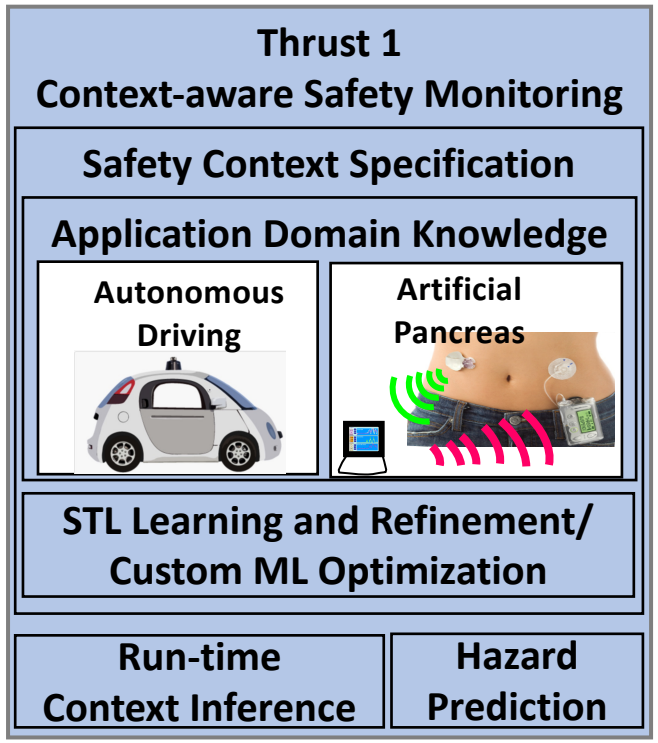
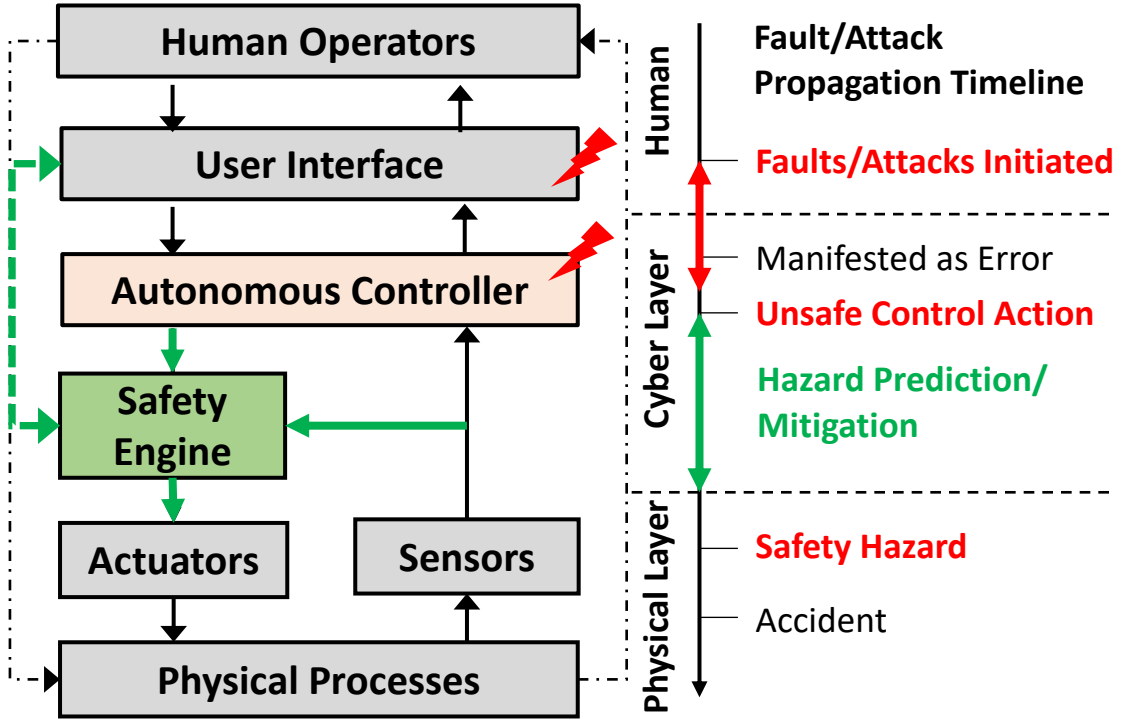
Context-aware Assurance in Cyber-Physical Systems

Xugui Zhou
Electrical and Computer Engineering
University of Virginia
Advisor: Homa Alemzadeh

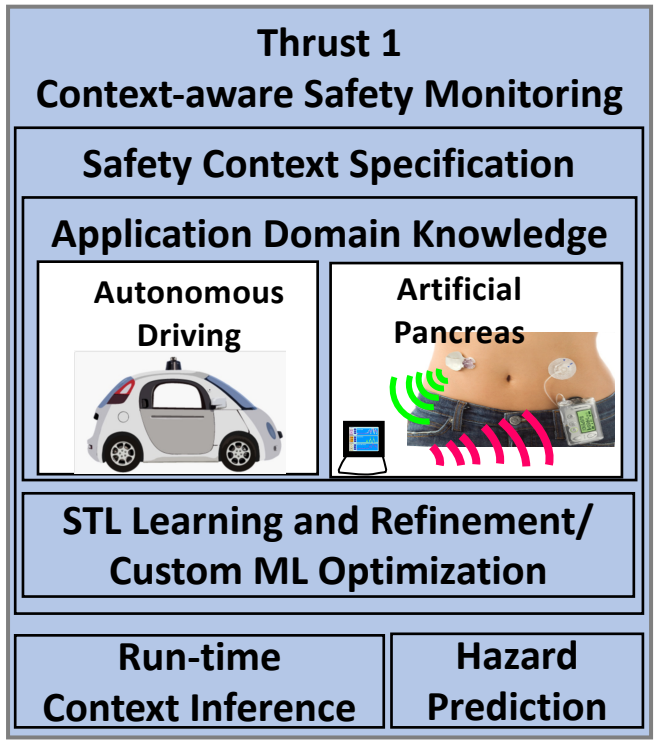
Overall Methodology



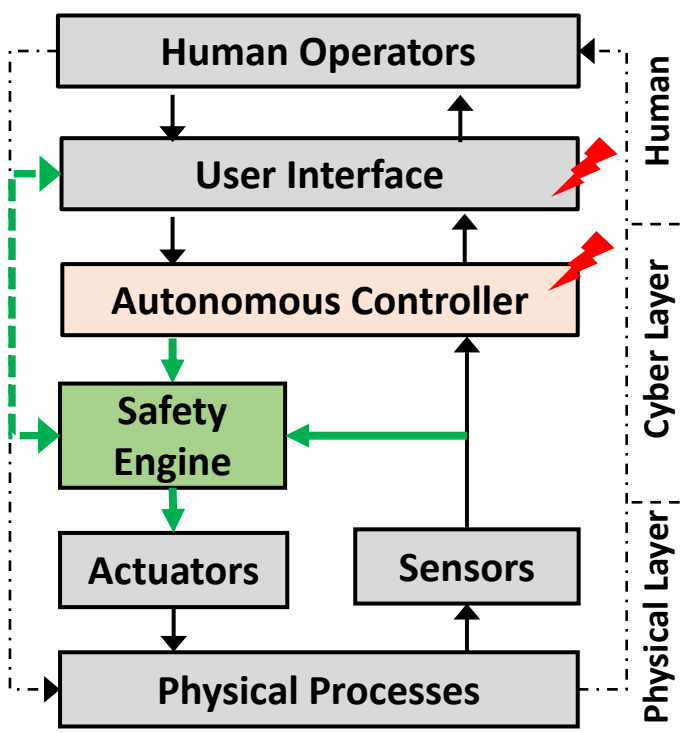
DSN21'



