

U.S. ARMY

DEVCOM ARMY RESEARCH LABORATORY

U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND ARMY RESEARCH LABORATORY

SHIELD: Secure Human-Machine Interaction with Evidential Learning and Dynamic Trust for Drone Swarm Control

10 AUGUST 2026

Terrence J Moore
Research Mathematician

Distribution Statement A: Approved for public release: distribution is unlimited.






SHIELD: Secure Human-Machine Interaction with Evidential Learning and Dynamic Trust for Drone Swarm Control

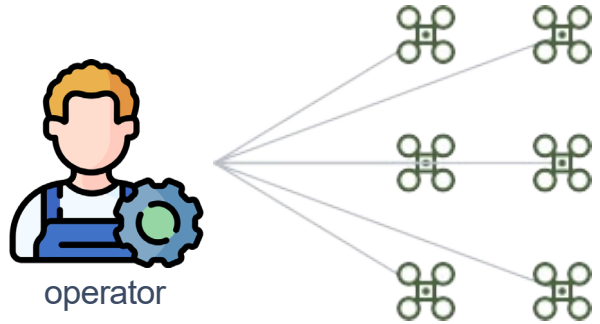
Dawood Wasif[1] **Terrence J. Moore**[2] Seunghyun Yoon[3] Hyuk Lim[3]
Dan Dongseong Kim[4] Frederica F. Nelson[2] Jin-Hee Cho[1]

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

USENIX VehicleSec 2026

The Dual Challenge of Drone Swarms

Humans do not scale



One operator cannot effectively watch more than **10 to 12** UAVs.

Autonomy is exposed



GPS spoof · jam · Byzantine · replay · malware

Removing the human opens a **cyber-physical attack surface**.

The gap: swarm autonomy must reduce operator workload *while* staying robust to compromised agents and uncertain environments. No prior system does both in one decentralized loop.

