Assured and Intelligent Cyber-Physical Systems (AICPS)

Dr. Fanxin Kong
Department of Computer Science and Engineering
University of Notre Dame
fkong@nd.edu

Intelligent Cyber-Physical Systems

- Traditional CPS tightly couple computation, networking and physical processes
- Intelligent CPS tightly couple **Super Intelligence**, networking and physical processes



Humanoid robots

2025: a Unitree H1 lunged toward a festival crowd in Tianjin; another flailed violently during a factory test, nearly hitting engineers.
2026: a dancing humanoid struck a child in the face at a Shaanxi mall.



Driver assistance (Tesla)

Oct 2025: NHTSA opened a probe into FSD on ~2.9 M Teslas after reports of running red lights and entering opposing lanes (14 crashes, 23 injuries).
Aug 2025: a federal jury held Tesla partly liable ($243 M) for a fatal Autopilot crash.



Robotaxis (Waymo)

Dec 2025: Waymo recalled software on 3,067 robotaxis after they drove past stopped school buses with red lights flashing; Austin schools documented 19 such cases.



Industrial robots

2023, South Korea: a box-handling robot arm crushed a worker it apparently took for a box.
2025: a Tesla technician sued for $51 M after a FANUC robot arm knocked him unconscious.

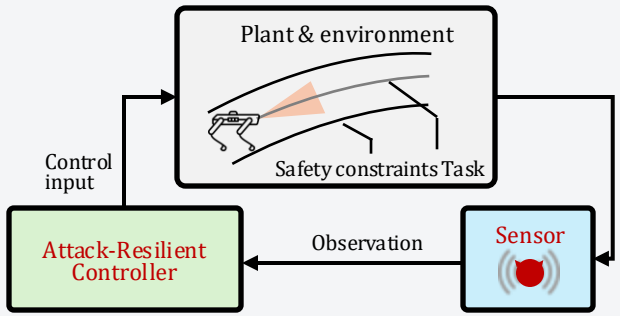
Assured and Intelligent CPS

- Assurance
  - Safety: identifying and reducing conditions that can lead to physical harm or environmental damage
  - Security: protection from intentional attacks

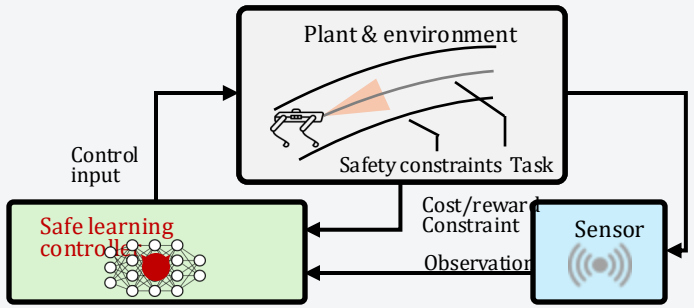
- Intelligence
  - Autonomy: perceiving environments, making decisions, and executing actions independently with little human intervention
  - Generalizability: ability to apply what is learned during training to new and unseen data, tasks, or environments
  - Adaptability: evolving with new data, tasks, and/or environments
  - Low SWAP: small size, light weight, and low power (compared to huge foundation models)