Overall Methodology

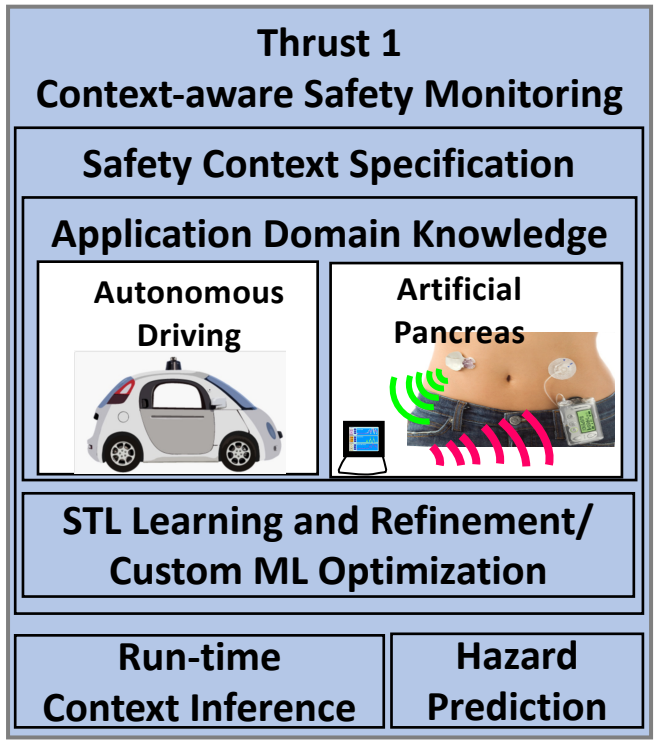


DSN21'

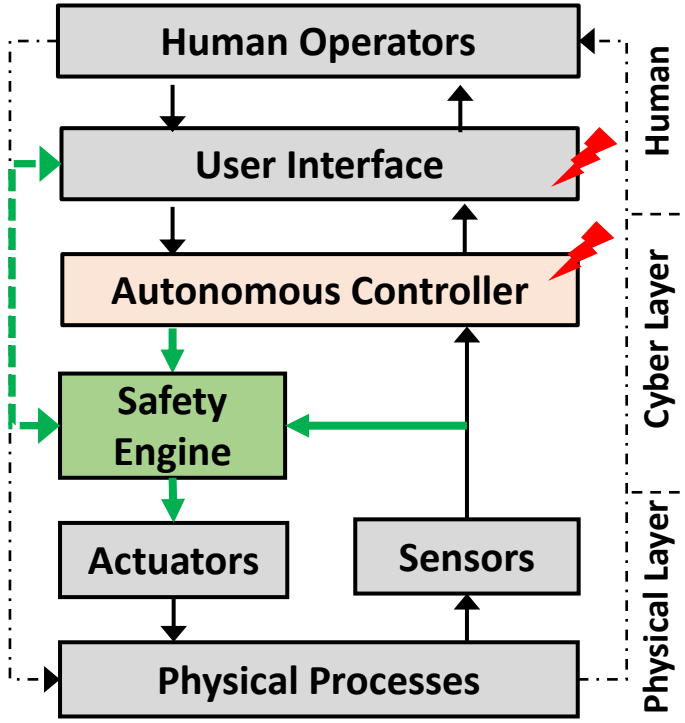


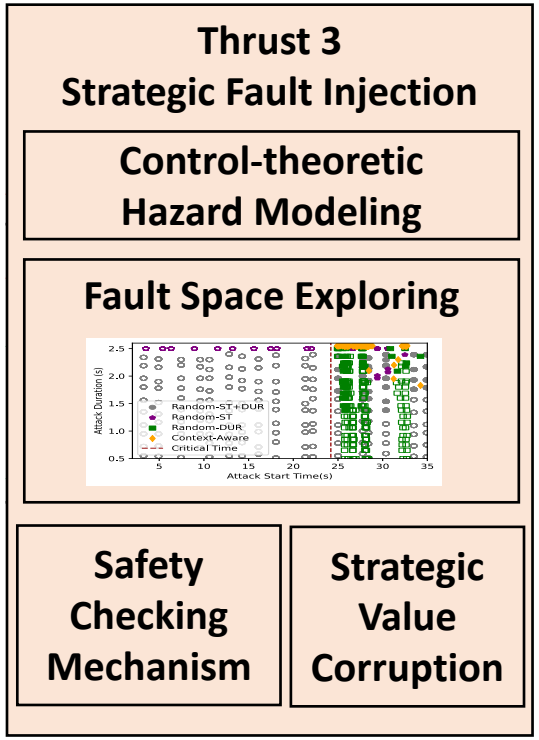
- Find safety defects in testing stage

Overall Methodology



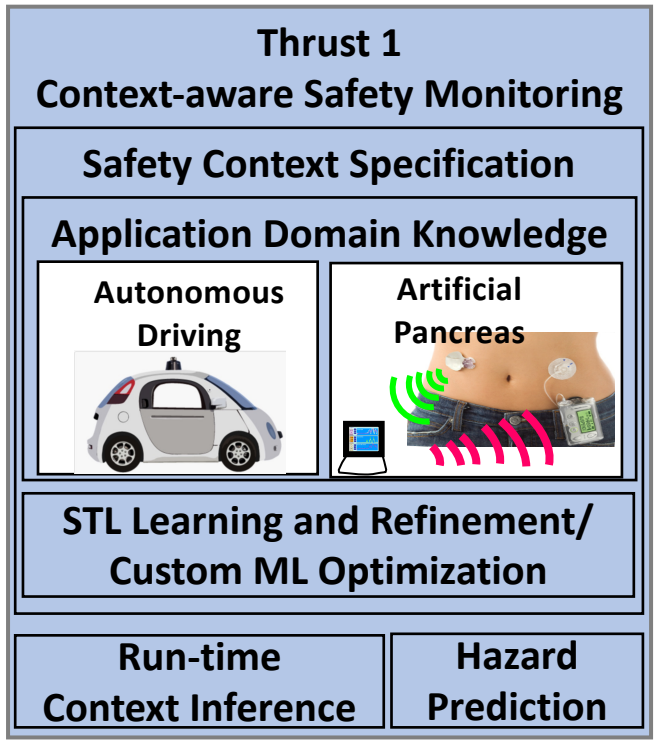
DSN21'



