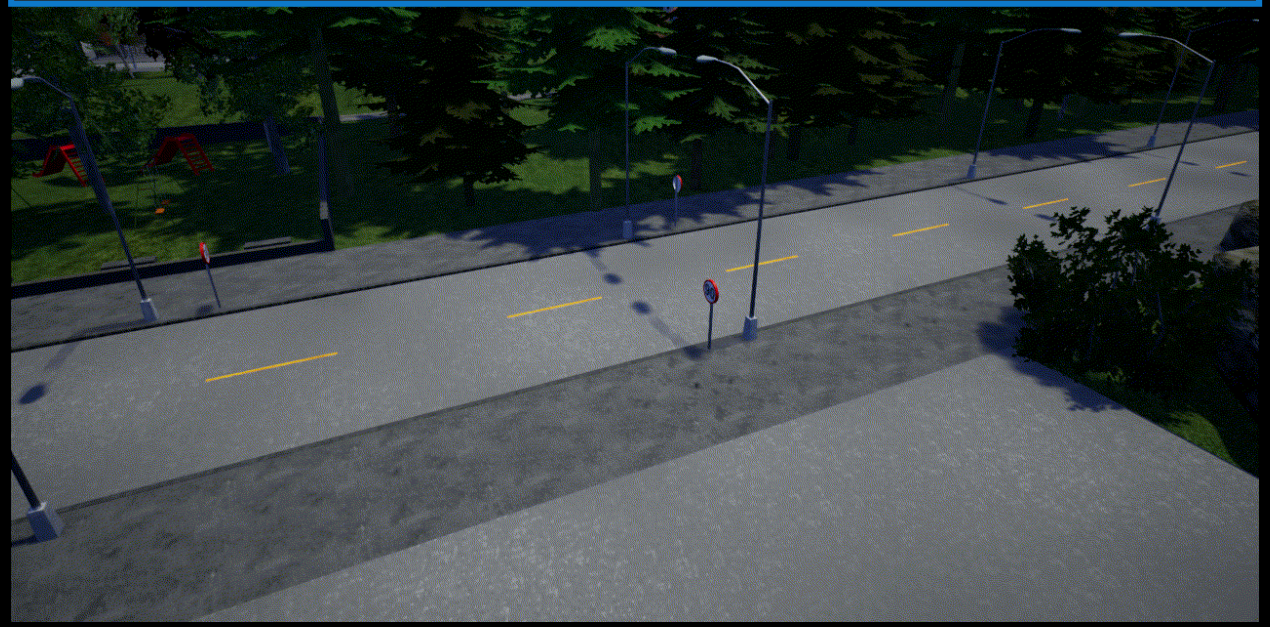


P. Du and K. Driggs-Campbell, *Finding Diverse Failure Scenarios in Autonomous Systems Using Adaptive Stress Testing,* SAE Journal of Connected and Automated Vehicles, 2019.
A. Corso,* P. Du,* K. Driggs-Campbell, and M.J. Kochenderfer, *Adaptive Stress Testing with Reward Augmentation for Autonomous Vehicle Validation,* ITSC 2019.

Typical Simulation

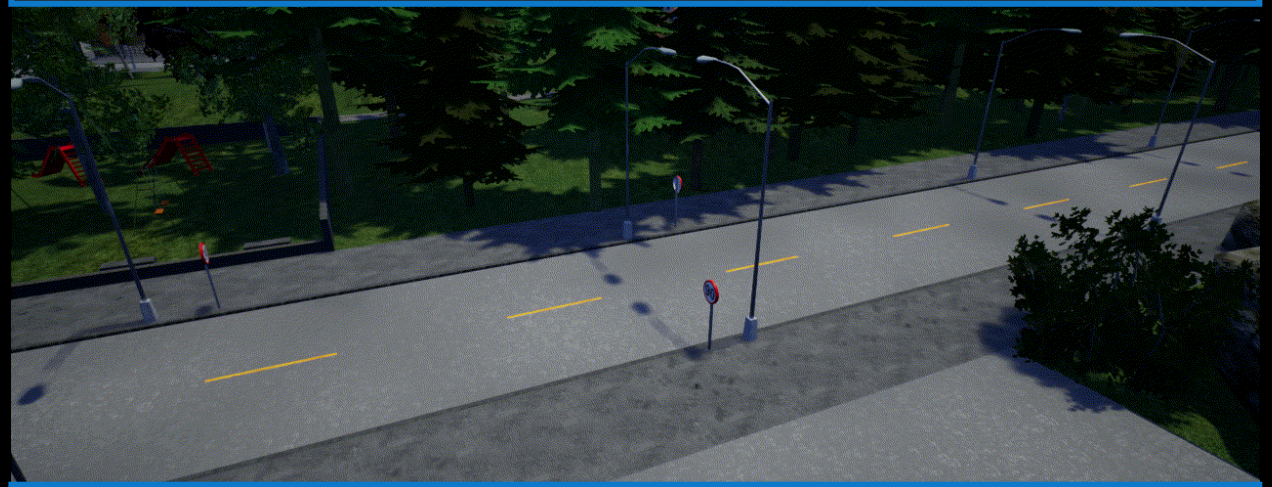


Worst Case Adversarial Example

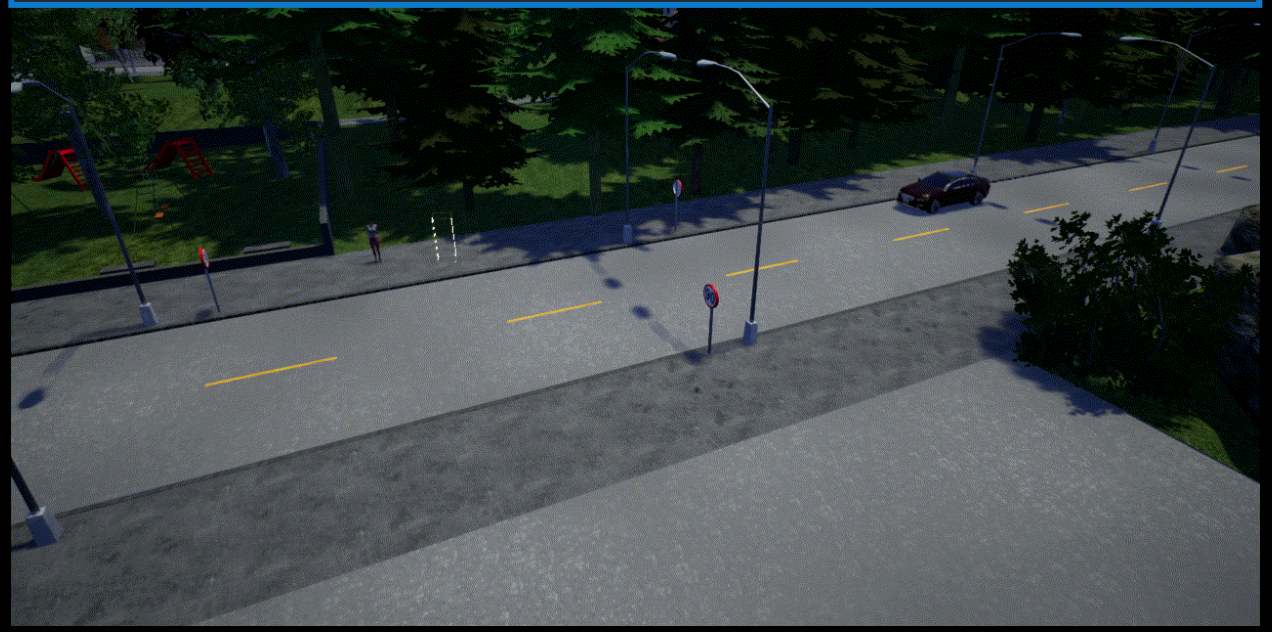


P. Du and K. Driggs-Campbell, *Finding Diverse Failure Scenarios in Autonomous Systems Using Adaptive Stress Testing,* SAE Journal of Connected and Automated Vehicles, 2019.
A. Corso,* P. Du,* K. Driggs-Campbell, and M.J. Kochenderfer, *Adaptive Stress Testing with Reward Augmentation for Autonomous Vehicle Validation,* ITSC 2019.