Journey to AICPS

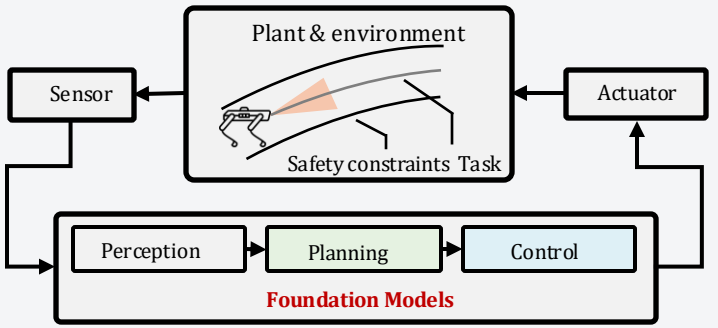
T1: Real-Time Attack-Resilient CPS



T2: Secure and Safe Reinforcement Learning

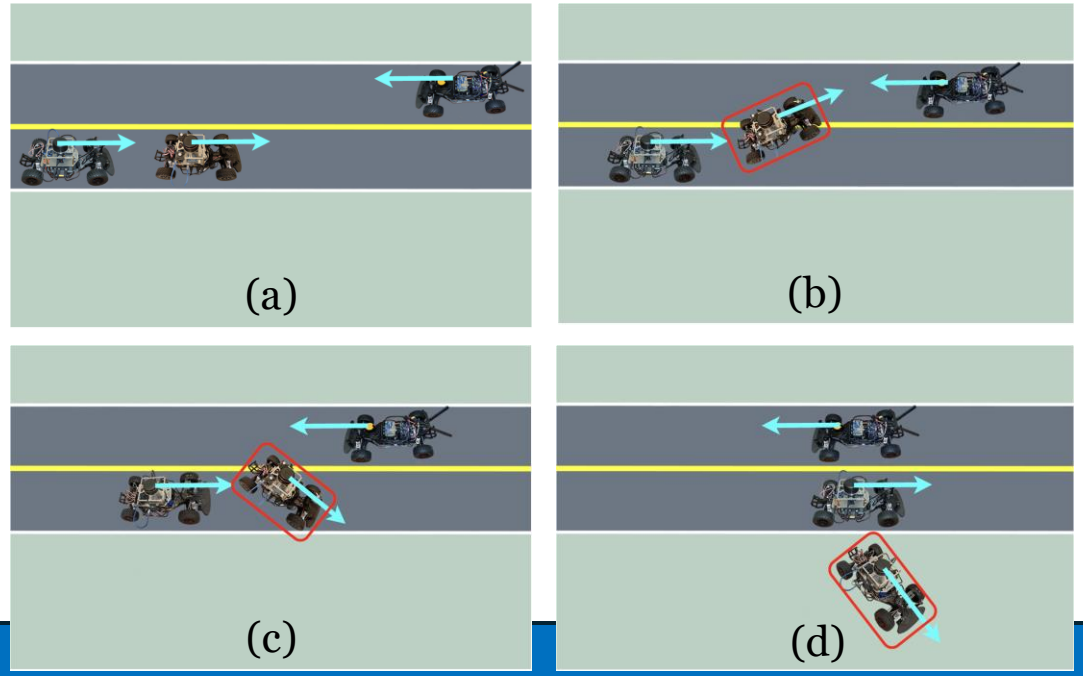


T3: Foundation Model-Enabled CPS



T1: Real-Time Attack-Resilient CPS (NSF CPS, CNS–2143256)