Why Existing Work Falls Short

Human-Swarm Teaming

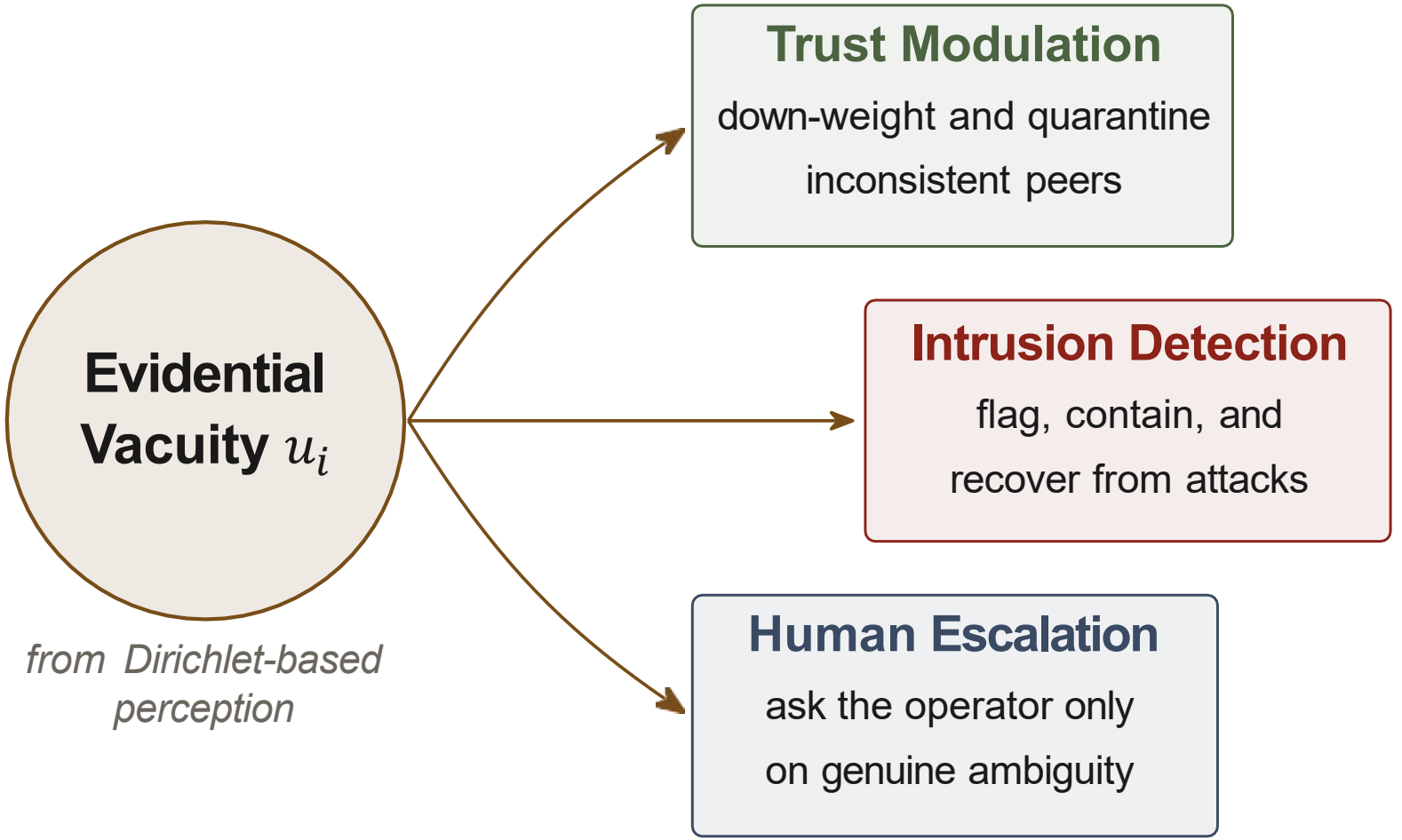
Improves interfaces and workload sharing but **assumes benign settings**. Every agent is trusted equally, with no way to flag or isolate a compromised UAV.

Multi-Agent Resilience

Isolates faulty agents, but **assumes centralized observers** and omits human oversight, targeting single-vehicle defenses, not the full human-swarm loop.

The missing piece: **no unified uncertainty signal** that jointly governs trust modulation, intrusion detection, and operator escalation inside one decentralized control loop.

Key Insight: One Signal, Three Jobs



A **single calibrated uncertainty quantity** drives all three functions at once, replacing three disconnected modules with one principled signal.

System and Threat Model

A cooperative swarm of N UAVs and **one operator** over a shared wireless network.

- Trusted, authenticated operator link.
- At most $f < \lfloor N/3 \rfloor$ compromised (Byzantine bound).
- Behavior-based trust, **no full PKI** on lightweight UAVs.