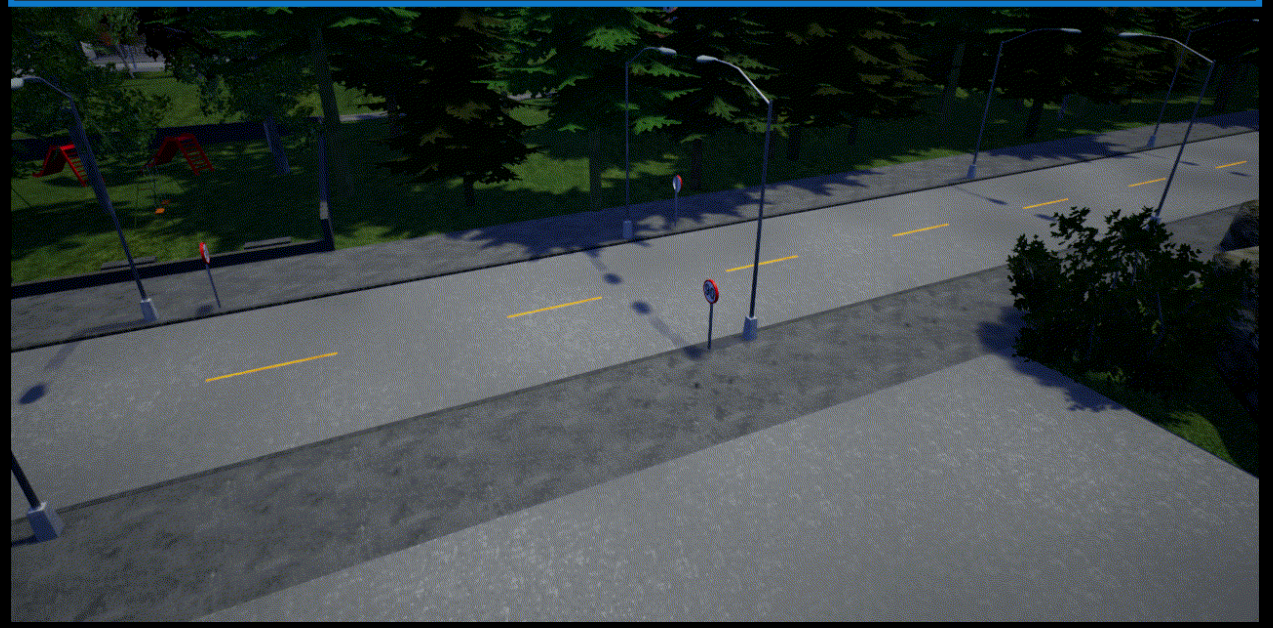
Typical Simulation



How to find relevant failures?



Worst Case Adversarial Example



AST phrases the validation as an RL problem to actively search and find likely failures.

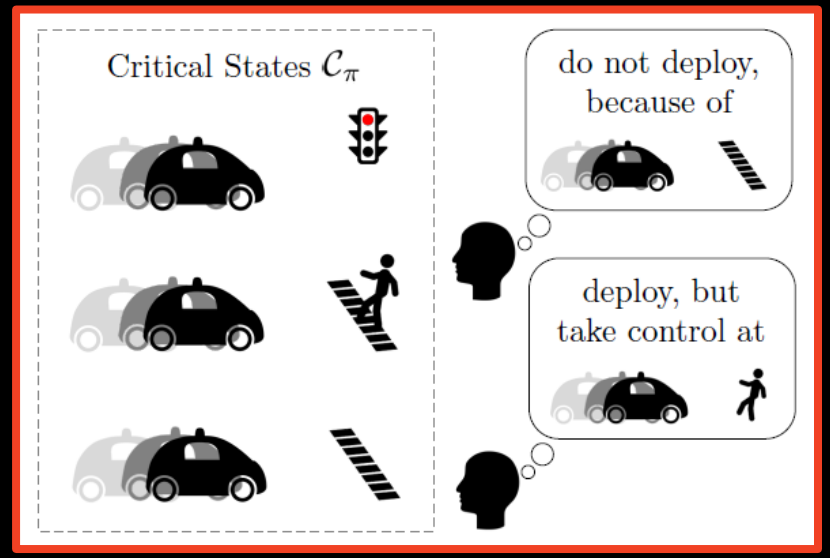
$$R(s,a) = \begin{cases} 0, & \text{if failure} \\ -\alpha - \beta f(s), & \text{if no failure and } t = T \\ -\log P(a) - \eta h(s), & \text{if no failure and } t < T \end{cases}$$

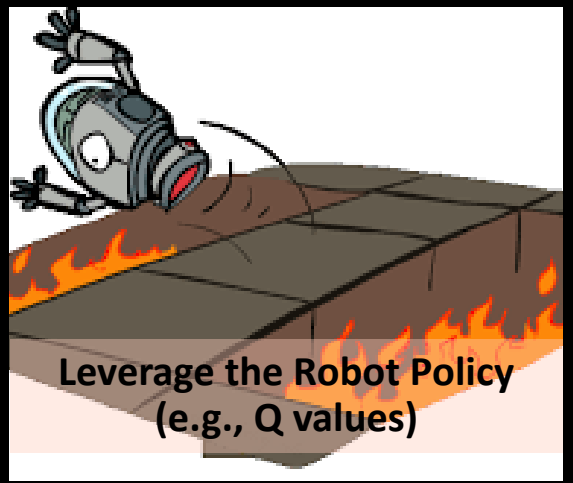
P. Du and K. Driggs-Campbell, *Finding Diverse Failure Scenarios in Autonomous Systems Using Adaptive Stress Testing,* SAE Journal of Connected and Automated Vehicles, 2019.

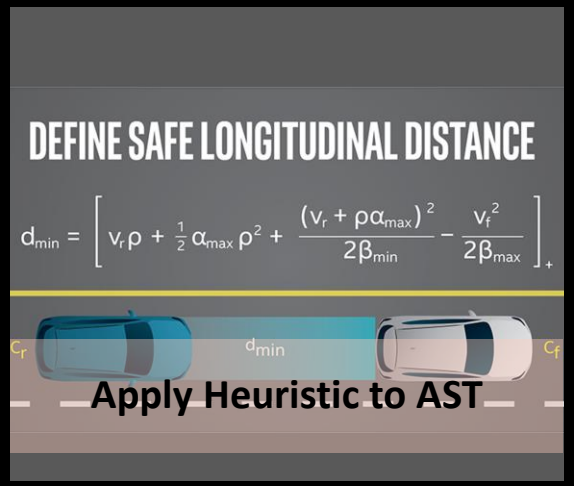
A. Corso,* P. Du,* K. Driggs-Campbell, and M.J. Kochenderfer, *Adaptive Stress Testing with Reward Augmentation for Autonomous Vehicle Validation,* ITSC 2019.

Critical States

- It is infeasible to examine the behavior of the system in all possible scenarios
  - **Key Idea:** Examine a smaller set of *critical states* that represent safety critical scenarios
- Huang et al. demonstrated that critical states can be used to help the user build an effective mental model of system behavior



