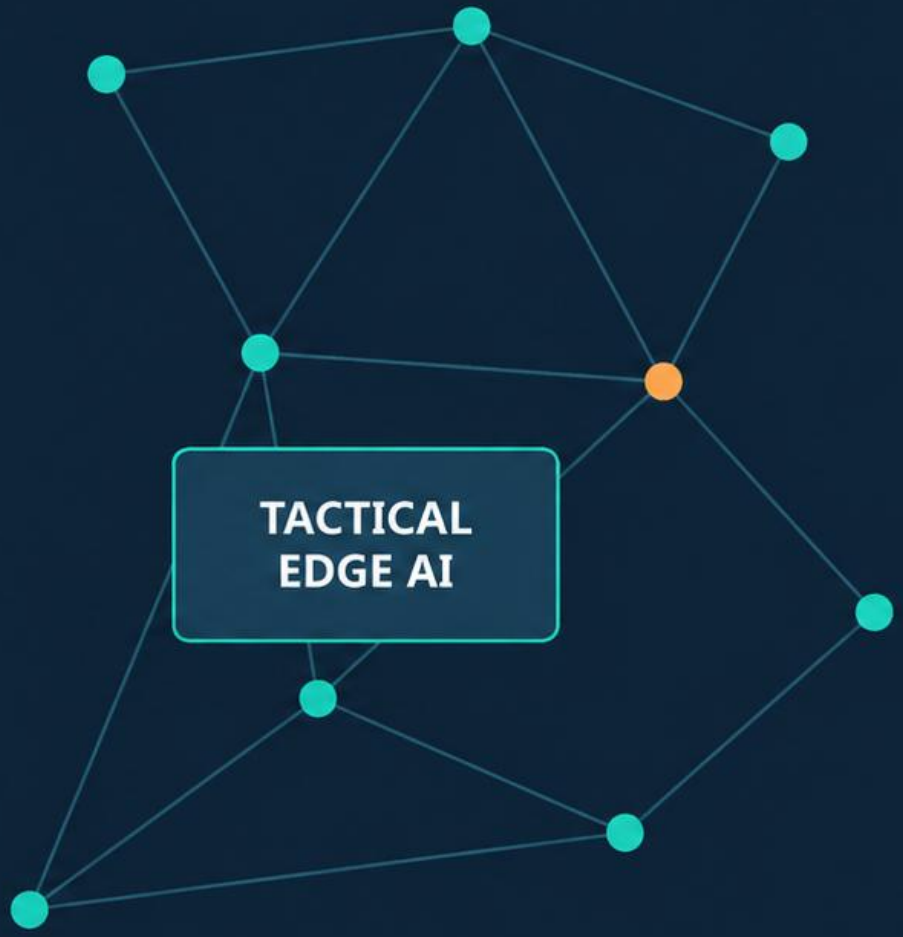
AI Safety and Robustness for the Tactical Edge