Six attack classes across five layers

A1 · GPS Spoofing
false position, perception

A2 · Comm. Jamming
dropped messages

A3 · Byzantine Faults
fabricated states

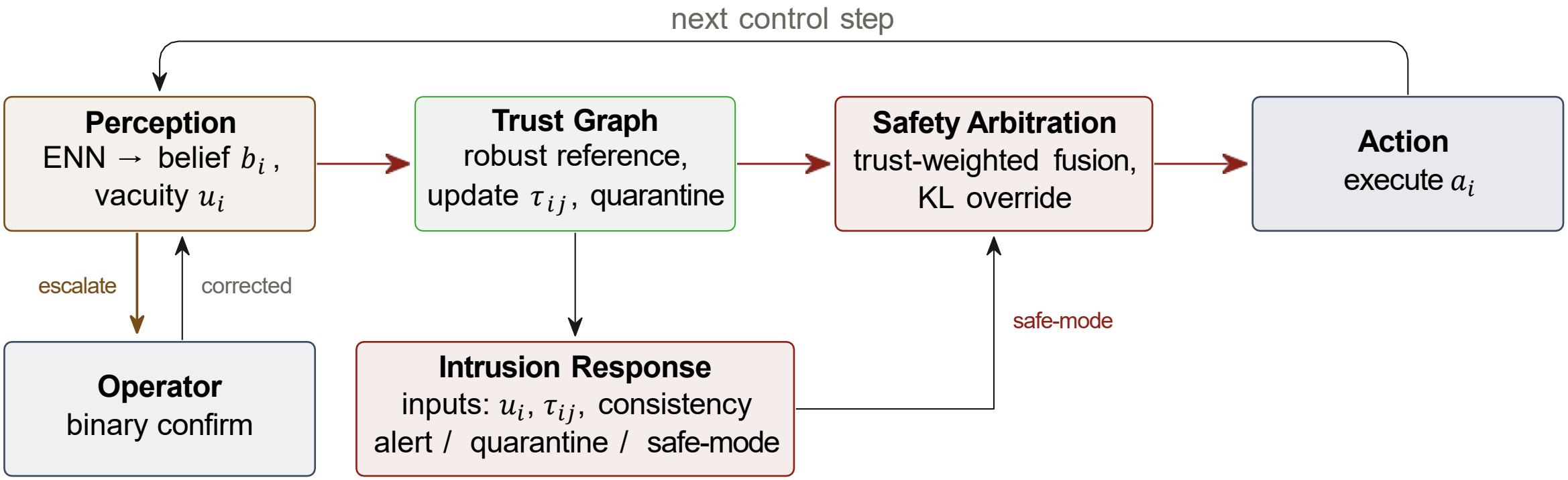
A4 · Replay
stale sensor data

A5 · Malware
altered policy, valid creds

A6 · Network Intrusion
packet inject / eavesdrop

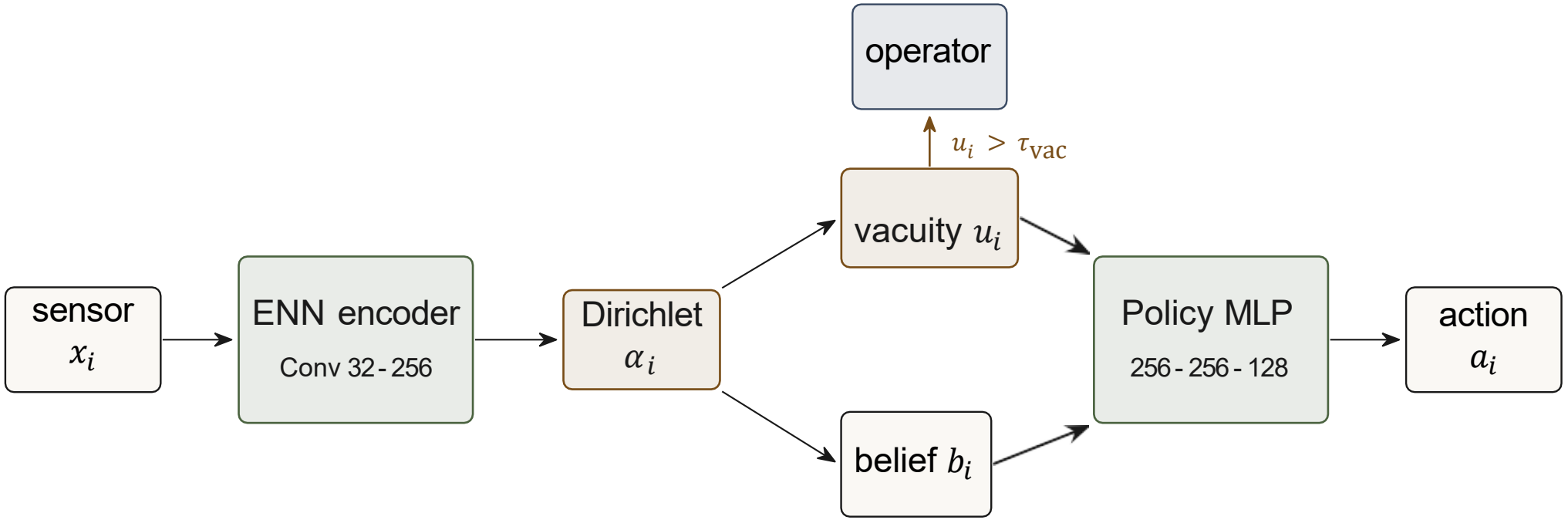
Plus three non-adversarial hazards (obstacles, congestion, emergencies) that must not be misread as attacks.