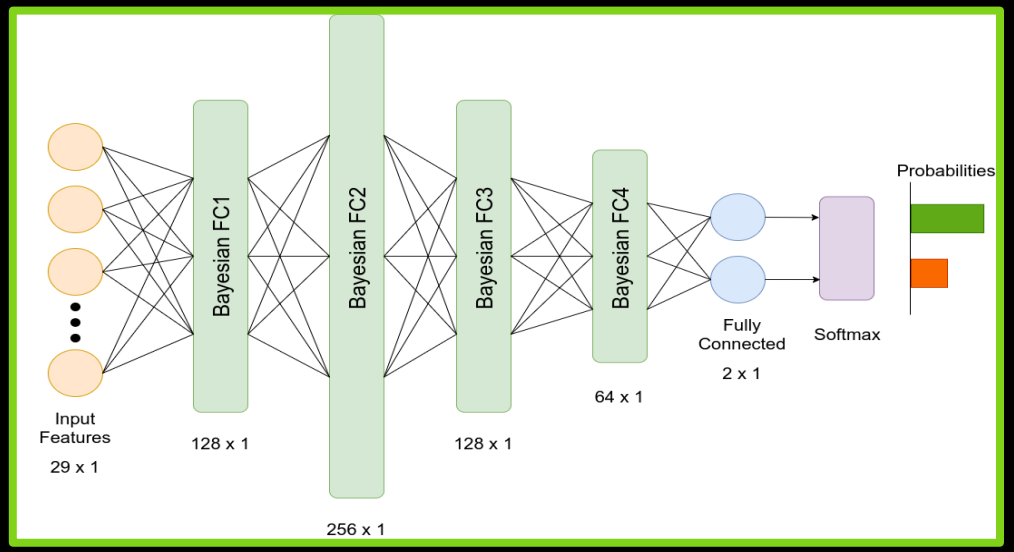




S. H. Huang, K. Bhatia, P. Abbeel, and A. D. Dragan, “Establishing Appropriate Trust via Critical States,” IROS 2018.

Classifying Critical States

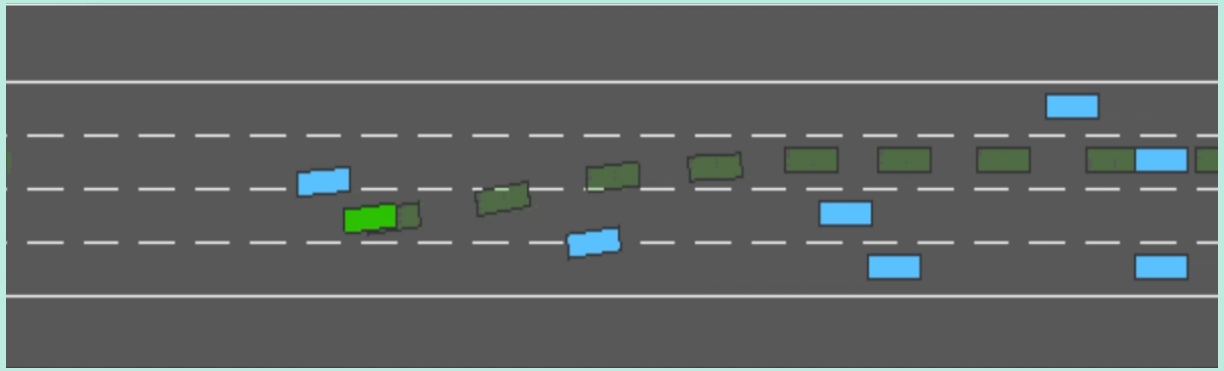
We seek to incorporate the notion of critical states from an *onlooker's* perspective into the failure search framework with human expert labels.



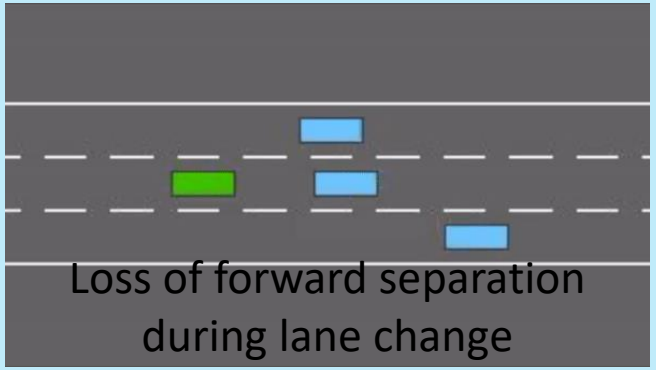
Human critical state (HCS) classifier

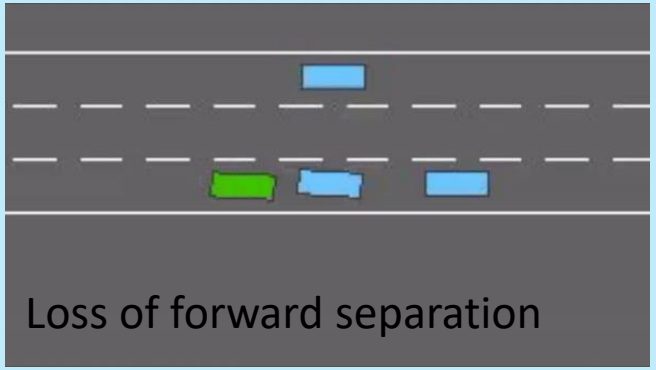
A Bayesian neural network is trained to predict the probability of a given state representing a critical scenario. Classification is based on the state of the environment and SUT.

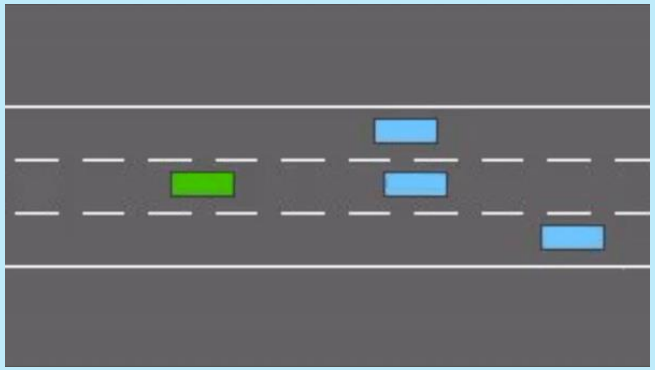
Test Environment



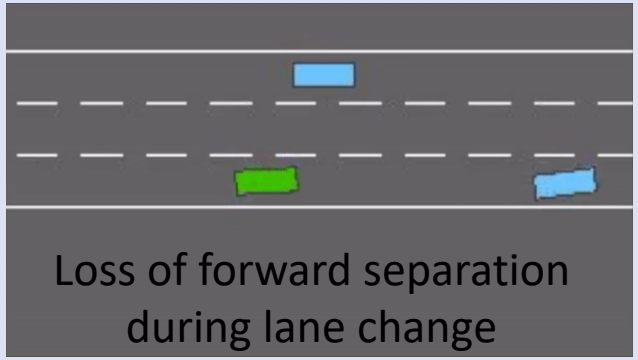
Heuristic Methods

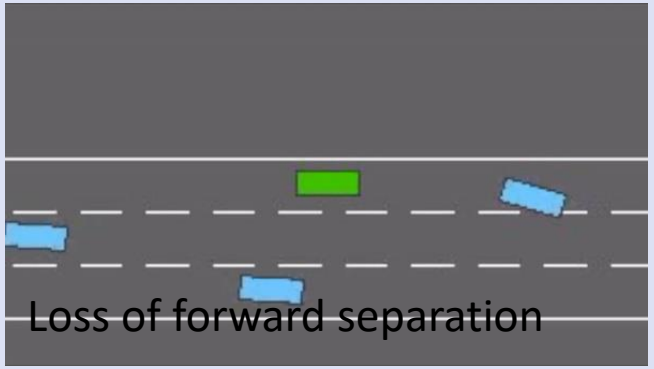


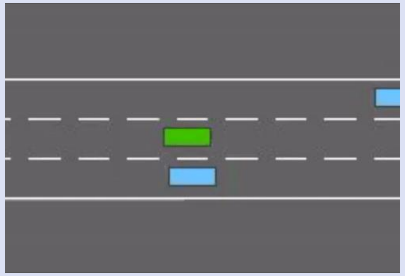


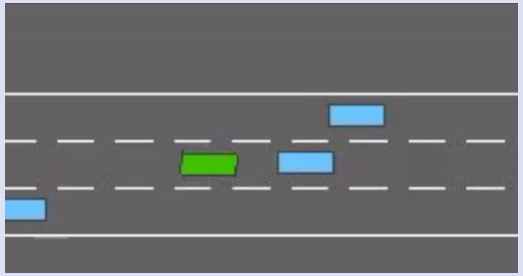


Human-in-the-Loop









E. Leurent, “An Environment for Autonomous Driving Decision-Making,” https://github.com/eleurent/highway-env, 2018.
P. Du and K. Driggs-Campbell. “Adaptive Failure Search Using Critical States from Domain Experts” ICRA 2021.