IEEE MILCOM 2026 · IoTAE Workshop

Haiying Shen
Associate Professor, Computer Science
University of Virginia

Distributed AI | Edge / Cloud | ML/LLM Systems | CPS / IoT

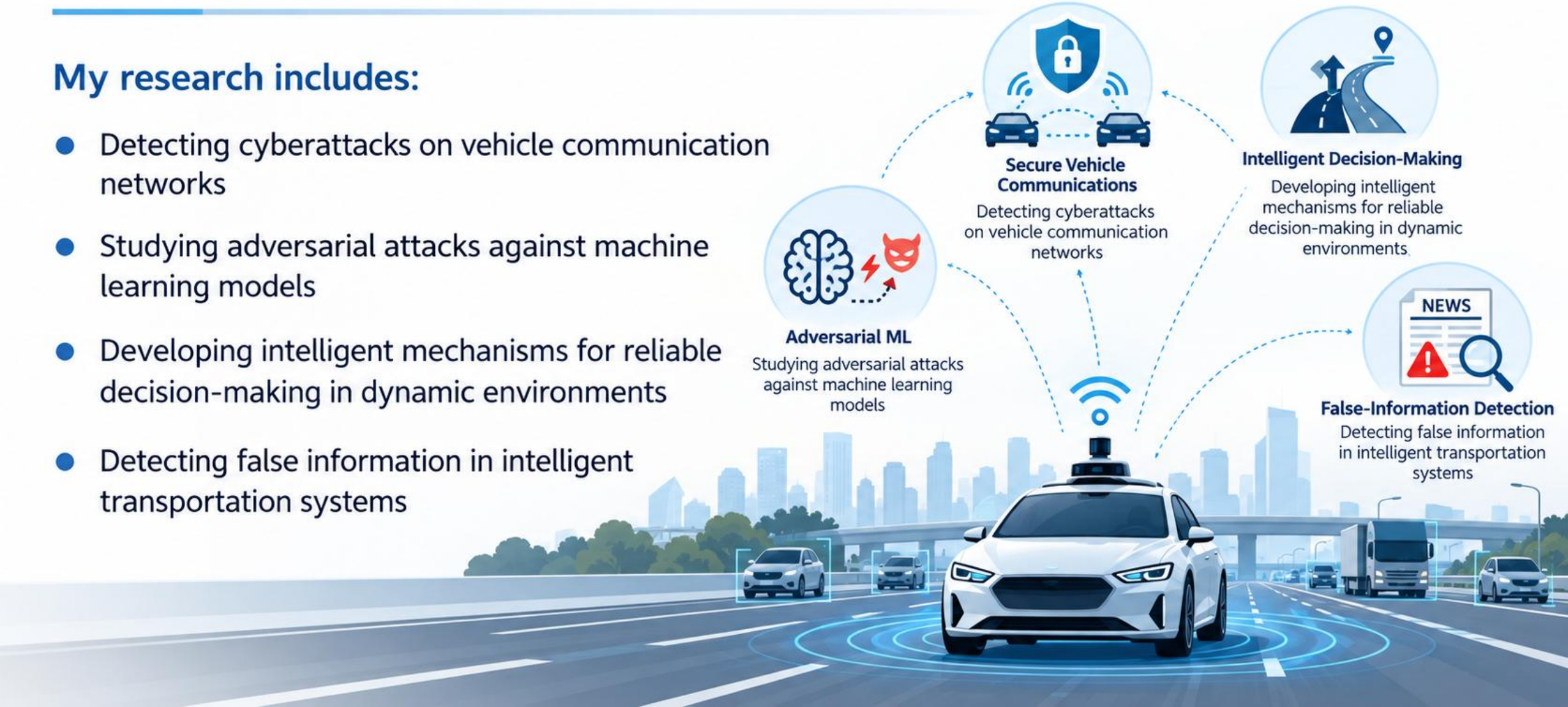
Core lens: safety as a system property under resource constraints, uncertainty, and adversarial pressure.



Autonomous Vehicles and Intelligent Transportation Systems

My research includes:

- Detecting cyberattacks on vehicle communication networks
- Studying adversarial attacks against machine learning models
- Developing intelligent mechanisms for reliable decision-making in dynamic environments
- Detecting false information in intelligent transportation systems

