VIRGINIA TECH

DEVCOM ARMY RESEARCH LABORATORY

RAIL: Risk-Aware Human-in-the-Loop Framework with Adaptive Intrusion Response for Autonomous Vehicles

Dawood Wasif[1] Terrence J. Moore[2] Seunghyun Yoon[3] Hyuk Lim[3]
Dan Dongseong Kim[4] Frederica F. Nelson[2] **Jin-Hee Cho**[1]

[1]Virginia Tech [2]U.S. DEVCOM Army Research Laboratory
[3]Korea Institute of Energy Technology [4]University of Queensland

USENIX VehicleSec 2026

When Detection, Control, and Learning Decouple



A compromised ECU injects a bounded steering bias on the CAN bus. The trajectory drifts within a **few control cycles**, yet the policy gets **no structured signal** attributing the fault, and the operator has **not yet perceived** it.

Detection, control, and learning stay **loosely coupled**: safety is penalized *after* a violation instead of governed as a runtime variable.

Why Existing Approaches Fall Short

Implicit Safety

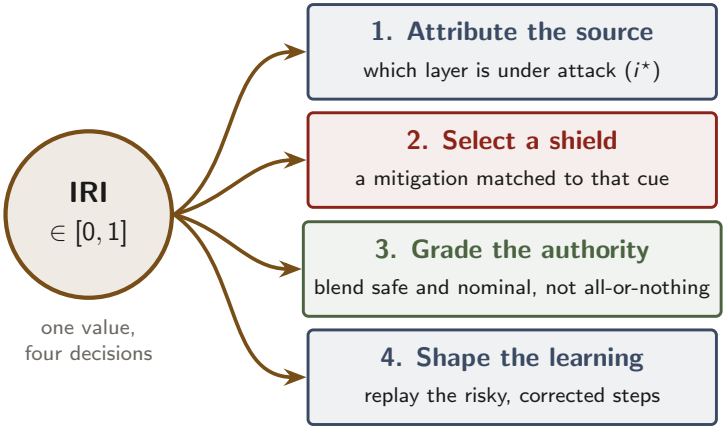
RL, Safe RL, offline, and imitation encode safety as **scalar penalties**. They degrade under sparse cost and cannot say *what* failed or *which* channel is compromised (Achiam17; Kumar20; Bain95).

Reactive Human Oversight

Human-in-the-loop methods treat takeover as a **binary override**, fired only after the operator perceives the threat. No preemptive, graded, or source-attributed response (Kelly19; Li22; Huang24).

Neither exposes **decomposed risk** nor a **graded response**, and neither provides an auditable trail for standards such as UN R155 and ISO/SAE 21434.

Key Insight: Risk as a First-Class Variable



One **interpretable, decomposable** risk index turns anomaly detection into calibrated, source-aware control, closing the **detect, respond, learn** loop in a single system.

System and Threat Model

An ego AV under a **runtime safety layer** with a human operator as backstop.

- **Black-box**, non-persistent adversary in bounded bursts.
- Two documented attack classes across the kill chain.

Three cues span the AV kill chain

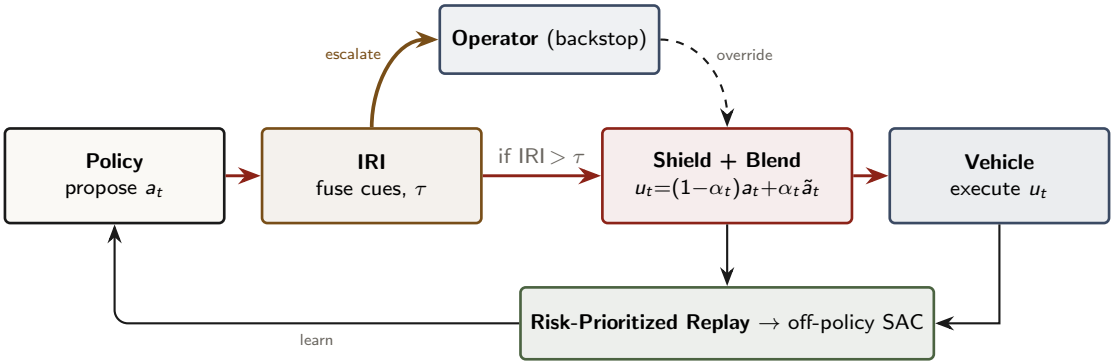
r_{CURV} **(actuation)** catches CAN injection and actuator faults