SHIELD: One Decentralized Control Loop



Routine decisions stay inside the swarm; sustained **vacuity** or **divergence** triggers containment or an operator query.

1. Evidential Perception and Escalation



$$\text{total evidence } S_i = \sum_{\ell=1}^{K} \alpha_{i,\ell} \qquad \text{belief } b_{i,k} = \frac{\alpha_{i,k}}{S_i} \qquad \text{vacuity } u_i = \frac{K}{S_i}$$

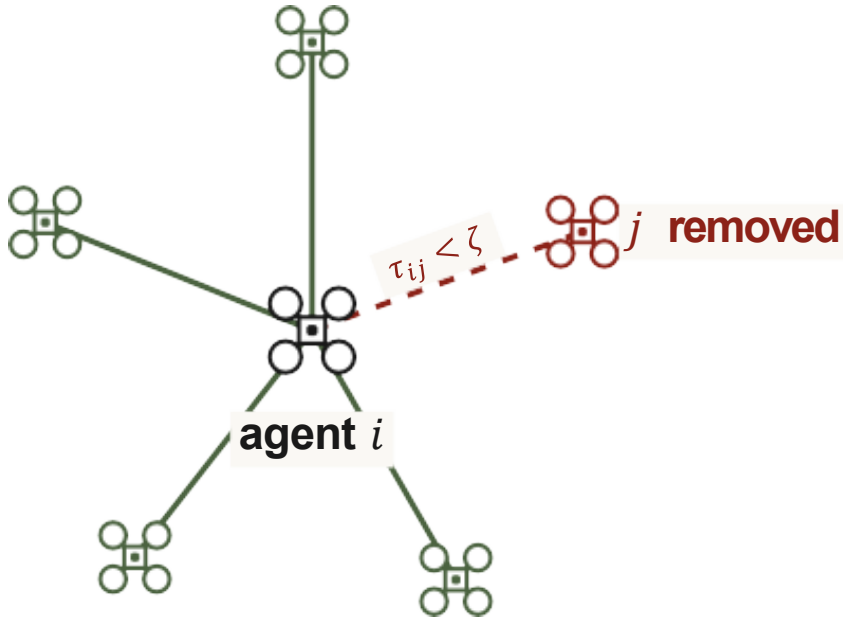
Scarce evidence $\Rightarrow$ small S_i $\Rightarrow$ high **vacuity**: an **out-of-distribution** flag that triggers a **binary** operator query.
A corrected belief resets u_i; no retraining at deployment.

2. Adaptive Trust and Structural Resilience

Each edge carries a trust weight τ_{ij} updated from **behavioral consistency** against a **robust reference** (coordinate-wise median of neighbors).

- The median tolerates up to f arbitrary values per coordinate, matching the fault bound.
- A spoofed position is **excluded before** it can bootstrap high trust.
- If $\tau_{ij} < \zeta$, the edge is removed and the agent is **quarantined**.