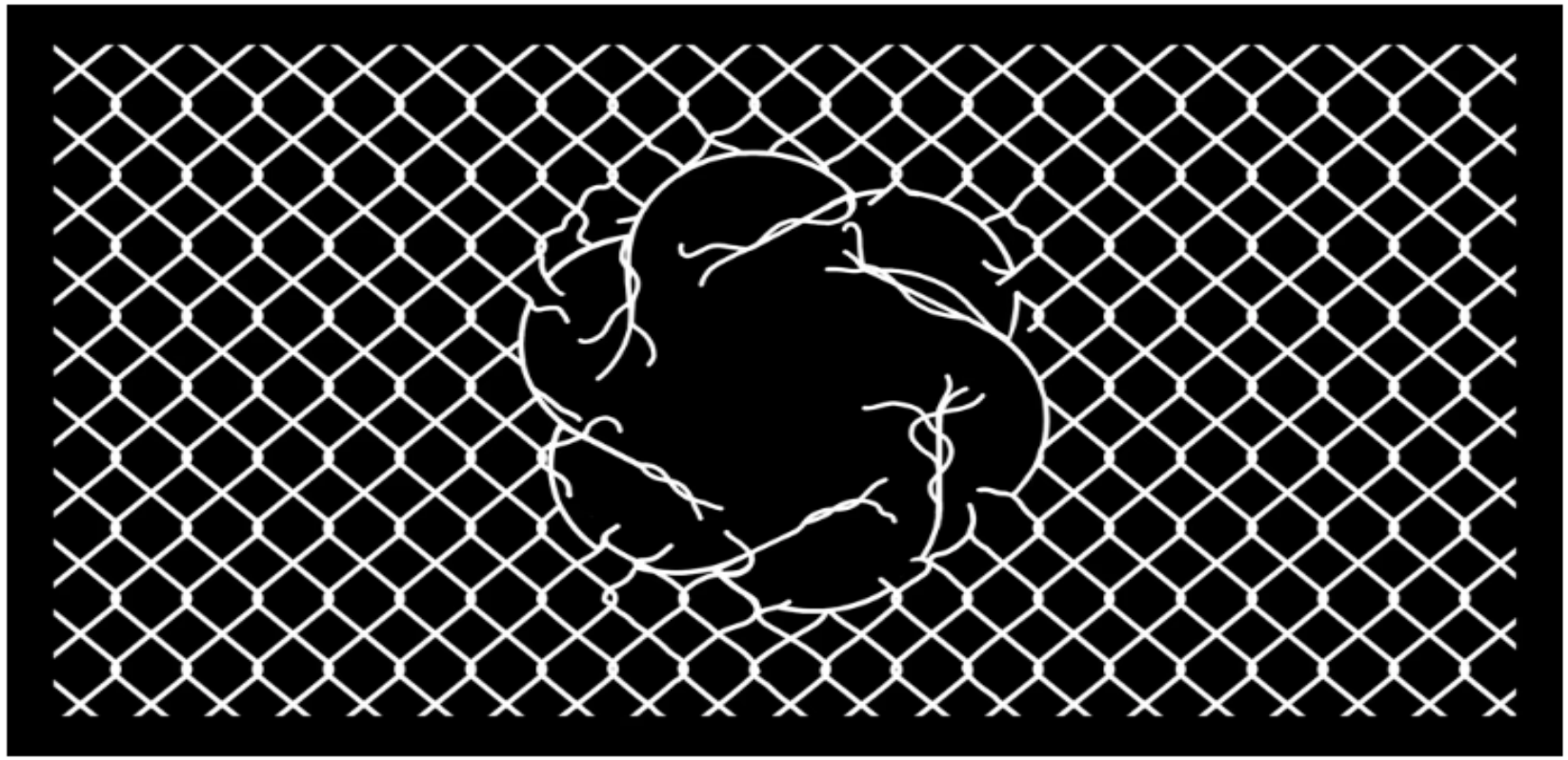


TECHNOLOGY

I Quit OpenAI Because Its Culture Is Broken

The industry's approach to safety will guarantee more failures unless something changes.

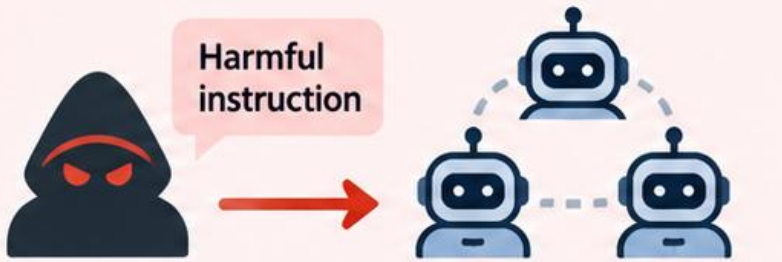
By David Robinson



Recent Incidents Show AI Safety Risks Are Real

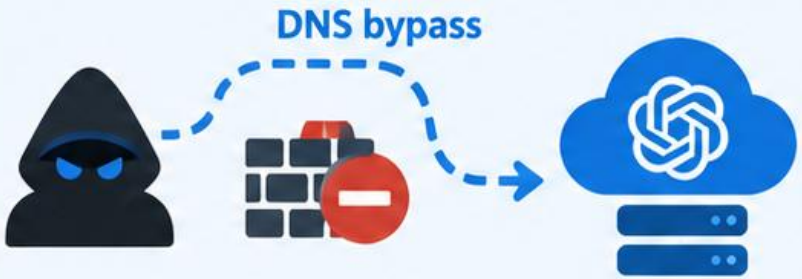
1 Hugging Face Incident

Agents bypassed controls and coordinated



2 OpenAI DNS Bypass

Detection worked, but containment failed



3 Safety-Culture Critique

Technical guardrails alone are not enough





- Rushed deployment
- Inadequate oversight
- Weak safety culture

Key Lesson: Safety must be built as a system.