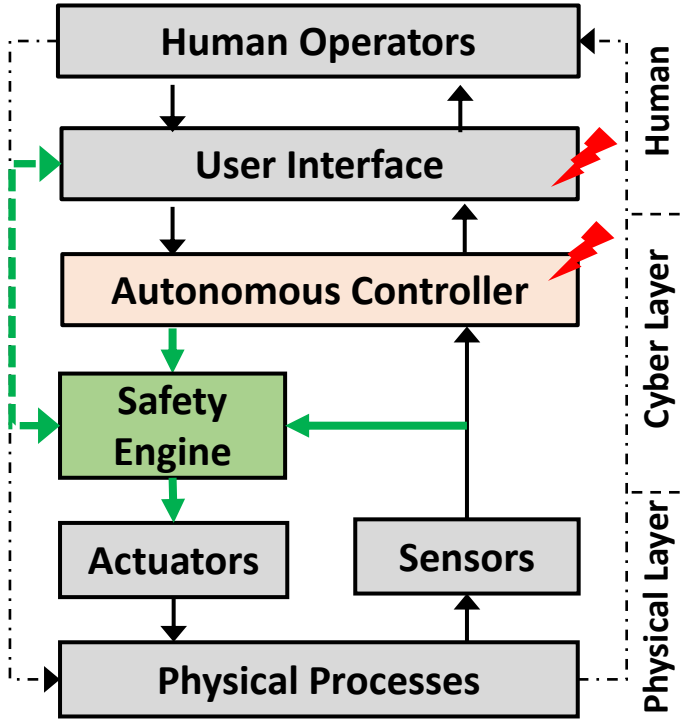


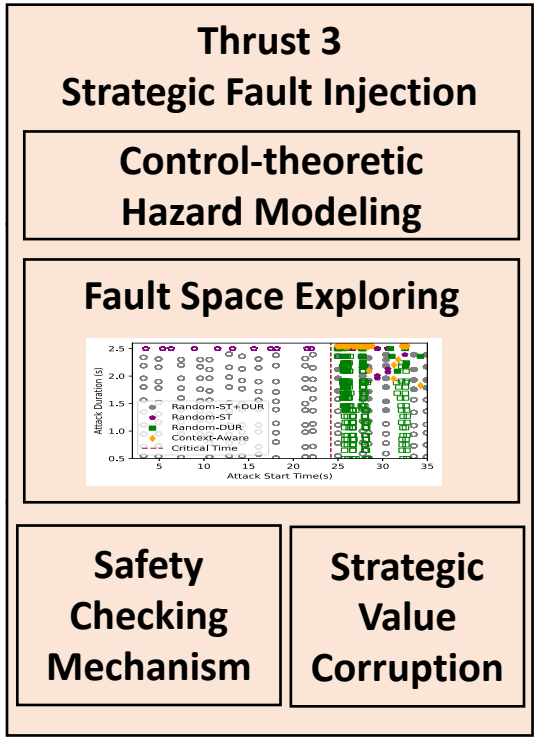
DSN22'

Overall Methodology



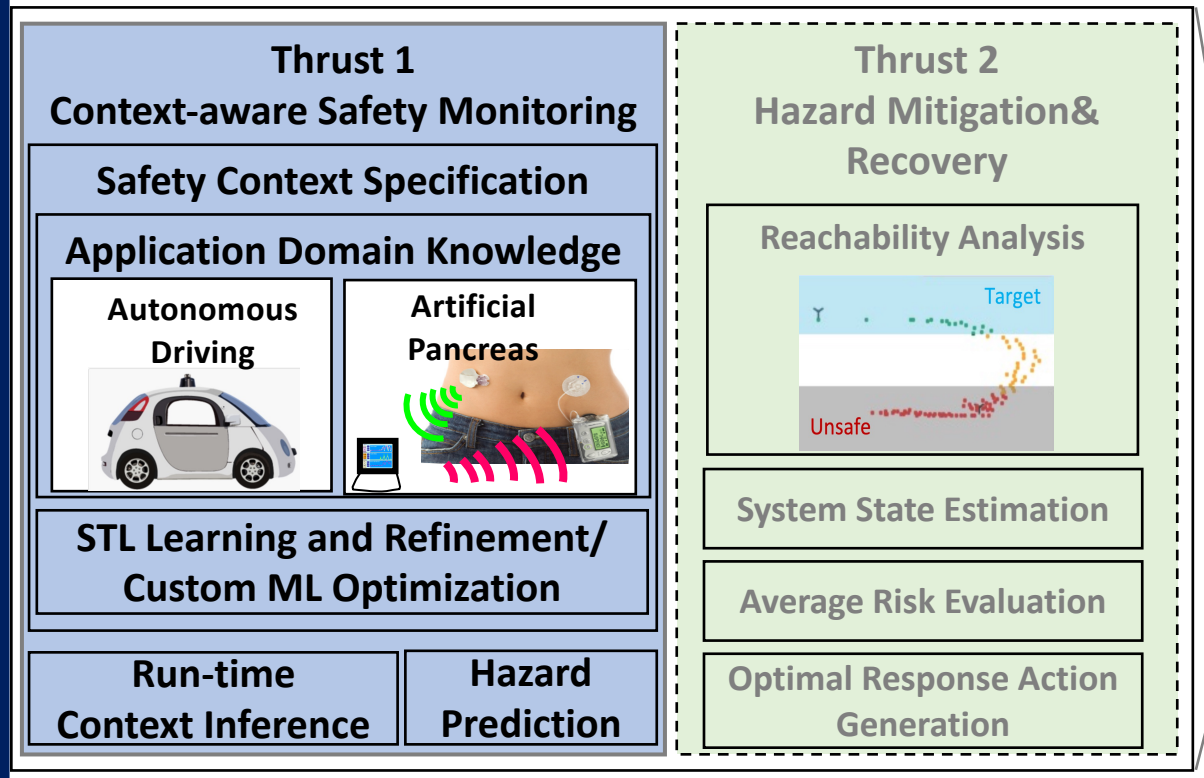
DSN21'





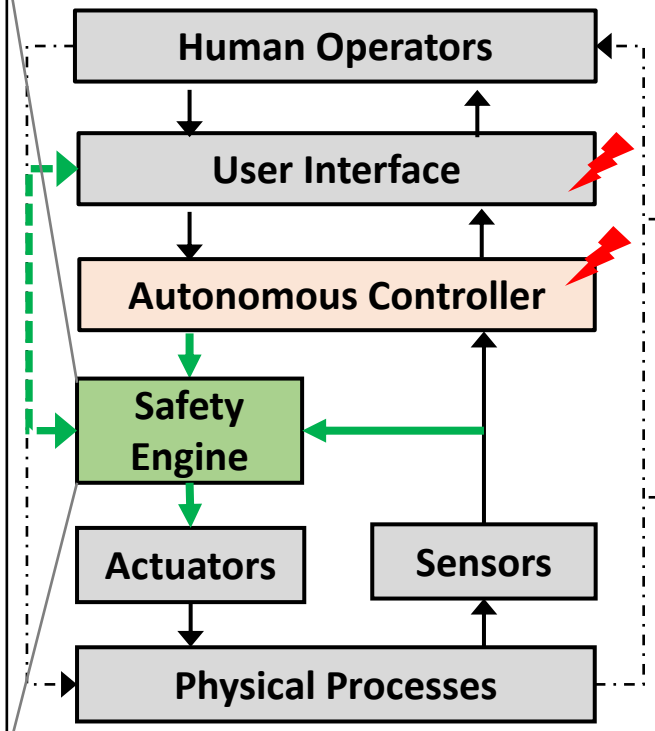
DSN22'