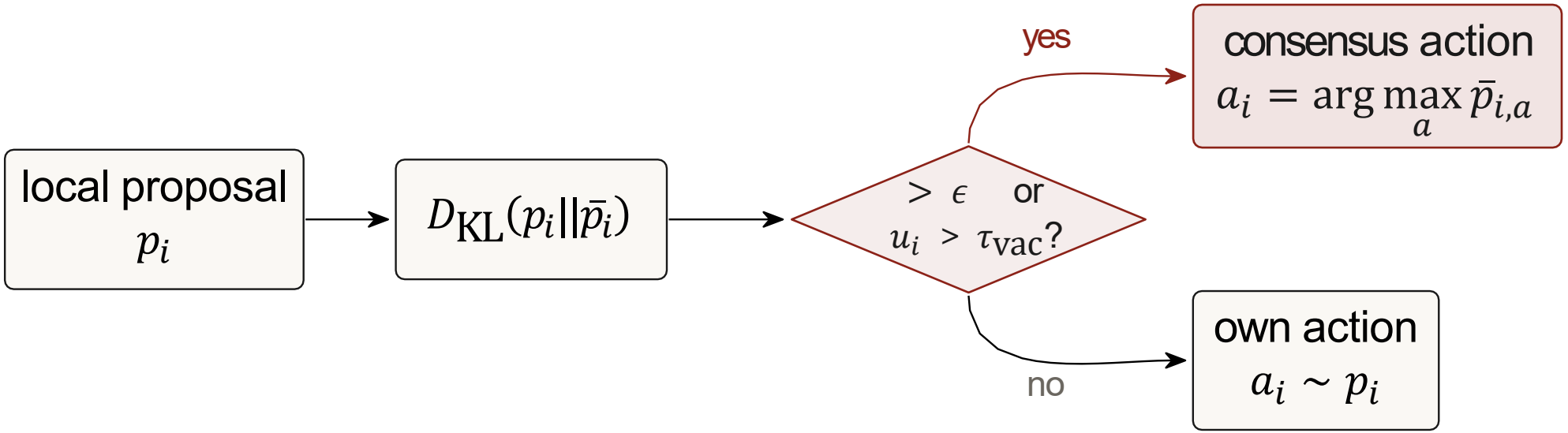
Compromised agent isolated



Coverage from a quarantined agent is reassigned to healthy neighbors in $O(\log N)$; recovered agents reintegrate gradually.

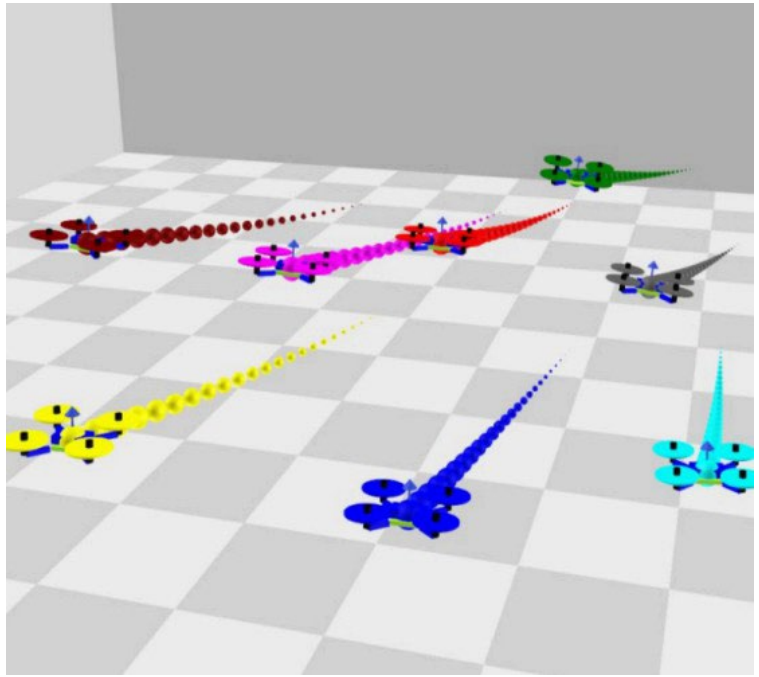
3. Consensus-Based Safety Arbitration

UAVs fuse neighbor proposals by **trust-weighted averaging**, then guard risky actions with a **divergence override**.