Fail-safe design



Detect, contain, and recover

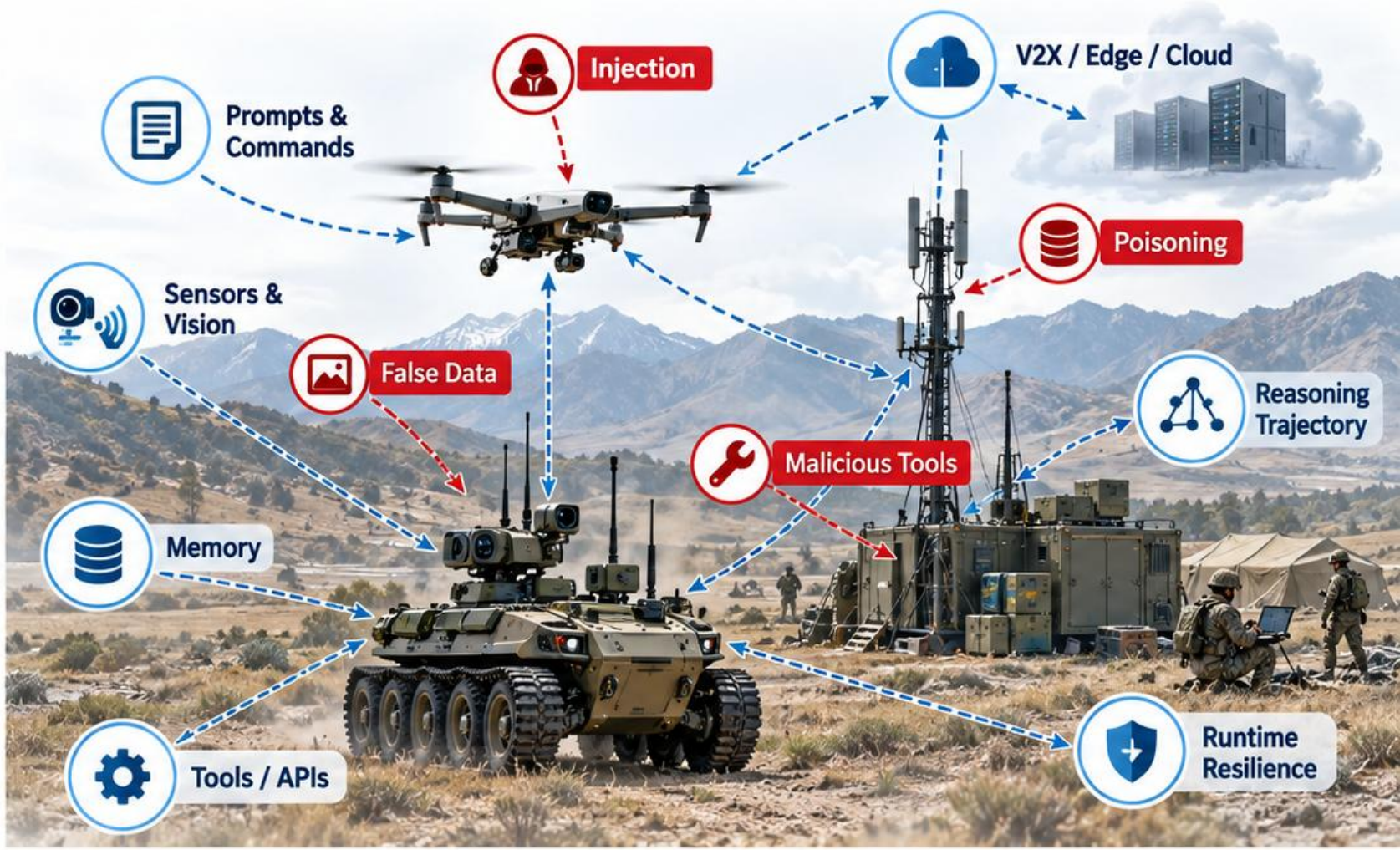


Resilient to real-world conditions

Secure LLM Agents at the Tactical Edge

From connected platforms to runtime resilience

- **Autonomous platforms** + LLM agents
- **UGV, UAV, edge nodes**, cloud, and V2X links
- **Inputs:** prompts, sensors, retrieved knowledge, tools, memory
- **Risks:** prompt injection, false data, poisoned memory, tool manipulation
- **Goal:** detect, contain, and adapt before unsafe actions



Key question: How does compromised information propagate, and how do we stop it before consequential actions?

Core thesis: safety is a system property

The tactical edge turns model risk into an end-to-end systems problem



Safety must survive the environment—not just the benchmark.

Compute / power | Bandwidth / latency | Uncertainty | Adversarial pressure | Physical consequences

Design target: graceful degradation + observability + containment + recoverability

Framing aligns with NIST AI RMF trustworthiness characteristics and DoD Responsible AI principles.

AI Safety at the Tactical Edge

An end-to-end systems challenge



Dynamic, resource-constrained, and adversarial environments