r_{TTC} **(planning)** catches phantom obstacles and disabled braking

r_{OOD} **(perception)** catches LiDAR spoofing and sensor blinding

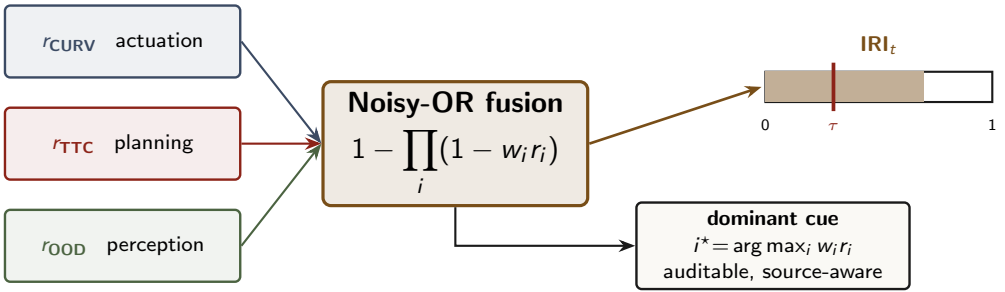
Evaluated attacks: **CAN injection** *on actuation and* **LiDAR spoofing** *on perception, in 5 s bursts every 30 s.*

RAIL: A Detect–Respond–Learn Loop



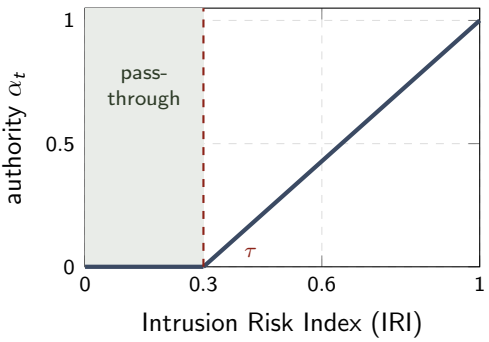
Low risk passes through ($u_t = a_t$); elevated **IRI** triggers graded shielding; the operator is a rare backstop; every transition feeds learning.

The Intrusion Risk Index (IRI)



A single strong anomaly is enough to raise IRI; weak consistent signals compound non-linearly. Fixed weights $\sum_i w_i = 1$; low cue correlation supports the Noisy-OR conditional-independence assumption. The index is **modular**: new cues plug in with a normalized r_i and a matching shield.

Graded Authority, Not Binary Takeover



Executor blends nominal and shielded actions:

$$u_t = (1 - \alpha_t)\, a_t + \alpha_t\, \tilde{a}_t$$

with α_t **monotone in IRI**. A contextual bandit picks the cue-specific shield $\tilde{a}_t$.

Guarantees: *passivity* (α_t=0 below τ), *monotonicity* (risk never lowers shielding), and *bounded deviation* ($\|u_t - a_t\| \leq \|\tilde{a}_t - a_t\|$).

Safety-Aware Learning

Dual reward shaping

Learns from both **safety failures** and **early warning signs**. Collisions, off-road events, and takeovers are penalized, while high-risk behavior and unnecessary shield intervention are also discouraged.

Risk-prioritized replay

Gives **more attention to dangerous experiences**. Near misses, high-risk situations, and takeovers are revisited more often so the AV can learn from them and avoid similar situations in the future.

Takeovers and near misses become **structured feedback** in an off-policy SAC backbone, accelerating safety alignment while preserving nominal behavior.