Two complementary signals. **Vacuity** captures self-uncertainty in perception; **trust** captures neighbor reliability. Either one alone triggers the fallback, preserving safety without giving up decentralized autonomy.

Evaluation Setup



QuadSwarm: waypoint surveillance, 8 to 32 UAVs

Task: waypoint surveillance in **QuadSwarm** at 50 Hz, across open and dense (20% obstacle) fields.

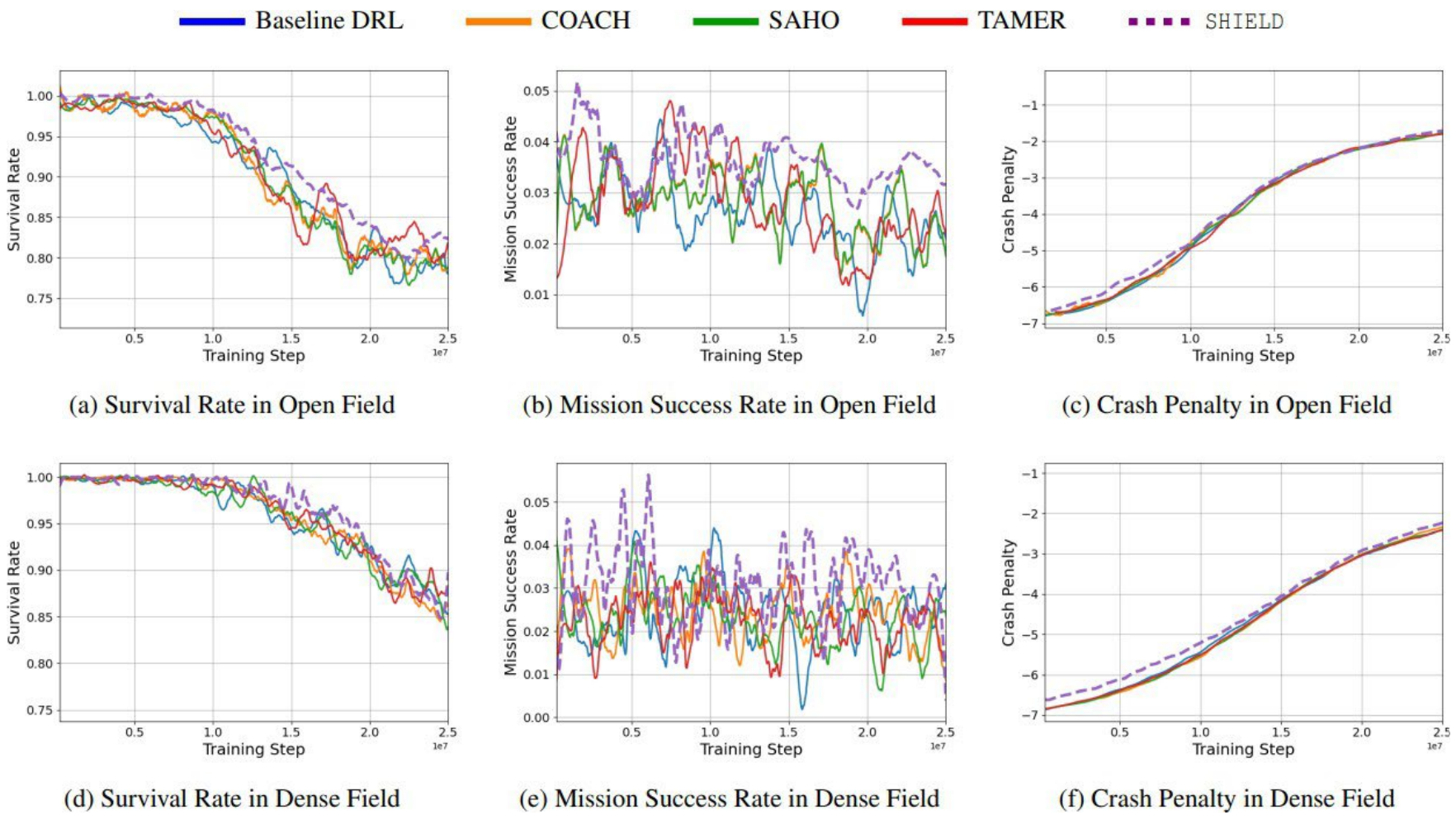
Baselines: Baseline DRL (APPO), COACH, SAHO, TAMER. Only SHIELD provides all four capabilities:

Method	Unc.	Trust	Intr.	Esc.
APPO	✗	✗	✗	✗
COACH	✗	✗	✗	△
SAHO	✗	✗	✗	△
TAMER	✗	✗	✗	△
SHIELD	✓	✓	✓	✓

✓ full △ partial ✗ none

Metrics: Mission Success (MSR), Survival (SR), Crash Penalty (CP), Detection (DR), False Positives (FPR), Recovery Time (RT), Human Interventions (HIC). Averaged over 30 trials.

Training Dynamics

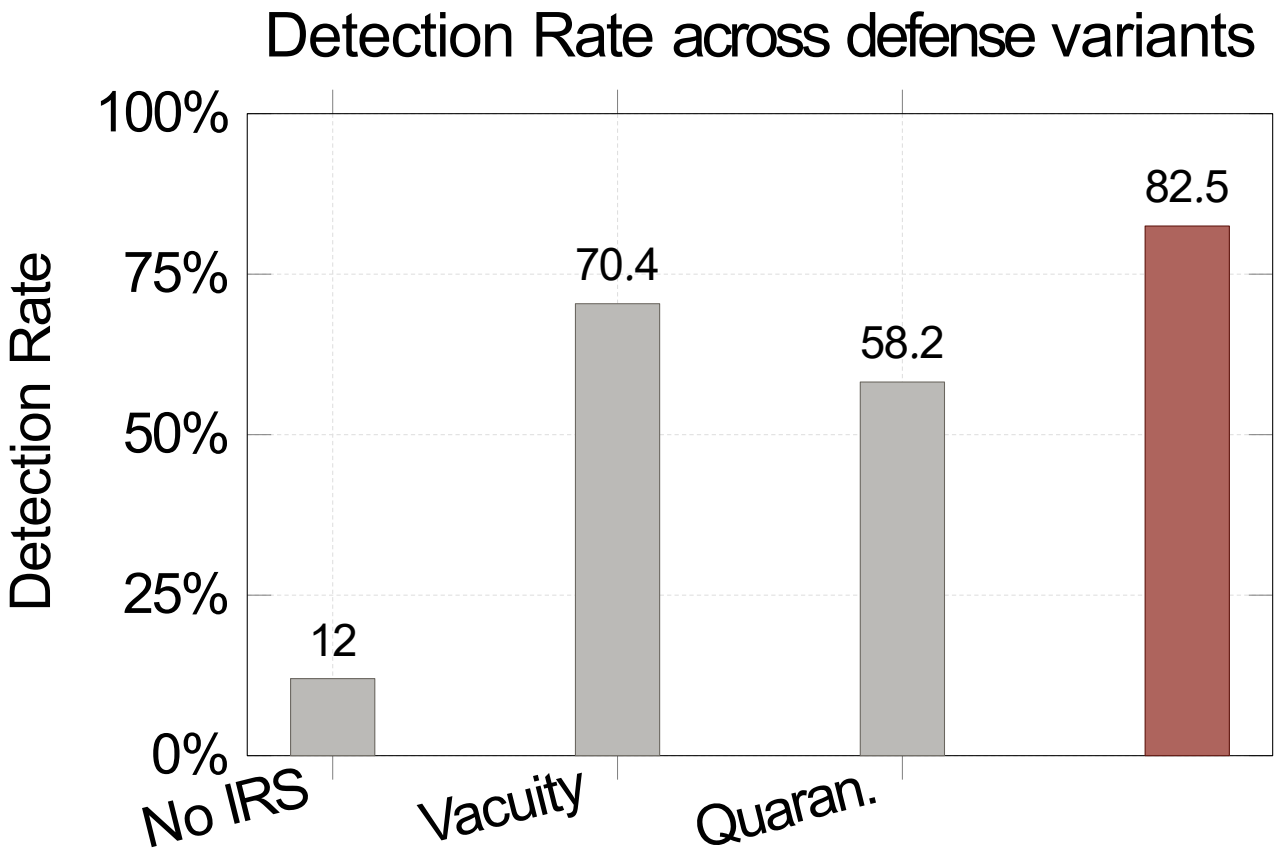


(a) Survival Rate in Open Field (b) Mission Success Rate in Open Field (c) Crash Penalty in Open Field

(d) Survival Rate in Dense Field (e) Mission Success Rate in Dense Field (f) Crash Penalty in Dense Field

SHIELD (dashed) leads on mission success and survival, with the largest gap early in training.

Results: Detection and Operator Workload



Variant	DR (%)	FPR (%)	RT (s)	HIC (/ep)
No IRS	12.0	n/a	n/a	28.7
Vacuity	70.4	18.3	20.5	20.8
Quaran.	58.2	22.4	15.4	19.8
SHIELD+IRS	**82.5**	**14.6**	**9.3**	**13.5**

Full SHIELD+IRS gives the highest detection and the lowest false positives, recovery time, and operator load, with the lowest variance on every metric.

Results: Ablation and Scalability

Every module contributes

Variant	CP	MSR	SR
Full SHIELD	**-3.85**	**0.040**	**0.120**
w/o Vacuity	-4.35	0.032	0.095
w/o Trust	-4.62	0.025	0.081
w/o Safety	-5.19	0.018	0.077

Monotonic decline confirms complementary, non-redundant modules.

Stable from 8 to 32 UAVs

Swarm	CP	MSR	SR
8	-3.85	0.040	0.120
16	-3.99	0.026	0.115
32	-4.02	0.021	0.114