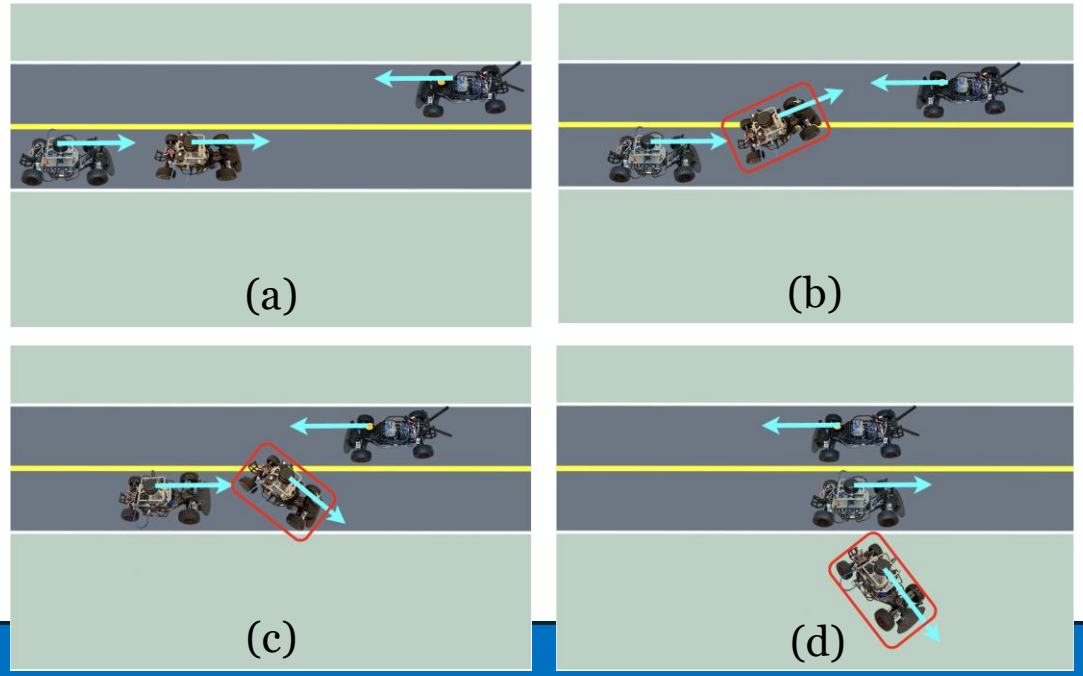
- Research problem: how to respond to attacks in CPS?
  - Importance: 1) need to stop drifting
    2) reduce caused negative impact
  - Challenges: 1) Timeliness: should detect and recover before irreparable consequences occur
    2) Safety: avoid unsafe states for response





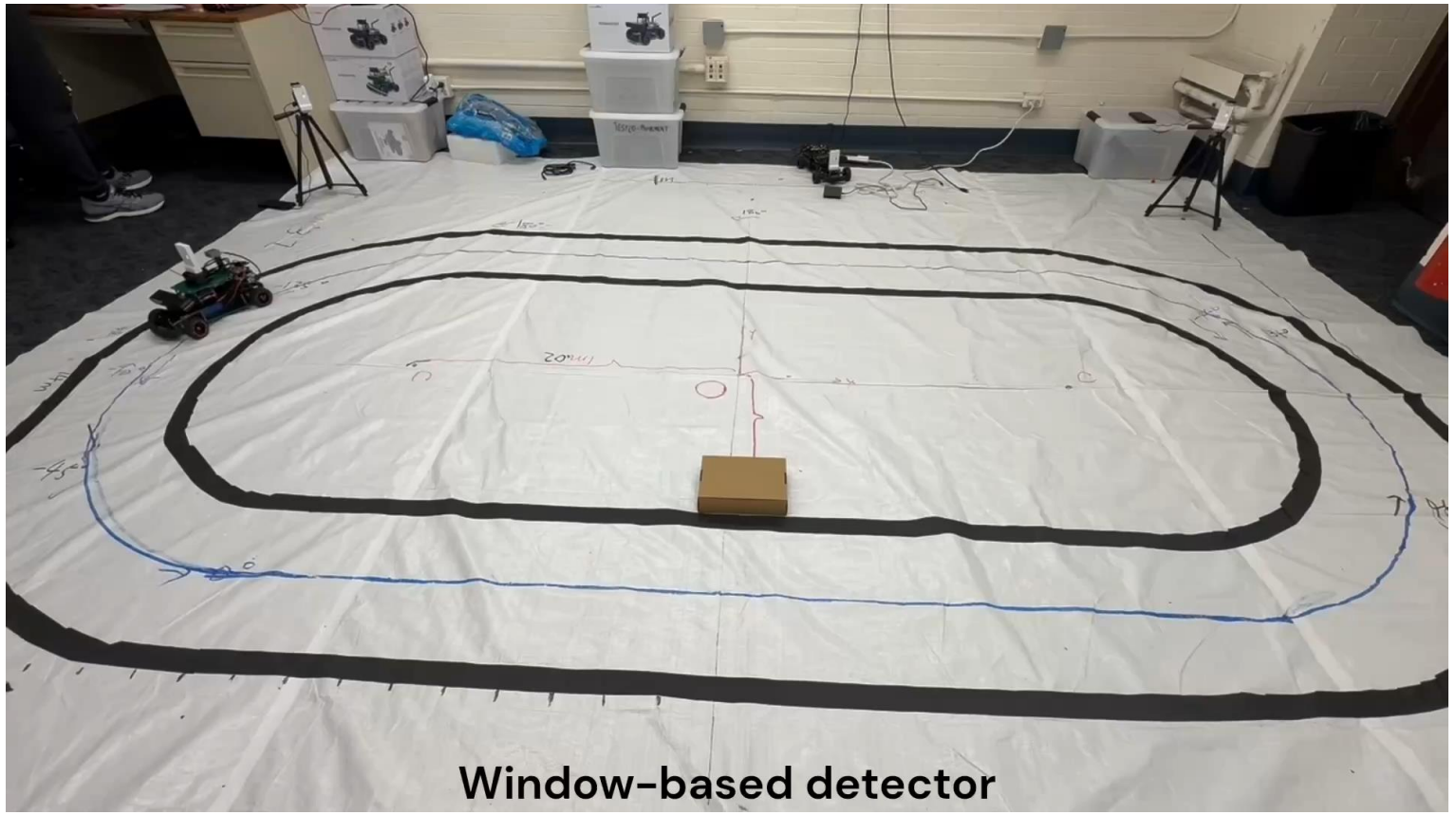
T1: Real-Time Attack-Resilient CPS (NSF CPS, CNS–2143256)