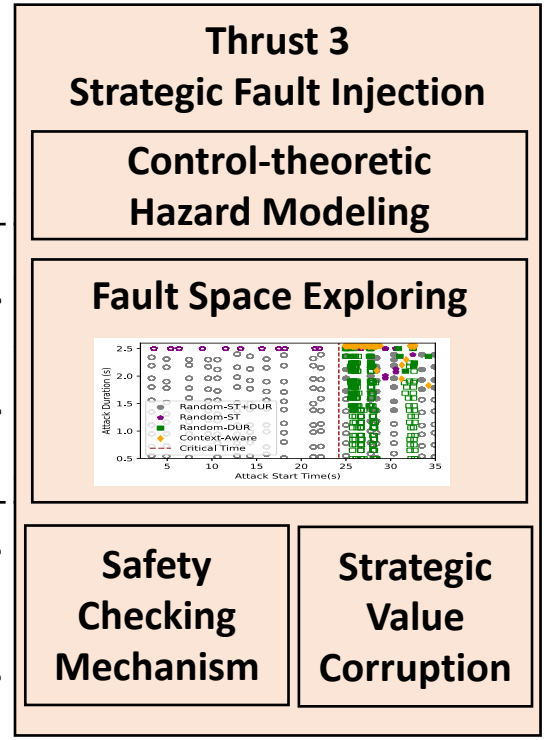
Overall Methodology



DSN21'

Data-driven Design of Context-aware Monitors for Hazard Prediction





DSN22'

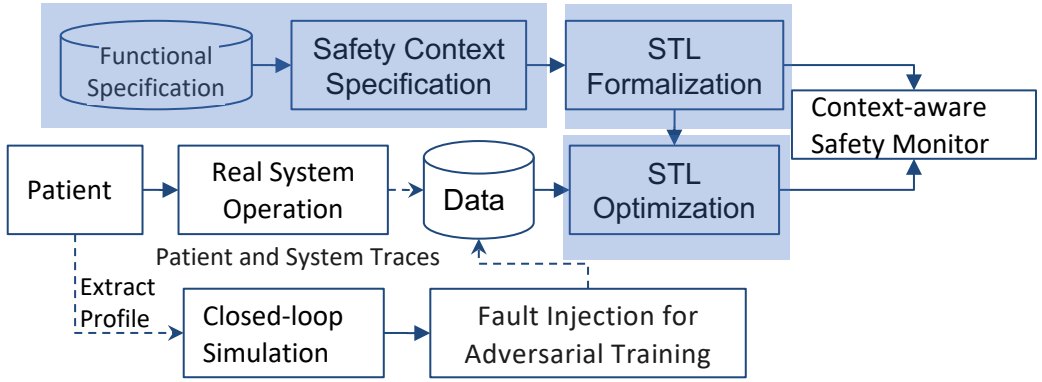
Strategic Safety-Critical Attacks Against an Advanced Driver Assistance System

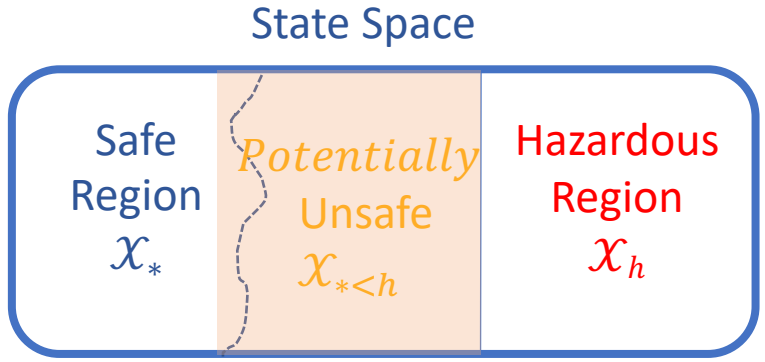
Safety Context Specification and Learning

- **Safety Context Specification (SCS)**
  - Control-theoretic Hazard Analysis Method
- **A Formal Framework to Generate SCS**
- **Formalization with Signal Temporal Logic**
- **Optimization of STL Formulas**
  - Tight Mean Exponential loss function (TMEE)

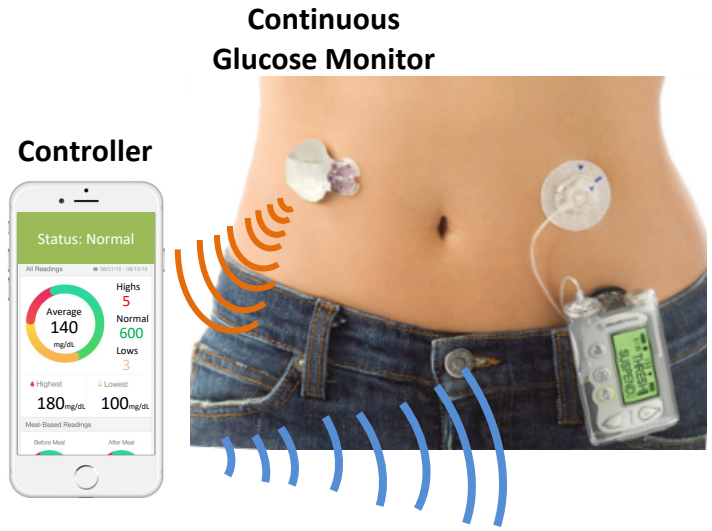
$$loss(r) = E[e^{-r} + r - \frac{1}{1+e^{-2r}}],\ r = \mu_i(d(t)) - \beta_i \quad (4)$$

  - Learn tight thresholds that
    - Under-approximate the safe region
    - Over-approximate the unsafe region

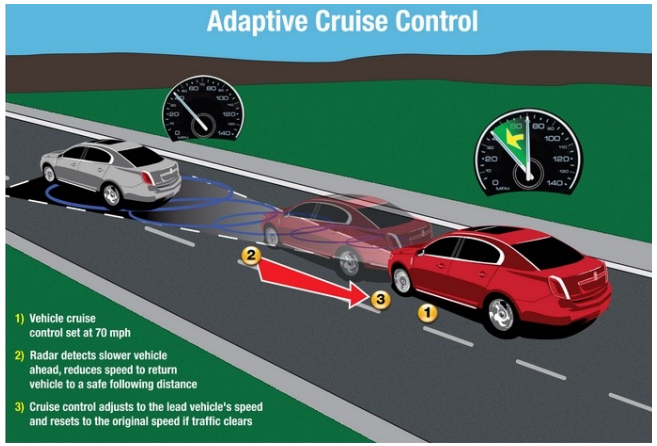




Case Studies



Artificial Pancreas System (APS)