Evaluation Setup

Two simulators, **MetaDrive** and **CARLA**, with **12 baselines** across three families:

- RL and Safe RL (SAC/PPO, Lagrangian, CPO)
- offline RL and imitation (CQL, BC, GAIL)
- conventional and modern HITL (HG-DAgger, IWR, HACO, HAIM-DRL)

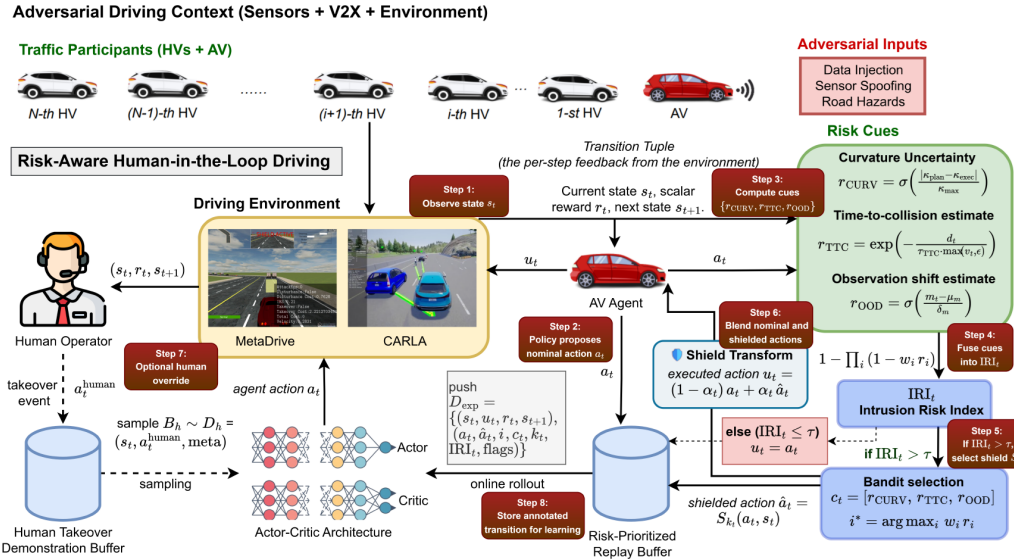
Only RAIL fuses, adapts, and decomposes

Capability	CBF	HJ	Simplex	Shield	RAIL
Multi-source fusion	✗	✗	✗	✗	✓
Online adaptation	✗	✗	✗	✗	✓
Human-feedback integration	✗	✗	✗	✗	✓
Graded blending	✗	✗	✗	△	✓
Risk decomposition	✗	✗	✗	△	✓
Model-free	✗	✗	△	△	✓

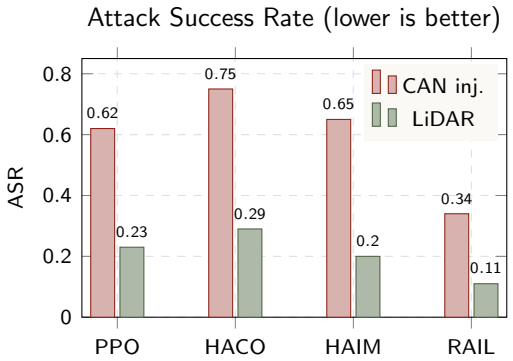
✓ full △ partial ✗ none

Runs the system 10 times per second, repeats each experiment five times for reliability, and tests it during normal driving and two types of attacks: CAN injection and LiDAR spoofing.

Overall RAIL Architecture



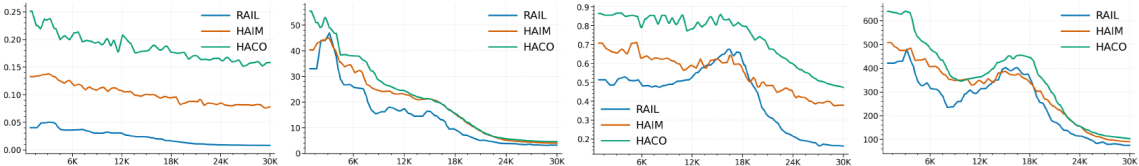
Results: Robustness Under Attack



CAN injection	TSR↑	ASR↓
PPO	0.23	0.62
HACO	0.58	0.75
HAIM	0.62	0.65
RAIL	**0.68**	**0.34**

Attack success cut by **up to 48%** and disengagement by **over 50%** vs. the strongest HITL baseline (HAIM), while matching expert-level success.