Validation with Human Insight

- AST is a tool for finding failures in complex systems
  - Allows stakeholders to assess performance of black-box / safety critical systems
  - Does not automatically generate intuitive (or relevant) results!
- Using critical states from domain experts provides our tool more intuition about what is risky
  - The resulting failures capture designer's internal model of risk and lead to a diverse set of scenarios

Today's Topics

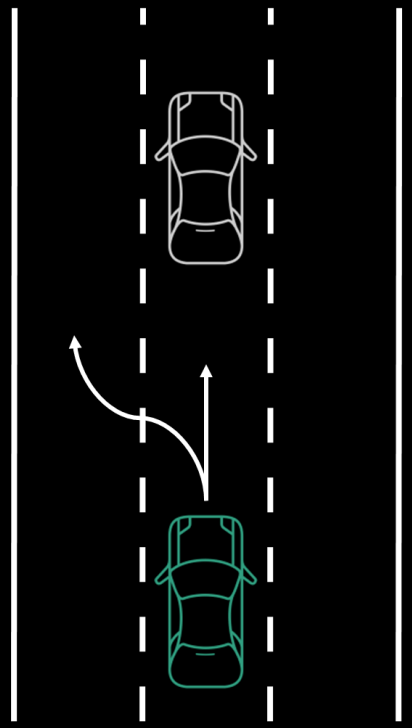
Critical and Timely Explanations (Ongoing work)

1. AutoPreview
2. Explanation Necessity Dataset

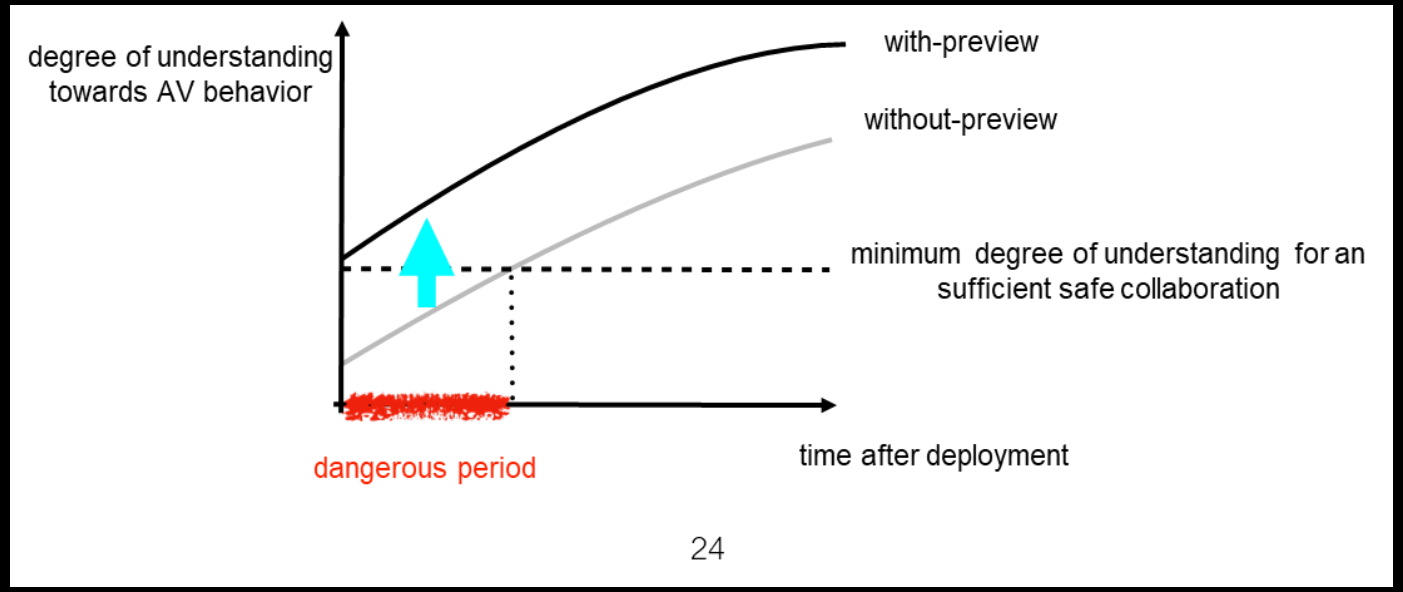
Offline Safety Validation

1. Adaptive testing frameworks
2. Integrating human insight into validation via critical states

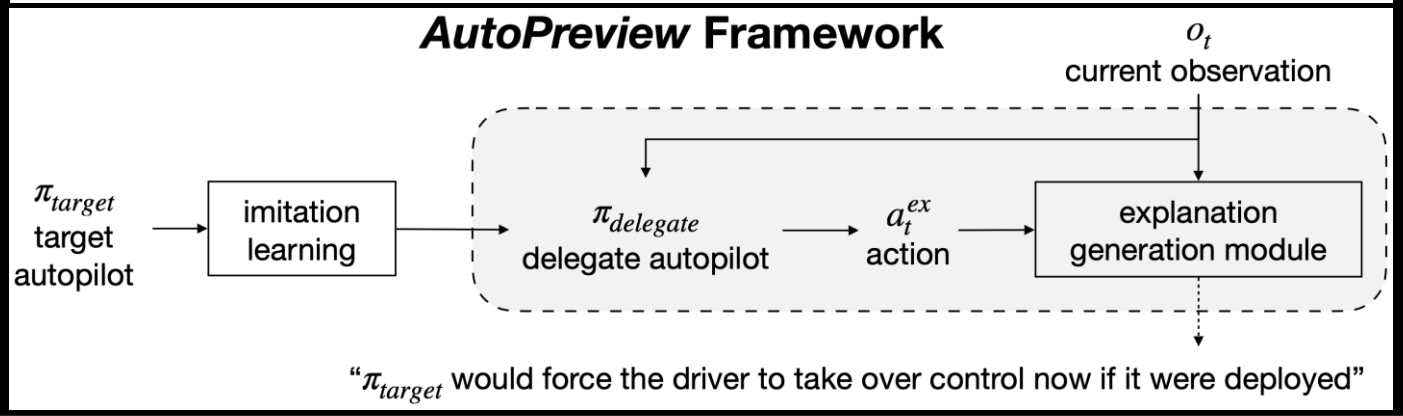
Building Mental Models Online



AutoPilot would overtake from the left lane



24



Explanations at Critical States

Explanation Generation for Autonomous Vehicles aims to:

- Improve awareness and trust of the passenger of the AV
- Take a human-centered approach:
  - *Timing:* Be minimally invasive
  - *Empathy:* Explanation content should not increase anxiety
  - *Personalization:* Consider preferred explanations & timing