- Research problem: how to respond to attacks in CPS?
  - Importance: 1) need to stop drifting
    2) reduce caused negative impact
  - Challenges: 1) Timeliness: should detect and recover before irreparable consequences occur
    2) Safety: avoid unsafe states for response



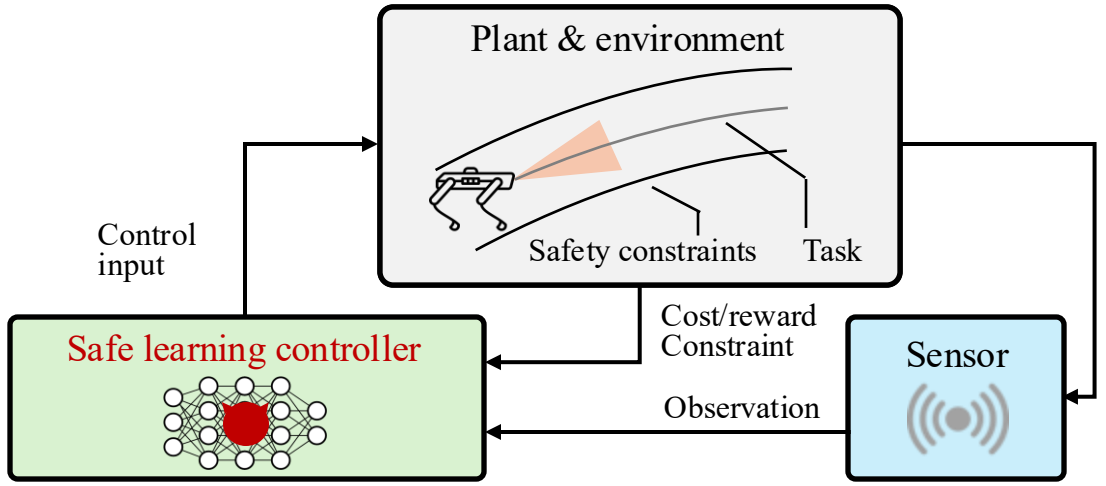


Demo: recovery-guaranteed recovery (RTAS'25)



T2: Secure and Safe Reinforcement Learning (NSF CAREER, CNS–2442914)

- Safe Reinforcement Learning (RL)
  - Maximizes the task reward while satisfying safety constraints
  - Generates control actions given observation
- Vulnerable neural networks
  - Small perturbation can cause drastic change in output
  - Compromise inference and training time
- Research problem