Autonomous Driving System (ADS)

Results

TABLE 5: Performance of CAWT Monitor vs. Non-ML Monitors

Simulator	Monitor	No. Sim.	Hazard%	FPR	FNR	ACC	F1 Score
Glucosym	Guideline	8820	33.90%	0.02	0.32	0.95	0.72
	MPC	8820	33.90%	0.02	0.34	0.95	0.71
	CAWOT	8820	33.90%	0.01	0.30	0.96	0.81
	CAWT	8820	33.90%	**0.01**	**0.02**	**0.99**	**0.96**
T1DS2013	Guideline	8820	39.30%	0.07	**<0.01**	0.93	0.75
	MPC	8820	39.30%	<0.01	0.02	**1.00**	0.96
	CAWOT	8820	39.30%	0.02	0.04	0.98	0.89
	CAWT	8820	39.30%	**<0.01**	0.03	**1.00**	**0.97**
OpenPilot	MPC	4800	39.9%	0.01	0.90*	0.79	0.17
	CAWOT	4800	39.9%	0.29	0.12	0.76	0.66
	CAWT	4800	39.9%	**<0.01**	**0.05**	**0.99**	**0.97**

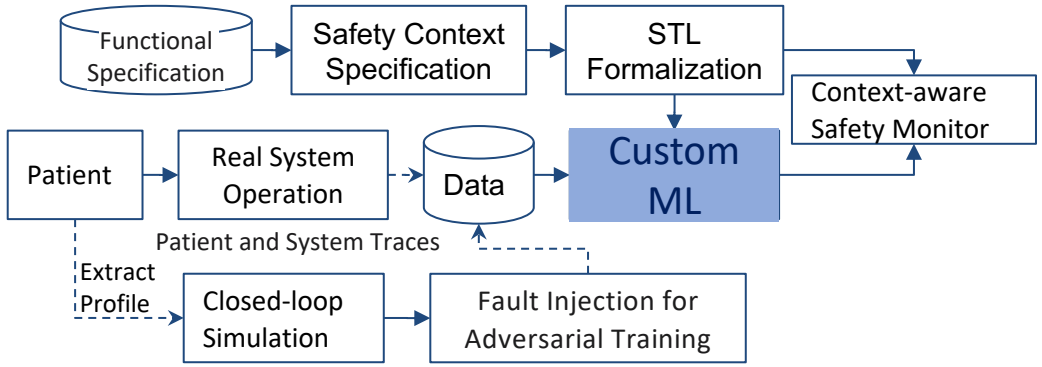
MPC: Model Predictive Control
Guideline: Medical Guideline with fixed non-patient-specific threshold
CAWT: Context-aware with refined thresholds
CAWOT: Context-aware without refined thresholds

Training with Custom Loss Functions

- ML model for anomaly detection

$$y_t = p(\exists t' \in [t, t+T] : x_{t'} \in \mathcal{X}_h | f(X_t), f(U_t))$$

  - Control action sequence U_t
  - System state sequence X_t
  - Binary output y_t: safe (0) or unsafe (1):



- Custom loss function

$$loss = loss_{ex} + w \left| y_t - I\left(\bigvee_{\Phi_h \in UCAS} f(\mu(X_t)) \models \Phi_h \right) \right|$$

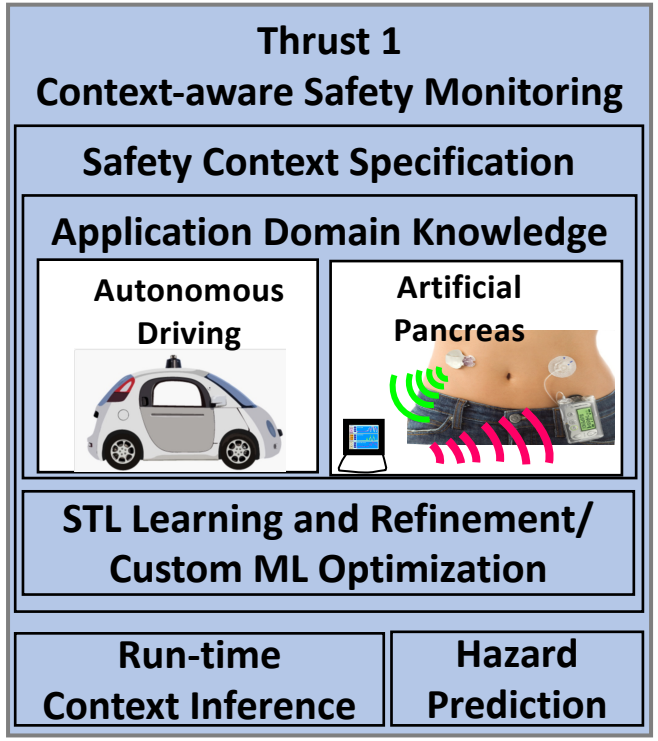
  - Indicates whether any of the STL formulas are satisfied over the measurement window
  - Enforces the satisfaction of safety properties also helps with interpretability of ML-based monitors

Xu, J., Zhang, Z., Friedman, T., Liang, Y. and Broeck, G., "A semantic loss function for deep learning with symbolic knowledge," *International conference on machine learning (PMLR)*, 2018.
X. Zhou, B. Ahmed, J. H. Aylor, P. Asare, H. Alemzadeh, "Hybrid Knowledge and Data Driven Synthesis of Runtime Monitors for Cyber-Physical Systems," *Under Review*