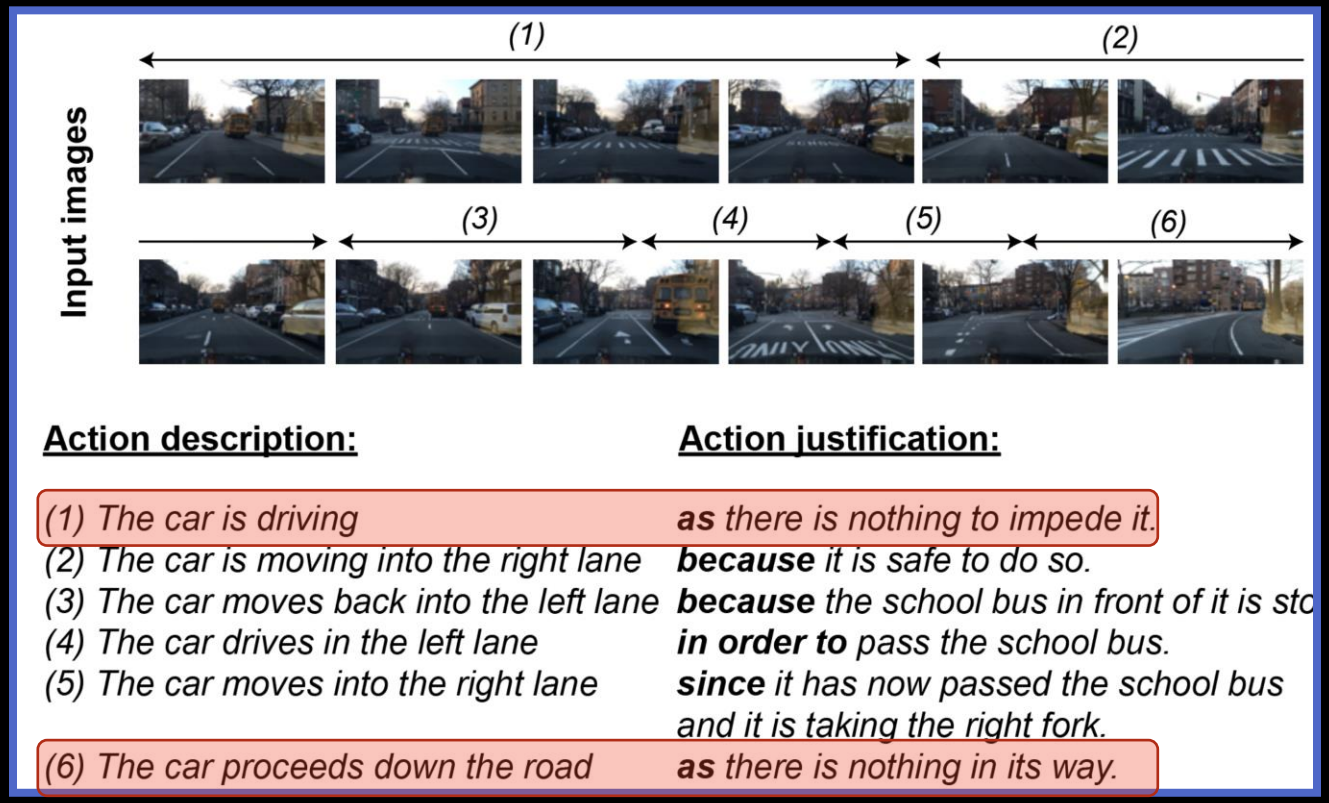
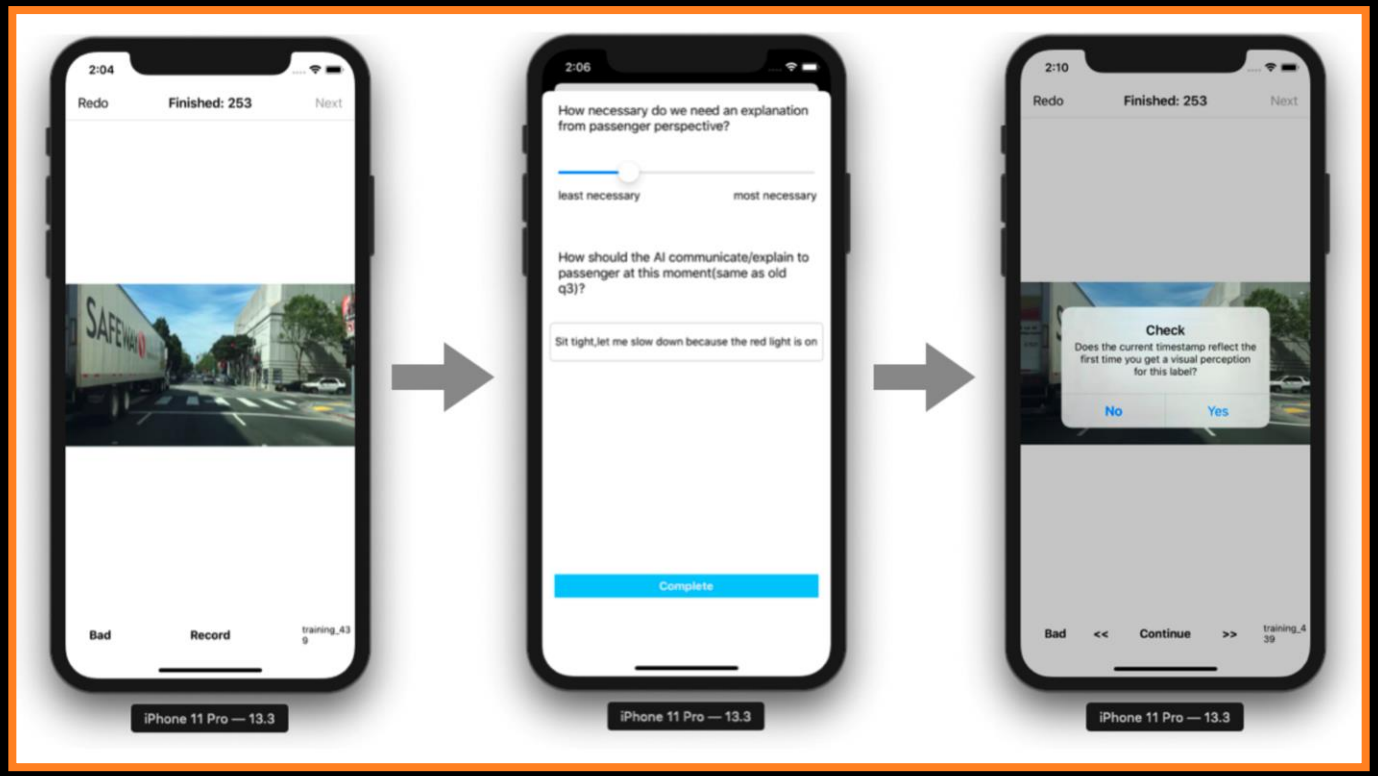


Image Credit: Kim et al. ECCV 2018.

Explanation Necessity Dataset



1103 driving video clips from BDD-A dataset can be annotated with necessity score, explanation moment, and explanation content

Necessity-based Explanations

Initial User Study tells us key factors are:

- The scenario under question
  - Near crash, merging, slowing, parking, pedestrian crossing, approaching intersection

Today's Topics

Critical and Timely Explanations (Ongoing work)

1. AutoPreview
2. Explanation Necessity Dataset

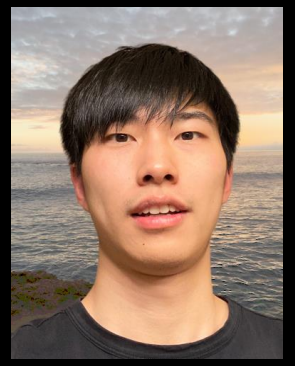
Offline Safety Validation

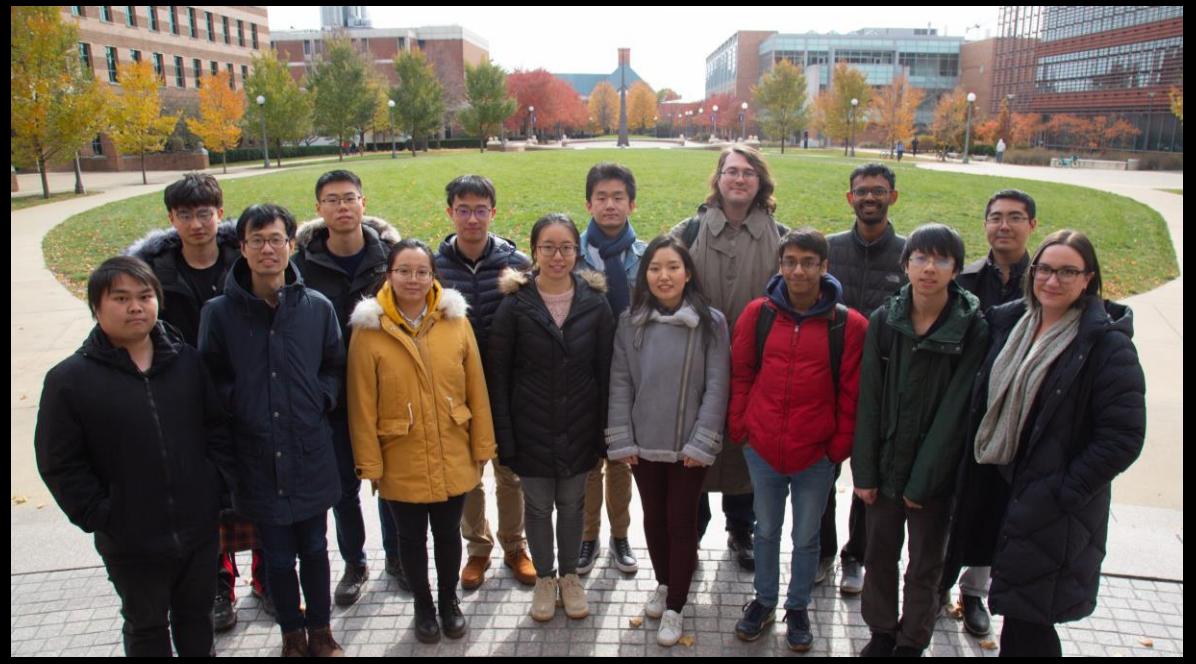
1. Adaptive testing frameworks
2. Integrating human insight into validation via critical states

Acknowledgements

ILLINOIS HEXAGON AUTONOMOUSTUFF
Center for Autonomy





Today's Topics

Critical and Timely Explanations (Ongoing work)

1. AutoPreview
2. Explanation Necessity Dataset

Offline Safety Validation

1. Adaptive testing frameworks
2. Integrating human insight into validation via critical states

Thank you!