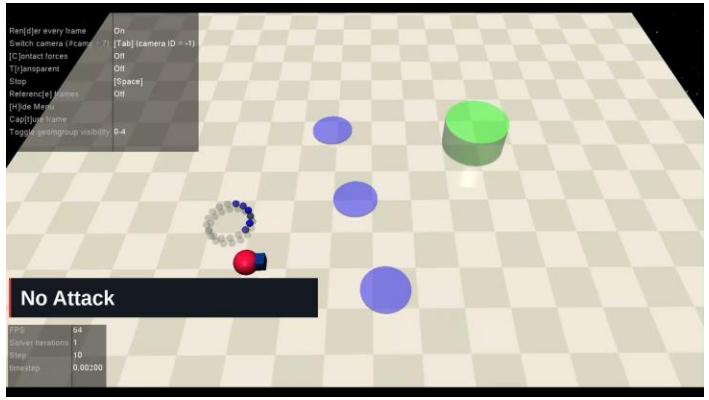
How to secure safe RL control polices?

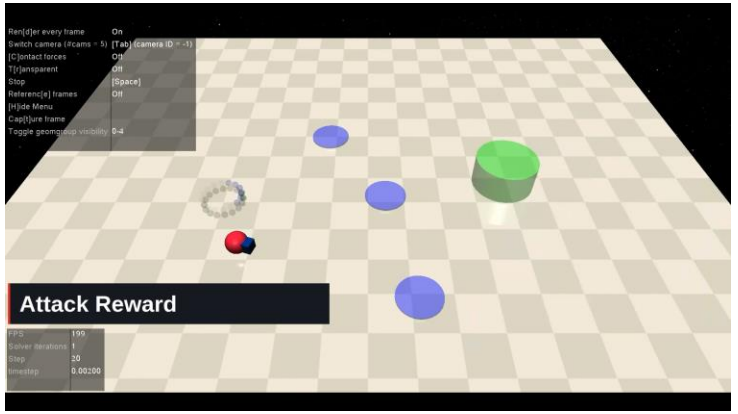
T2: Secure and Safe Reinforcement Learning (NSF CAREER, CNS–2442914)

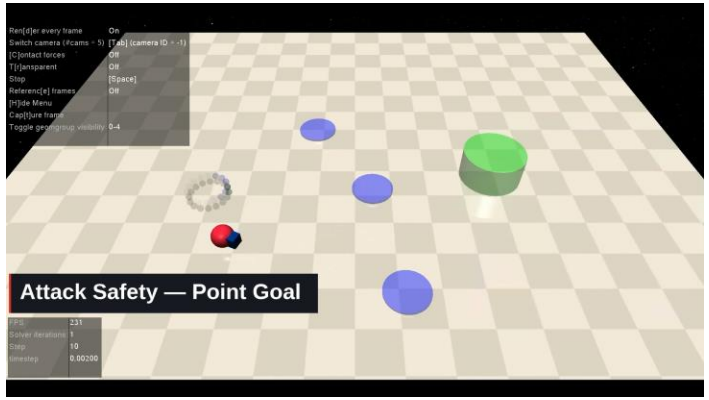
- Why to study safety-oriented vulnerability?
  - Reward-decreased attack ≠ safety-violated attack
- What needed?
  - New definitions of vulnerabilities and attacks that involve safety
  - New vulnerability exploration, mitigation, and defense methods

Finding 1: Attacks decreasing reward may not violate safety.

Finding 2: Attacks violating safety may not change or even increase reward.