Survival stays flat; inference is **7.8 ms** *at* N=64, *inside the 20 ms control budget.*

Thresholds and attack parameters swept ±20 to 50%: DR stays in **[78, 86]%**, confirming results are not artifacts of tuning.

Limitations and Future Work

- Evaluation is **simulation-based**. Real platforms add unmodeled aerodynamics, sensor drift, and latency that affect vacuity calibration and trust dynamics.
- Current attacks are **parameterized and non-adaptive**. Adaptive, RL-based adversaries that game the quarantine cycle are deferred to future work.
- Next steps: **physical-platform validation**, adaptive thresholding, richer low-cognitive-load feedback beyond binary confirmation, and communication-efficient trust updates at larger scale.

Conclusion

- One signal, **evidential vacuity**, unifies trust modulation, intrusion detection, and human escalation in a single decentralized loop.
- Across six attack classes: **82.5% detection**, 14.6% false positives, 9.3 s recovery, and **53% fewer** operator confirmations.
- Up to **60% relative improvement** in mission success, with graceful degradation to 32 UAVs.

Thank you!

Questions?

Presented by: Terrence J. Moore terrence.j.moore.civ@army.mil
U.S. DEVCOM Army Research Laboratory, Adelphi, MD

Joint work with Dawood Wasif, Seunghyun Yoon, Hyuk Lim, Dan Dongseong Kim, Frederica F. Nelson (ARL), and Jin-Hee Cho (Virginia Tech)

Artifact (Zenodo): 10.5281/zenodo.20208768

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