Training Dynamics



(a) Disturbance rate (b) Disturbance count (c) Takeover rate (d) Takeover count

On MetaDrive, **RAIL** reduces disturbance and operator takeovers **fastest** and reaches the lowest steady state across all four metrics, compared with HAIM-DRL and HACO.

Ablation, Sensitivity, and Cost

Every component earns its place

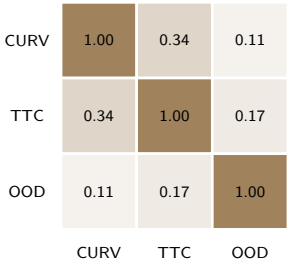
(SR: Success Rate; SV: Safety Violation; DR: Disturbance Rate)

Variant	SR↑	SV↓	DR↓
w/o TTC cue	0.67	1.02	0.022
w/o CURV cue	0.78	0.80	0.014
w/o bandit	0.79	1.10	0.018
w/o shield pen.	0.74	0.72	0.004
RAIL full	**0.85**	**0.75**	**0.003**

Dropping TTC hurts most. Shielding, not the operator, drives robustness.

Cue correlations ρ

(pairwise Spearman correlation)



All off-diagonal $\rho \leq 0.34$: cues are complementary.

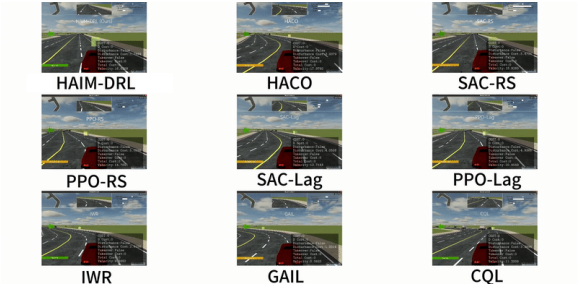
Runtime & workload

(MTBT: Mean Time Between Takeover)

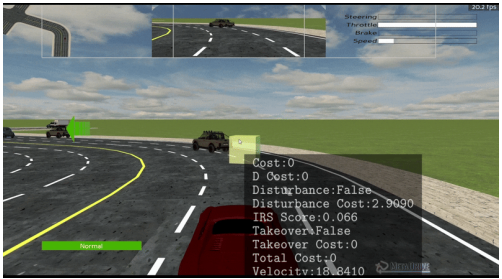
Overhead	0.22% cycle
Shield precision	85.8%
MTBT	310 steps
Operator active time	3.1 s

Interpretable risk at negligible cost, with rare, well-targeted takeovers.

Qualitative Comparison in MetaDrive



Nine methods drive the same route under attack



RAIL runtime view with on-screen risk telemetry

RAIL holds the lane with a **low risk score** and **no operator takeover**, while several baselines drift or disengage under the same attack.

Conclusion

- One **decomposable Intrusion Risk Index** unifies detection, graded shielding, escalation, and learning in a single runtime loop.
- Under CAN injection and LiDAR spoofing: attack success reduced **up to 48%**, disengagement **over 50%**, matching expert-level success (0.85).
- Risk becomes **graded, auditable control** rather than binary fail-safes, aligned with UN R155 and ISO/SAE 21434.
- Artifact **Available and Functional**, archived on Zenodo.

Thank you!

Questions?

Presented by: Jin-Hee Cho (jicho@vt.edu)
Virginia Tech, Department of Computer Science

Joint work with Dawood Wasif (Virginia Tech); Terrence J. Moore and Frederica F. Nelson (ARL); Seunghyun Yoon and Hyuk Lim (KENTECH); and Dan Dongseong Kim (Queensland)

Artifact (Zenodo): 10.5281/zenodo.20451598

Acknowledgement: Army Research Office Grant No. W911NF-24-2-0241.