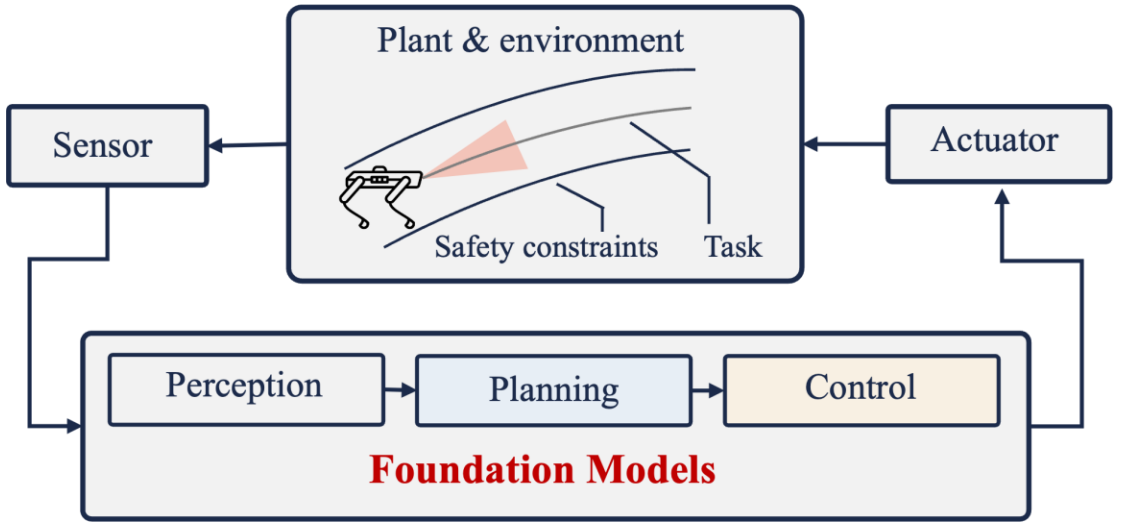


(Demo from ICCPS'24)

T3: Foundation Model-Enabled CPS

- Foundation Models (FMs) pretrained on massive data
  - High generalization capability and adaptability
  - Large Language Models (LLMs), Vison Language Models (VLAs)

- FMs transforming landscape of CPS autonomy
  - Application in perception, planning, and control
  - LLMs for task planning
  - VLAs for end-to-end control



T3: Foundation Model-Enabled CPS

- Foundation Models (FMs) pretrained on massive data
  - High generalization capability and adaptability
  - Large Language Models (LLMs), Vison Language Models (VLAs)

- FMs transforming landscape of CPS autonomy
  - Application in perception, planning, and control
  - LLMs for task planning
  - VLAs for end-to-end control

- Problem?



T3: Foundation Model-Enabled CPS

- Foundation Models (FMs) pretrained on massive data
  - High generalization capability and adaptability
  - Large Language Models (LLMs), Vison Language Models (VLAs)

- FMs transforming landscape of CPS autonomy
  - Application in perception, planning, and control
  - LLMs for task planning
  - VLAs for end-to-end control

- Research Problem

How to evaluate, enhance, and assure safety in FM-enabled CPS?

Preliminary results: SafeVLA-Bench (arXiv)

- SafeVLA-Bench: A Benchmark for the Success–Safety Gap in Vision-Language-Action Models



π0.5 · LIBERO-10
SUCCESS ✓ — knocks over the milk carton ✗



π-RL-130 · LIBERO-Object
SUCCESS ✓ — knocks over a bottle ✗



π-RL-130 · LIBERO-10
SUCCESS ✓ — tilts the moka pot in transit ✗

- Leaderboard: https://safevla.org/

Preliminary results: SafeVLA-Bench (arXiv)

LIBERO · mean over four suites

#	Model	Train	Safety ↑	SR ↑	SBU ↓	VSI ↓
1	VLA-Adapter	SFT	82.0%	97.6% #3 by SR	17.0%	0.104
2	InternVLA-M1	SFT	81.5%	96.4% #10 by SR	16.8%	0.118
3	VLANeXt	SFT	81.5%	97.6% #4 by SR	17.1%	0.110
4	X-VLA	SFT	81.0%	97.2% #5 by SR	17.1%	0.113
5	MolmoAct2	SFT	80.9%	97.9% #1 by SR	17.5%	0.107
6	SimVLA	SFT	80.2%	96.9% #8 by SR	18.2%	0.123
7	OpenVLA-OFT-7B	SFT	80.1%	97.8% #2 by SR	18.1%	0.122
8	UniVLA	SFT	79.5%	94.8% #12 by SR	17.1%	0.127
9	GR00T-N1.6	SFT	78.4%	97.1% #6 by SR	20.0%	0.125