Results

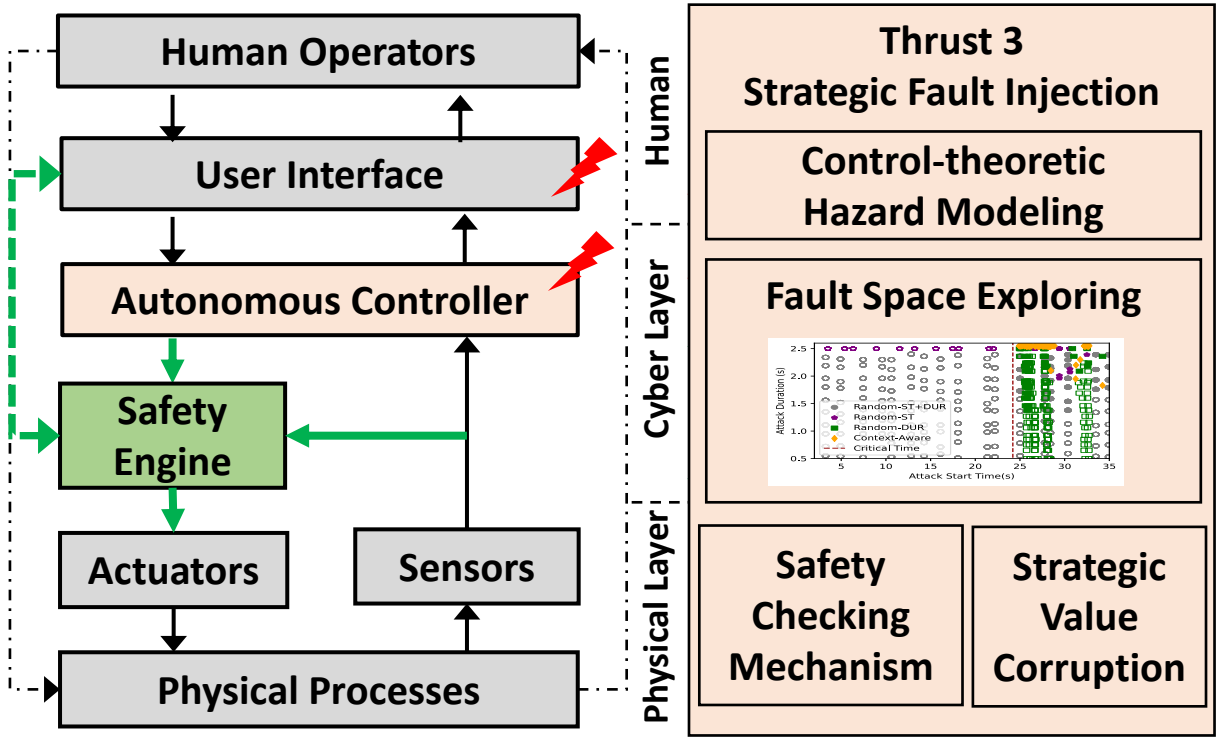
Simulator	Metric	Sample Level (Tolerance Window)				Simulation Level (Two Regions)			
	Monitor	FPR	FNR	ACC	F1 Score	FPR	FNR	ACC	F1 Score
Glucosym	MLP	0.02	0.07	0.97	0.89	0.13	0.06	0.89	0.79
	LSTM	0.04	0.06	0.96	0.81	0.16	0.06	0.87	0.78
	CAWT	0.01	**0.02**	**0.99**	**0.96**	0.10	**0.01**	0.92	0.86
	MLP_Custom	0.02	0.05	0.98	0.91	0.12	0.05	0.90	0.81
	LSTM_Custom	**0.00**	0.23	0.97	0.86	**0.03**	0.15	**0.94**	**0.87**
T1DS 2013	MLP	<0.01	0.56	0.94	0.71	0.05	0.28	0.90	0.78
	LSTM	<0.01	0.06	0.99	0.95	0.08	0.06	0.93	0.87
	CAWT	<0.01	**0.03**	**1.00**	**0.97**	0.06	**0.02**	0.95	0.91
	MLP_Custom	0.01	0.27	0.96	0.82	0.10	0.18	0.88	0.78
	LSTM_Custom	**0.00**	0.17	0.98	0.90	**0.02**	0.10	**0.96**	**0.92**
Open Pilot	MLP	0.01	0.11	0.97	0.93	0.06	0.09	0.94	0.88
	LSTM	0.01	<0.01	**1.0**	**0.99**	0.05	<0.01	**0.96**	**0.93**
	CAWT	**<0.01**	0.05	0.99	0.97	**0.04**	0.05	**0.96**	**0.93**
	MLP_Custom	0.01	0.19	0.96	0.88	0.06	0.15	0.91	0.84
	LSTM_Custom	0.03	**0.00**	0.98	0.95	0.18	**0.00**	0.87	0.80

Overall Methodology



DSN21'

Data-driven Design of Context-aware Monitors for Hazard Prediction



DSN22'