Extra slides

Heuristic Reward Function

The baseline reward is influenced by the discovery of a failure state and a heuristic reward to guide failure search.

$$R_{heur}(s) = \begin{cases} 0 & s \in E \\ -\infty & s \notin E, t \geq T \\ h(s) & s \notin E, t < T \end{cases}$$

Set of failure states

Time horizon of failure search

Heuristic signal

HCS Reward Function

When considering critical states, the reward function used to train the failure generation policy contains both a reward signal provided by the HCS classifier and a heuristic measure.

$$R_{hcs}(s) = \begin{cases} 0 & s \in E \\ -\infty & s \notin E, t \geq T \\ (1-\sigma^2)\beta\mu + \sigma^2 h(s) & s \notin E, t < T \end{cases}$$

Set of failure states

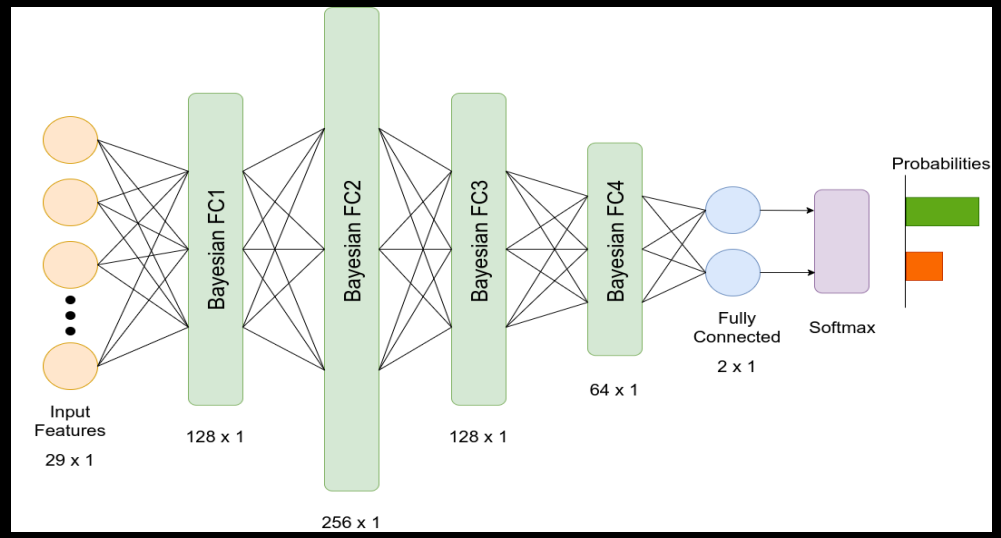
Time horizon of failure search

Variance of distribution from classifier

Mean of distribution from classifier

Classifying Critical States

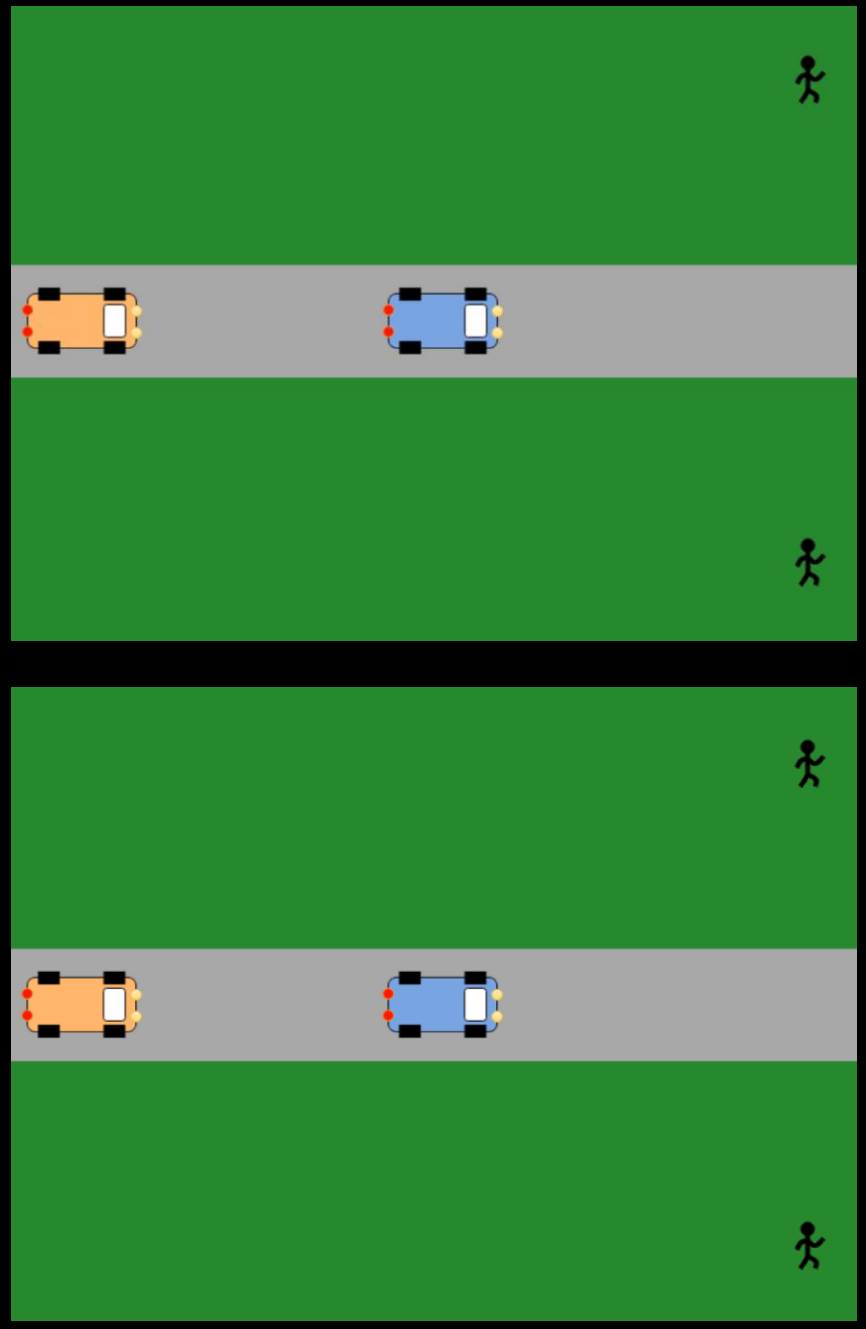
We seek to incorporate the notion of critical states from an *onlooker's* perspective into the failure search framework with human expert labels.



Human critical state (HCS) classifier

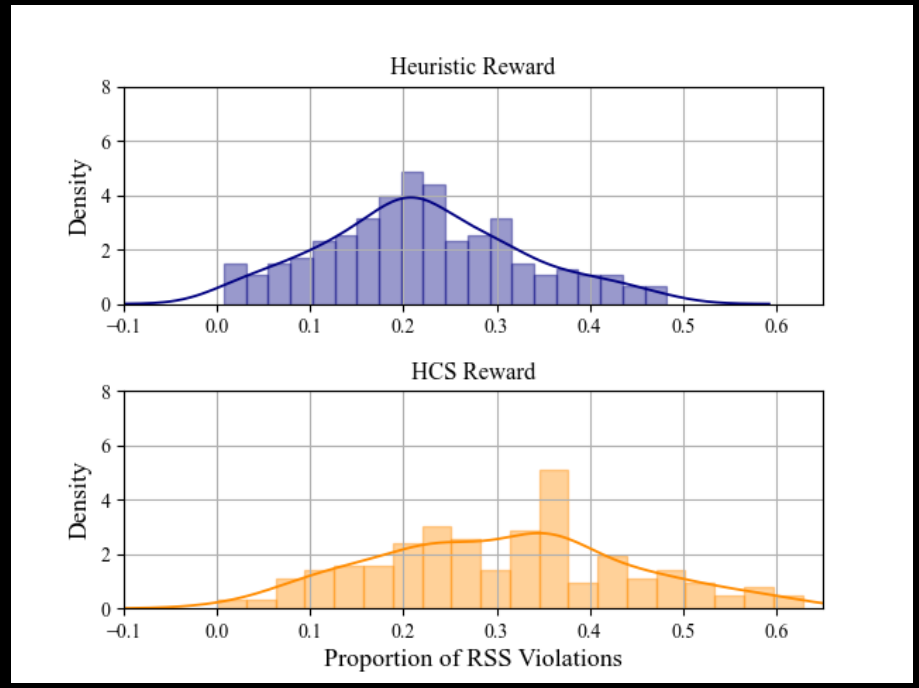
A Bayesian neural network is trained to predict the probability of a given state representing a critical scenario. Classification is based on the state of the environment and SUT.