LIBERO: 18 policies, mean over 4 suites (top 9 shown)

RoboCasa-365 · Atomic-Seen

#	Model	Safety ↑	SR ↑	SBU ↓	VSI ↓
1	RLDX-1-FT-RC365	53.1%	58.4% #4 by SR	24.0%	0.222
2	GigaWorld-Policy-0.1	52.0%	43.1% #7 by SR	20.0%	0.255
3	$\pi_{0.5}$	51.2%	42.3% #8 by SR	19.8%	0.195
4	PRTS	51.2%	64.1% #3 by SR	30.4%	0.241
5	GR00T-N1.6	51.0%	53.6% #5 by SR	24.6%	0.253
6	Xiaomi-Robotics-1	50.1%	81.9% #1 by SR	37.7%	0.233
7	WorldDreamer	48.9%	68.3% #2 by SR	31.1%	0.276
8	π_0	44.1%	32.4% #9 by SR	12.3%	0.318
9	GR00T-N1.5	40.7%	47.2% #6 by SR	26.3%	0.268

RoboCasa-365: 9 policies, 18 task families

All results are public at safevla.org: Safety, SR, SBU and VSI per policy, sortable by any metric.

Assured and Intelligent CPS

- Assurance
  - Safety: identifying and reducing conditions that can lead to physical harm or environmental damage
  - Security: protection from intentional attacks

- Intelligence
  - Autonomy: perceiving environments, making decisions, and executing actions independently with little human intervention
  - Generalizability: ability to apply what is learned during training to new and unseen data, tasks, or environments
  - Adaptability: evolving with new data, tasks, and/or environments
  - Low SWAP: small size, light weight, and low power (compared to huge foundation models)

Other preliminary results

- Adaptivity of AICPS
  - Safe-RULE: Safe Reinforcement UnLEarning (arXiv)
- Low SWaP FM for CPS
  - SLoSa: Small, Low-Power, and Safe Foundation Models for Robotic Task Planning (arXiv)

Acknowledgements

Collaborators


Prof. Insup Lee
Univ of Penn


Prof. Oleg Sokolsky
Univ of Penn


Prof. Danny Chen
Univ of Notre Dame


Prof. Sharon Hu
Univ of Notre Dame


Prof. Alvaro A. Cardenas
UC Santa Cruz


Prof. Qinru Qiu
Syracuse Univ


Prof. Russell Joseph
Northwestern Univ


Prof. Xin Chen
Univ of New Mexico


Dr Steve Drager
AFRL

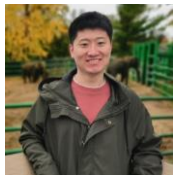

Dr Matt Anderson
AFRL


Prof. Vir V. Phoha
Syracuse Univ


Prof. M. Cenk Gursoy
Syracuse Univ

Former PhD Students

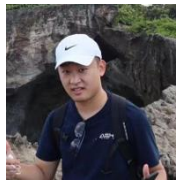

Prof. Francis Akowuah
South Dakota Mines


Dr. Lin Zhang
Eli Lily


Prof. Mengyu Liu
Washington State Univ


Dr. Zifan Wang
Meta

Current PhD Students


Shixiong Jiang
Univ of Notre Dame


Weizhe Xu
Univ of Notre Dame


Jialiang Fan
Univ of Notre Dame


Taozheng Zhu
Univ of Notre Dame


Zijun Wang
Univ of Notre Dame


Sicheng Chen
Univ of Notre Dame

Thank you!

Q&A