TRUSTWORTHY AI FOR SAFE CYBER-PHYSICAL SYSTEMS

Mansur M. Arief

Postdoc in AI Safety and Sustainability at Stanford Intelligent Systems Lab (SISL) and Mineral-X

PhD in Mechanical Engineering at Safe AI Lab, Carnegie Mellon University (CMU)

INTRO

- AI adoptions are increasingly more prevalent in CPS.
- Their performance is extremely hard to guarantee, or to explain.
- Adoption is conditional on safety guarantee and public acceptance.

FRAMEWORK

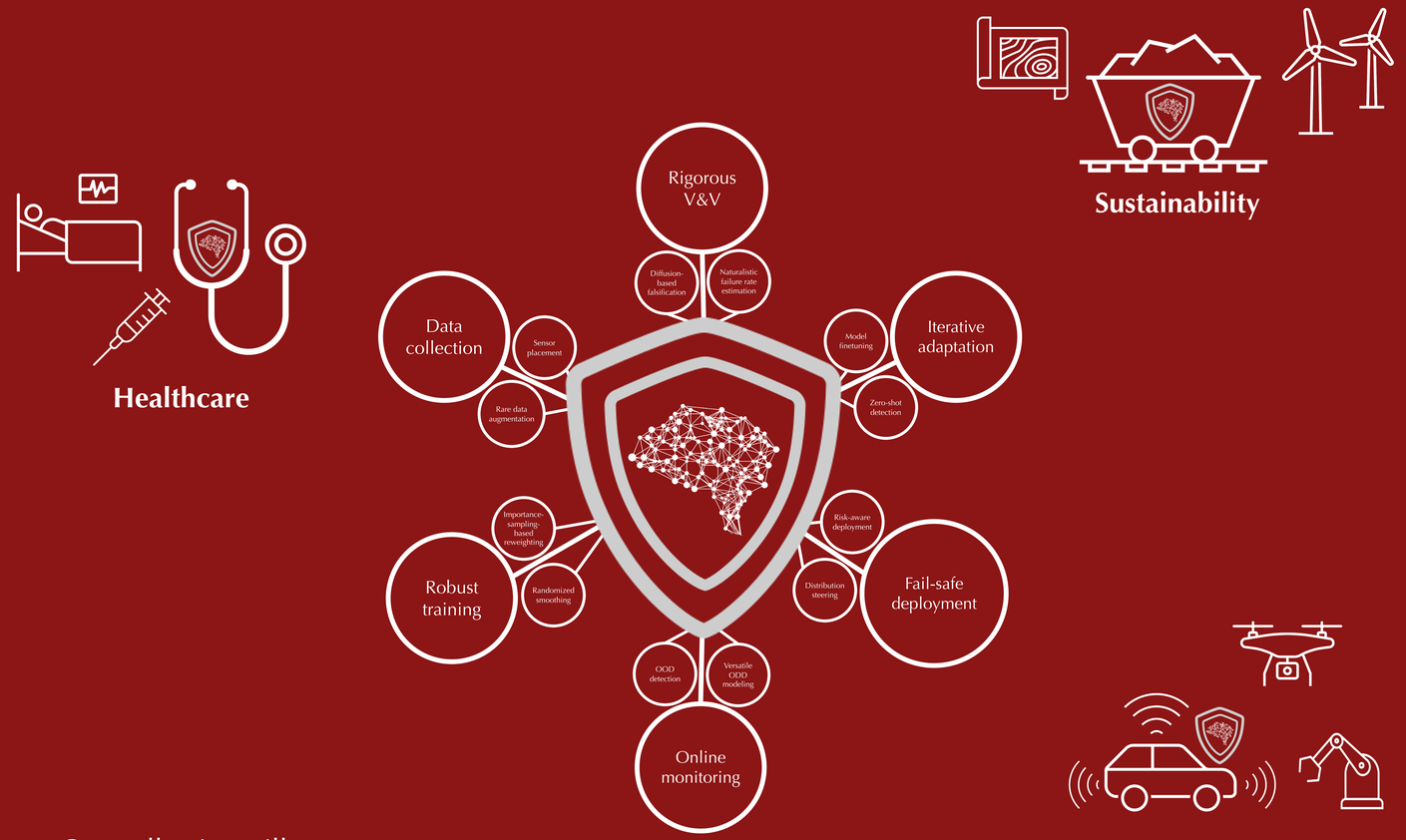
1. **Trustworthy-centric development**
   - Optimized sensor placement [1]
   - Failure-augmented training [2]
   - Novel scenario monitoring [3]
   - Multi-platform testing synthesis [4]
   - Risk-aware deployment [5]
2. **Iterative online adaptation**
   - Point-cloud scene annotation [6]
   - Rare-data-augmented retraining and adaptation [2, 3]
3. **Scalable AI V&V approach**
   - Critical scenario generation [8]
   - Diffusion failure sampling [9]
   - Deep and latent-space importance sampling [10-13]

Scan for link to the **references** and **download full poster**

Stanford University | SISL Stanford Intelligent Systems Laboratory

CPS Rising Stars 2024 Workshop

Trustworthy AI is crucial for safe CPS across domains

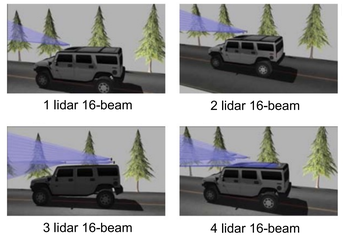


Our collective will, our responsibility, is to create trustworthy AI. Now A.I. is doing really good work, making scientific discoveries, finding new materials, and medical. The real issue now, is how to develop and deploy the technology thoughtfully, whether it is in the classroom or industry. *Fei Fei Li*

Our trust in technology relies on understanding how it works. It's important to understand why AI makes the decisions it does. … develop tools to make it more explainable, fair, robust, private, and transparent. *IBM Research*

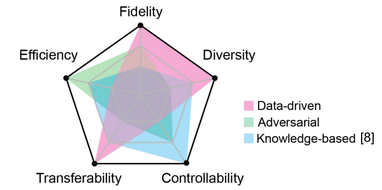
CURRENT RESULTS

- **[Design]** Sensor placement with good redundancy reduces stringent downstream requirements.



[1]

- **[Model & Prototype]** The "long tail" of scenario distribution poses huge challenge, but recent AI advances alleviates some practical issues.



- **[Evaluation & Monitoring]** Failures rarity + catastrophic consequences require V&V rigor to ensure safety.

[12]

FUTURE RESEARCH

1. **AI-based closed-loop design + V&V:** feedback mechanisms for rapid explainability and design iterations
2. **Ultra-safe CPS for safety-critical systems**: mobility, healthcare, and sustainability
3. **Trustworthy CPS**: trustworthy AI + systems design + safety engineering

LET'S STAY CONNECTED!

Website: www.mansurarief.github.io

Email: mansur.arief@stanford.edu