Adaptive Stress Testing with Reward Augmentation



P. Du and K. Driggs-Campbell, *Finding Diverse Failure Scenarios in Autonomous Systems Using Adaptive Stress Testing,* To appear in SAE Journal of Connected and Automated Vehicles, 2019.
A. Corso,* P. Du,* K. Driggs-Campbell, and M.J. Kochenderfer, *Adaptive Stress Testing with Reward Augmentation for Autonomous Vehicle Validation,* ITSC 2019.

- The heuristic used for baseline AST is based on the longitudinal separation between the ego vehicle and closest environment vehicle which is a good fit for the RSS criteria.
- Despite this, we see the median proportion of improper response (w.r.t. RSS) increase by approx. 50% from 21% (heuristic) to 32% (HCS).



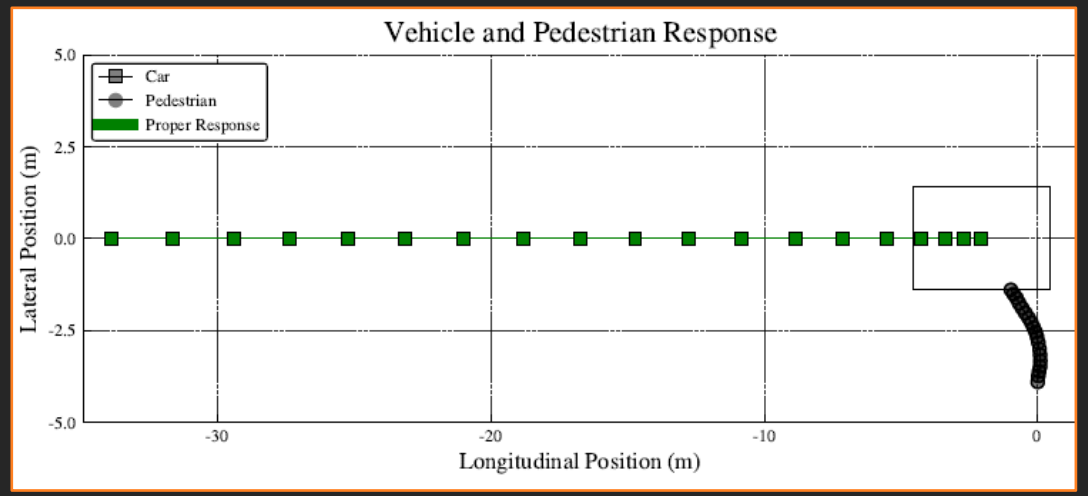
Adaptive Stress Testing with Reward Augmentation

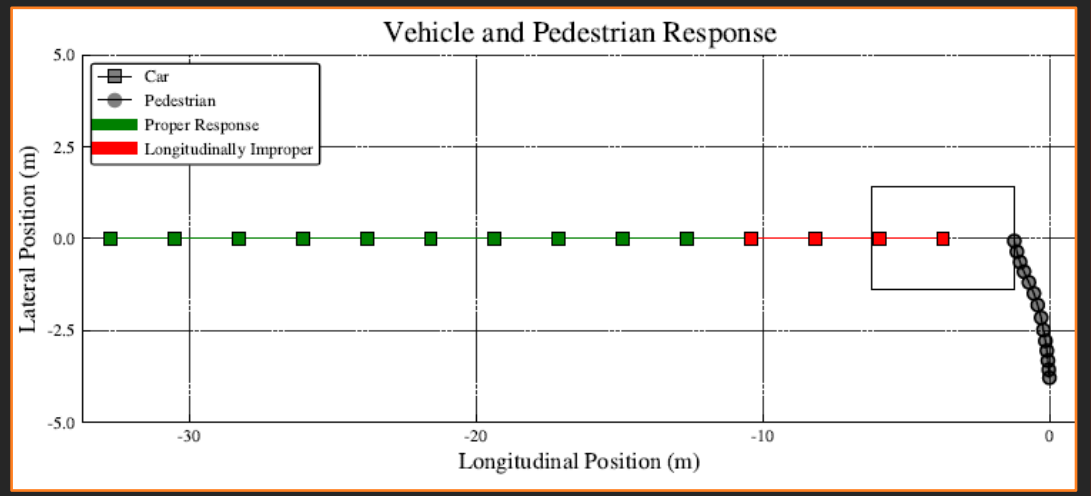
Actively search the system to find *diverse and meaningful* likely failures in your system:

$$R(s,a) = \begin{cases} R_{TD}, & \text{if failure} \\ -\alpha - \beta f_{RSS}(s), & \text{if no failure and } t = T \\ -\log P(a), & \text{if no failure and } t < T \end{cases}$$

Failure Type * indicates fault	Generic AST	AST++
Vehicle/Pedestrian*	25	15
Vehicle*/Pedestrian	0	4
Vehicle*/Vehicle	0	6