Strategic Safety-Critical Attacks Against an Advanced Driver Assistance System

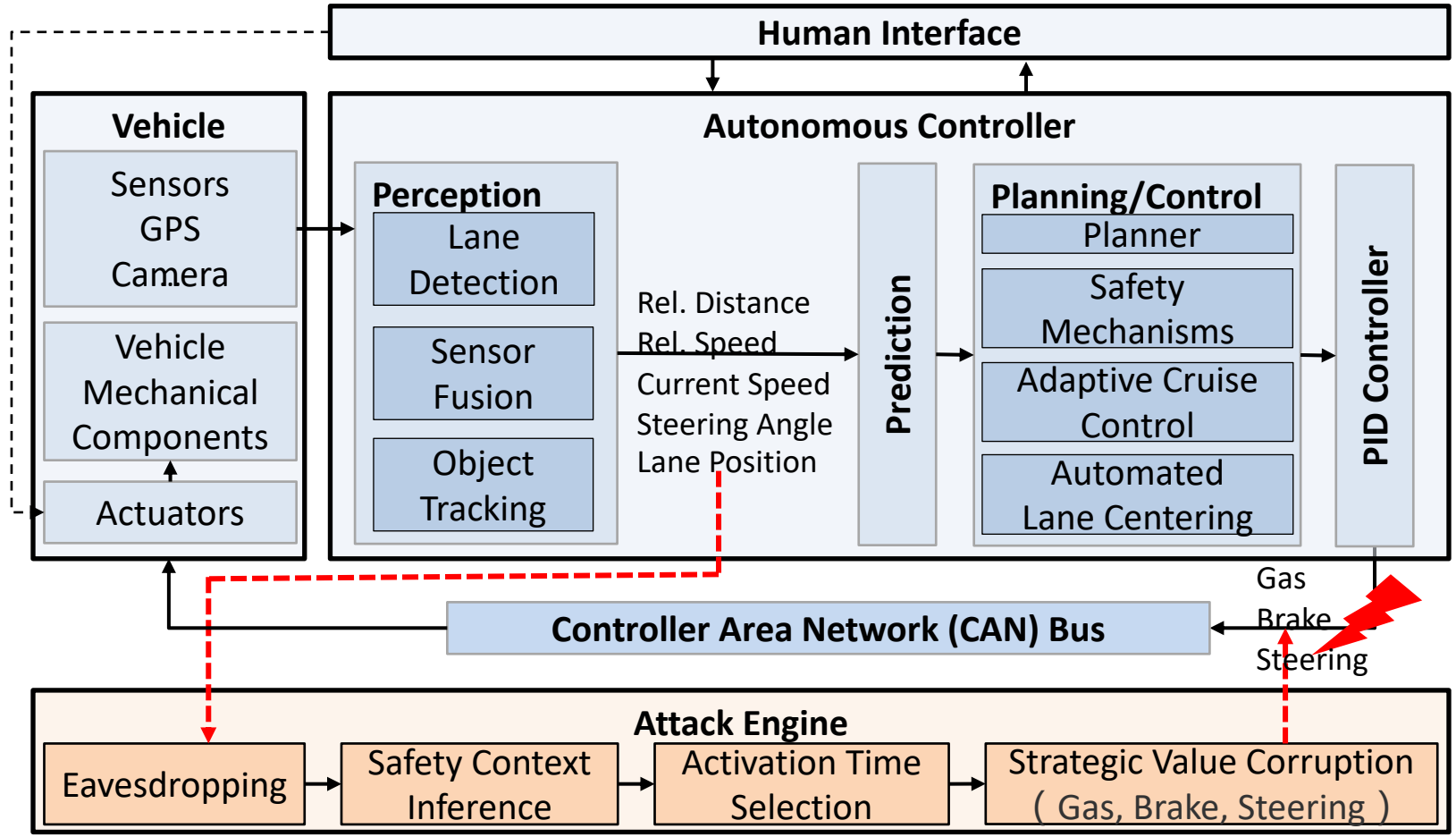
Strategic Safety-Critical Attacks

- **Technical Problem**
  - Find the most salient safety-critical scenarios from the fault/attack parameter space (e.g., activation time, duration, error value) as **efficiently** and **realistically** as possible.
- **Limitations of Existing Solutions on AV Safety**
  - Real road testing is time and resource consuming with high risk
  - More efficient works relying on simulation
    - Focus on level 3+
    - Without considering driver intervention
    - Without using realistic control software
    - Require a large amount of training data (ML-based approach)

Contributions

- **Propose a context-aware safety-critical attack strategy that can find the most critical contexts and attack types to corrupt the ADAS (L2) outputs:**
  - **Goals:**
    - Maximize the chance of hazards
    - Cause hazards as soon as possible without being detected
  - **Merits**:
    - Exploring the fault parameter space which is impossible to mine using random techniques
    - Model-based method, less training data requirements than ML-based approaches
  - **Applications**:
    - Safety checking for validation
- **Develop a closed-loop simulation platform with "a real ADAS control software" and "a driver simulator" to assess the resilience of a widely-used L2 ADAS, OpenPilot.**
- **Demonstrate the effectiveness of the proposed attack strategy in comparison to several random attacks.**

Overview

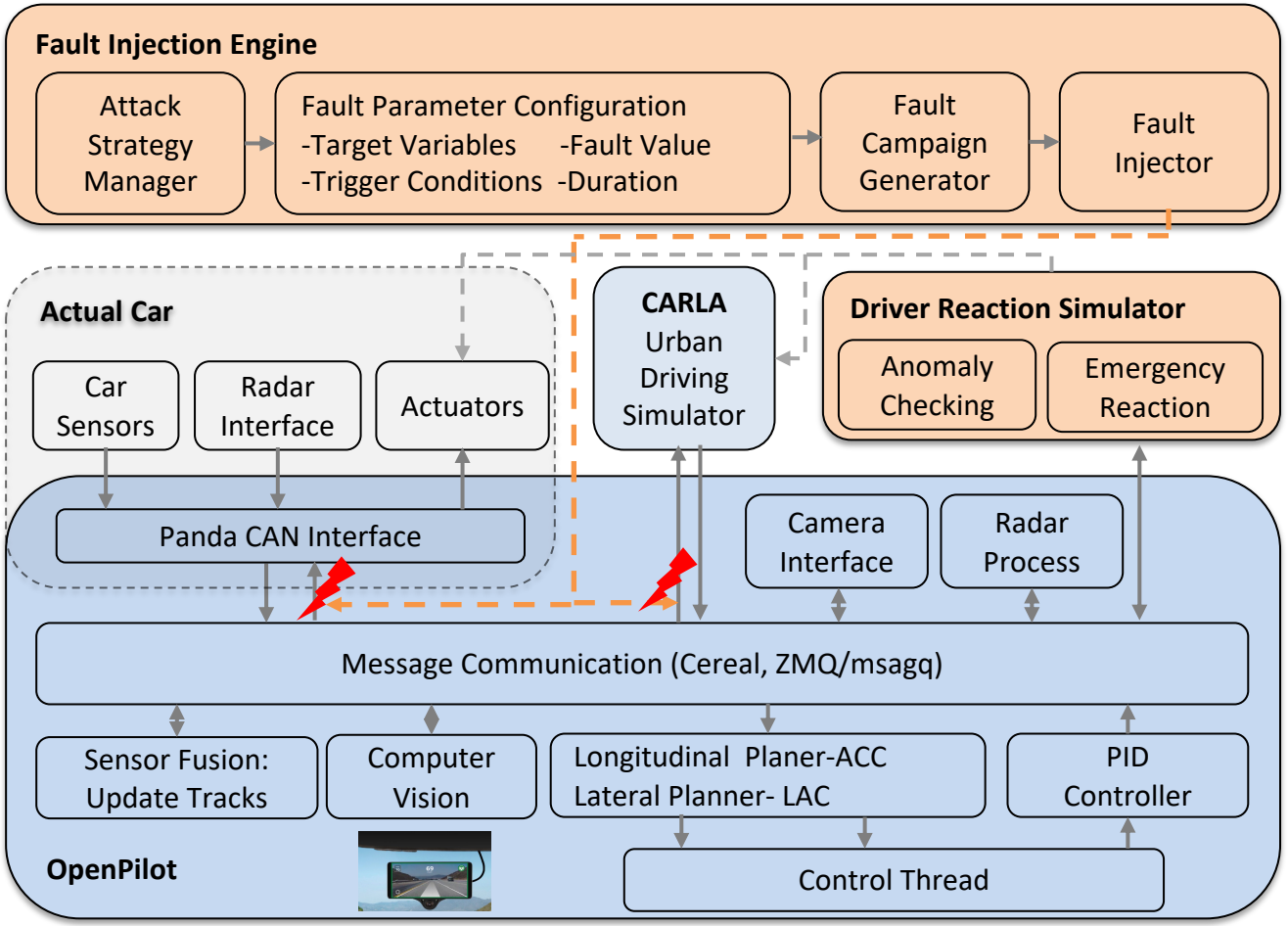

Human Interface
Vehicle
Sensors
GPS
Camera
Vehicle
Mechanical
Components
Actuators
Autonomous Controller
Perception
Lane
Detection
Sensor
Fusion
Object
Tracking
Rel. Distance
Rel. Speed
Current Speed
Steering Angle
Lane Position
Prediction
Planning/Control
Planner
Safety
Mechanisms
Adaptive Cruise
Control
Automated
Lane Centering
PID Controller
Gas
Brake
Steering
Controller Area Network (CAN) Bus
Attack Engine
Eavesdropping
Safety Context
Inference
Activation Time
Selection
Strategic Value Corruption
(Gas, Brake, Steering)

Approach: Attack Model

- **Assumptions**
  - Access to the sensor measurements
  - The capability of modifying the actuator commands with faulty values
- **Possible Entries**
  - Wireless networks (e.g., over-the-air updates)
  - In-vehicle networks (CAN, FlexRay, Ethernet, Bluetooth, or telematics devices)
  - Vehicle to everything communication
  - Vulnerable components supplied by different vendors