Responsibility-Sensitive Safety





AST: Failure Assessment Comparison

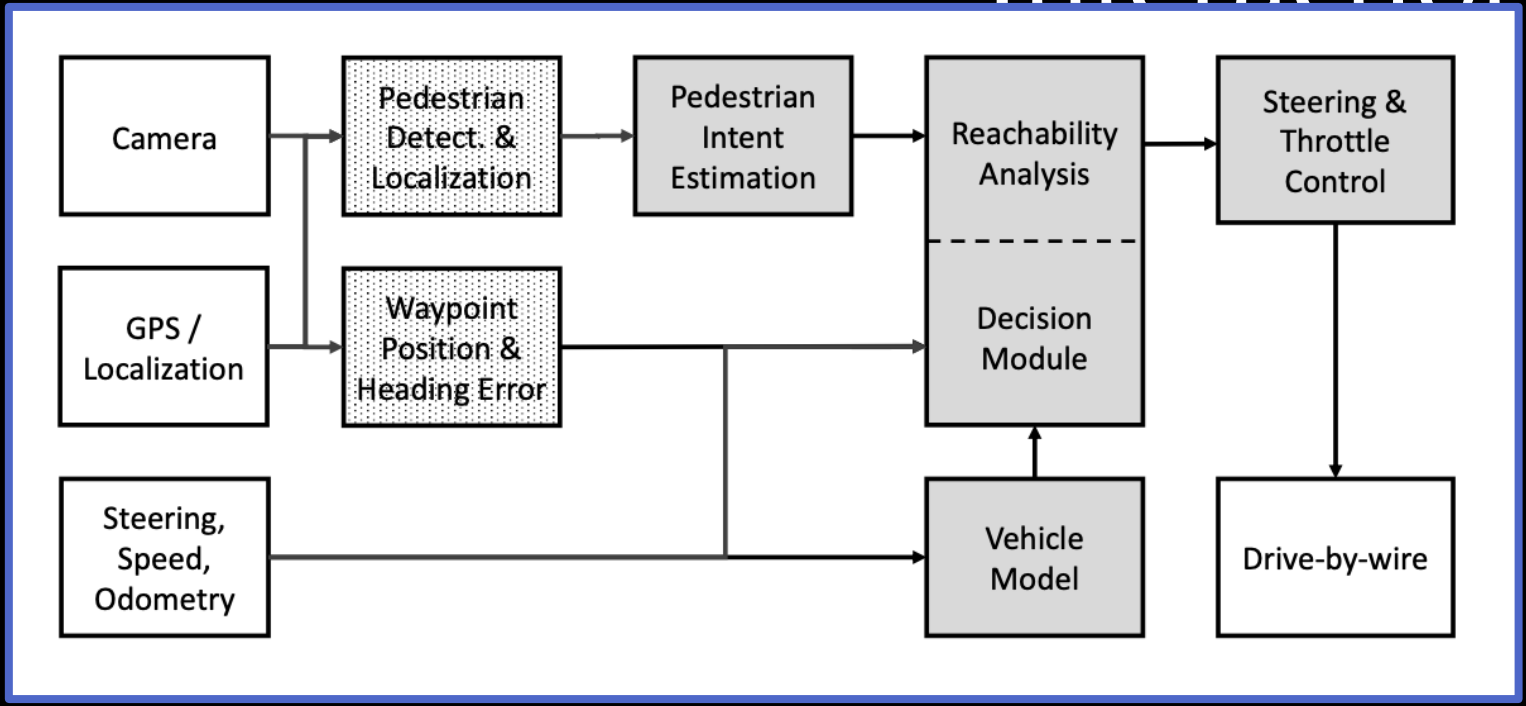
Heuristic / Rule-based System

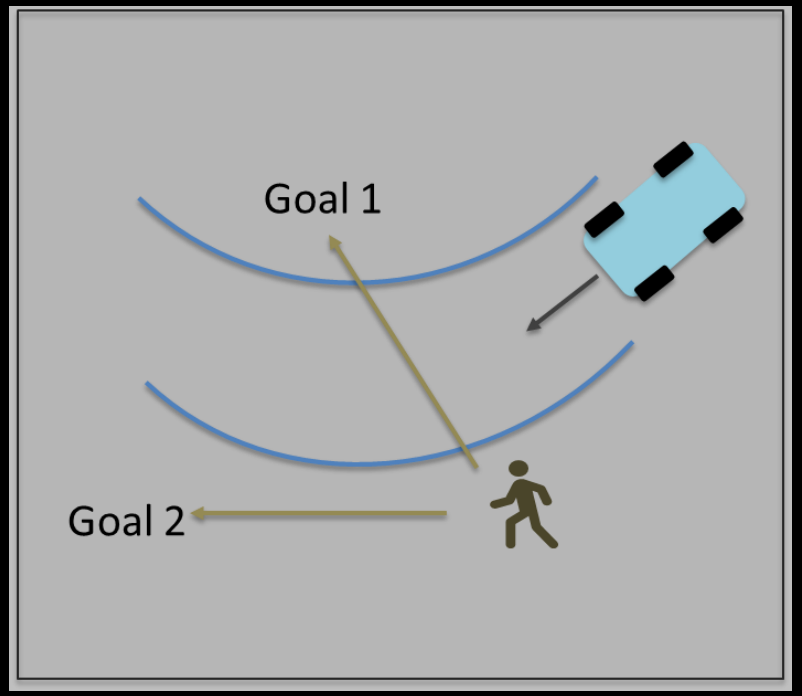
```
prob: 1.0 isevent: false dist: 14.846573992753521
traj: 1 t: 2
```

System with Tracking Errors

```
prob: 1.0 isevent: false dist: 6.463812923468552
traj: 1 t: 2
```

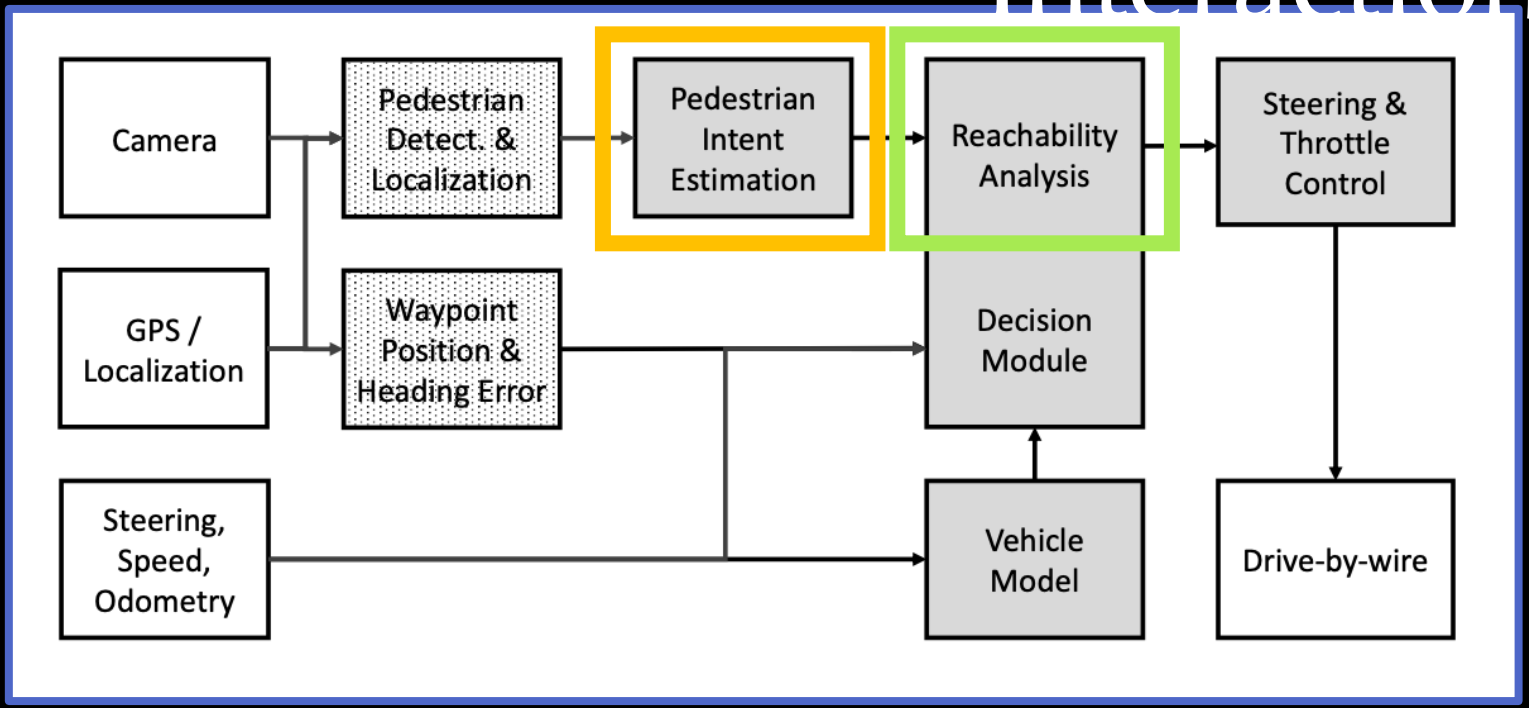
Online Monitoring for Vehicle-Pedestrian Interaction

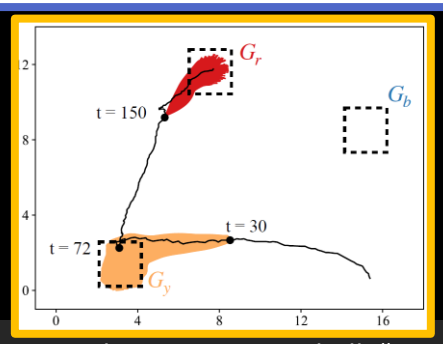


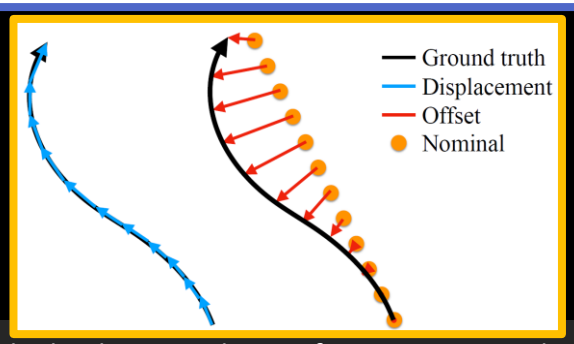


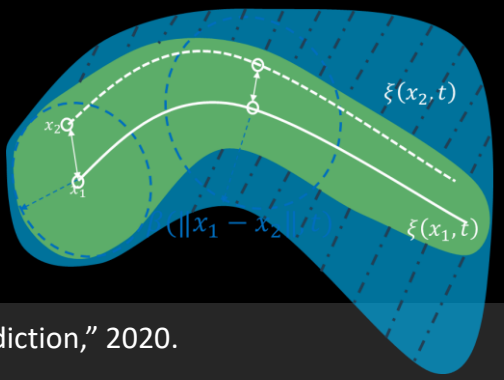
P. Du, Z. Huang, T. Liu, T. Ji, K. Xu, Q. Gao, H. Sibai, K. Driggs-Campbell, and S. Mitra. “Online Monitoring for Safe Pedestrian-Vehicle Interactions.” ITSC 2020.

Online Monitoring for Vehicle-Pedestrian Interaction









Z. Huang, A. Hasan, and K. Driggs-Campbell. "Intention-aware Residual Bidirectional LSTM for Long-term Pedestrian Trajectory Prediction," 2020.
S. Duggirala, S. Mitra, and M. Viswanathan, "Verification of annotated models from executions," EMSOFT 2013.
C. Fan, B. Qi, and S. Mitra, "Data-driven formal reasoning and their applications in safety analysis of vehicle autonomy features," IEEE Design & Test, 2018.