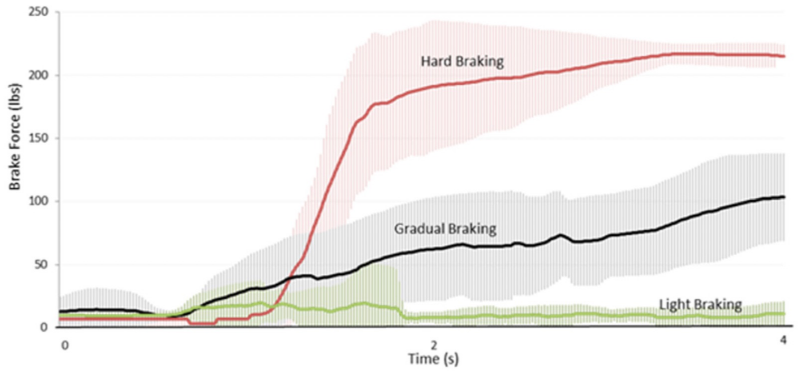
Experiment Setup: Open-source Research Platform

- Real ADAS control software
  - Panda is not activated
- CARLA urban driving simulator
  - Radar is not integrated
- **Driver behavior simulator**
- **Fault-injection engine**



Experiment Setup: Driver Reaction Simulator

Activation Condition	Driver Reaction	Reaction Time
• ADAS Safety Warnings	Emergency brake for FCW	2.5 seconds (the average driver reaction time reported in AV literature)
• Hard brake	Stop brake, output regular gas amount, without changing the steering angle	
• Unexpected increase in acceleration • Unexpected increase in steering angle • Unsafe cruising speed	Emergency brake (gas=0) $brake = e^{10t-12}/(1+e^{10t-12})$ (4)	



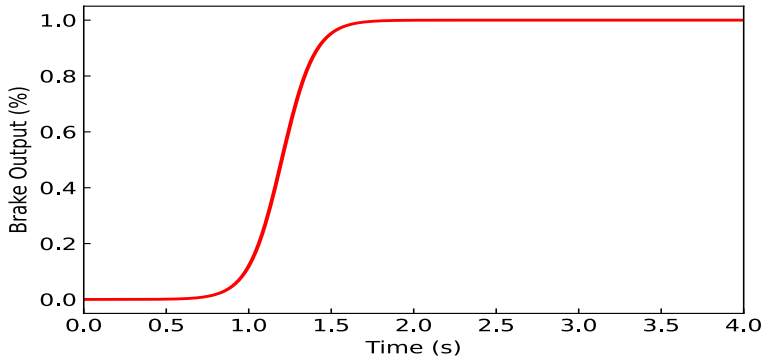


Figure Source: J. G. Gaspar and D. V. McGehee, "Driver brake response to sudden unintended acceleration while parking," Transportation Research Interdisciplinary Perspectives, vol. 2, p. 100039, 2019.

Results: Overall Performance

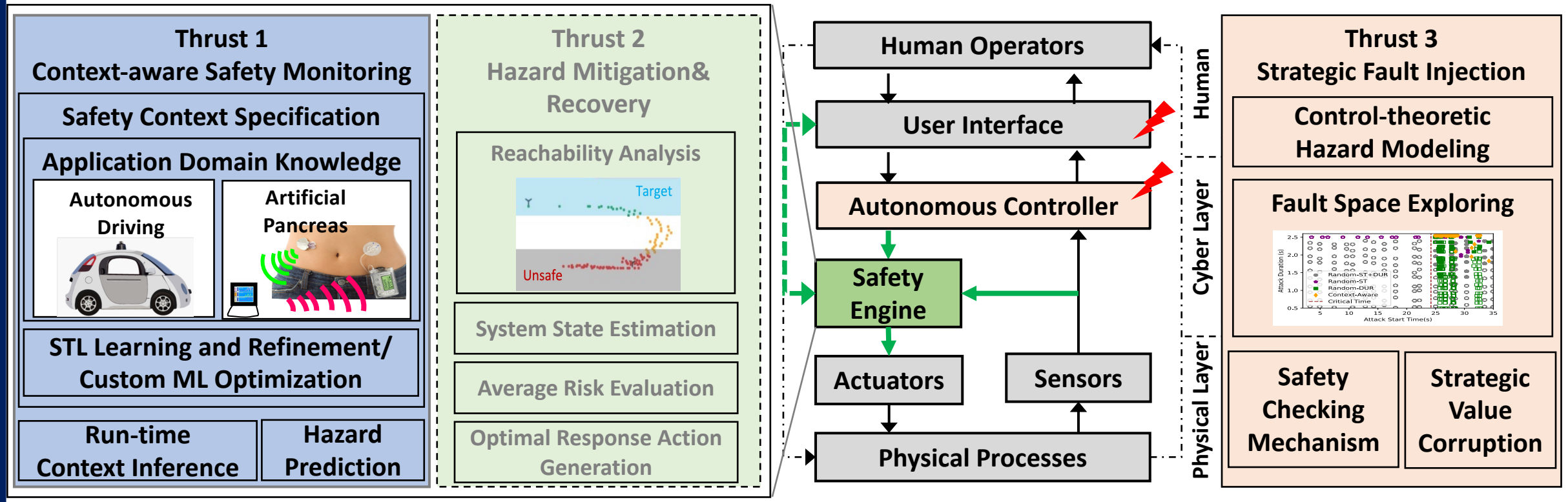
Attack Strategy	Alerts	Hazards	Accident	Hazards& no Alerts	LaneInvasion (No. Event/s)	TTH(s) (Avg. ± Std.)
No Attacks	2 (0.1%)	0	0	0	0.46	
Random-ST+DUR	3248 (22.6%)	5727 (39.8%)	3293 (22.9%)	3083 (21.4%)	1.03	1.61±1.96
Random-ST	346 (24.0%)	771 (53.5%)	516 (35.8%)	474 (32.9%)	0.68	1.49±0.73
Random_DUR	210 (14.6%)	388 (26.9%)	332 (23.1%)	229 (15.9%)	0.46	1.92±1.17
Context-Aware	4 (0.3%)	1201 (83.4%)	641 (44.5%)	1197 (83.1%)	0.66	2.43±1.29

- **Random-ST+DUR:** Random Start Time and Duration
- **Random-ST:** Random Start Time and Fixed Duration
- **Random-DUR:** Random Duration with Context-aware Start time
- **Context-Aware:** Context-aware Start Time and Duration

Key Observations

- Lane invasions happen without any attacks.
- Forward collision warning does not get activated.
- Human intervention helps prevent hazards and accidents.
- Steering angle is the most vulnerable attack target.
- The Context-Aware attack strategy is efficient in
  - Exploring safety critical states in the fault space.
  - Evading human driver detection and ADAS safety checks through strategic value corruption.
- **Broader Impact:**
  - Help improve safety checks of ADAS.

Overall Methodology



DSN21'

Data-driven Design of Context-aware Monitors for Hazard Prediction

DSN22'

Strategic Safety-Critical Attacks Against an Advanced Driver Assistance System

Thank you!

xugui@virginia.edu

Test Platform Available Online:
https://github.com/UVA-DSA/openpilot-CARLA/
https://github.com/UVA-DSA/ContextSafetyMonitorAPS