# Runtime Safety for AI-Enabled Cyber-Physical Systems

Lu Feng

Department of Computer Science

University of Virginia

QEST+FORMATS 2025, Aarhus, Denmark

*My very first research paper was published at QEST 2010.*

2010 Seventh International Conference on the Quantitative Evaluation of Systems

# Compositional Verification of Probabilistic Systems using Learning

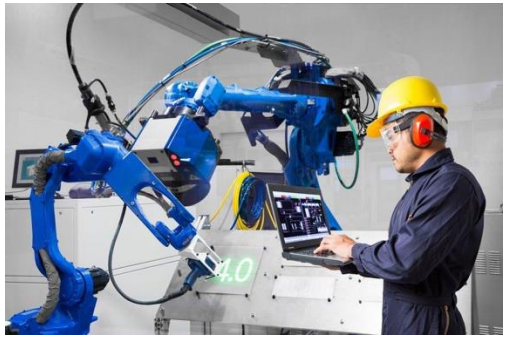
Lu Feng, Marta Kwiatkowska, David Parker
*Oxford University Computing Laboratory, Parks Road, Oxford, OX1 3QD*
*Email: {lu.feng, marta.kwiatkowska, david.parker}@comlab.ox.ac.uk*

# Cyber-Physical Systems (CPS)

Tight integration of networked computation with the physical world.

Self-driving Car

Medical Devices



Robotics

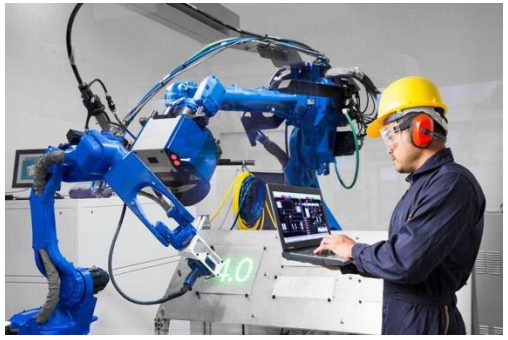
Smart Cities

# Cyber-Physical Systems (CPS)

Tight integration of networked computation with the physical world.
Safety is paramount: CPS failures are costly and dangerous.

Self-driving Car
→ accidents can cost lives

Medical Devices
→ errors can harm patients

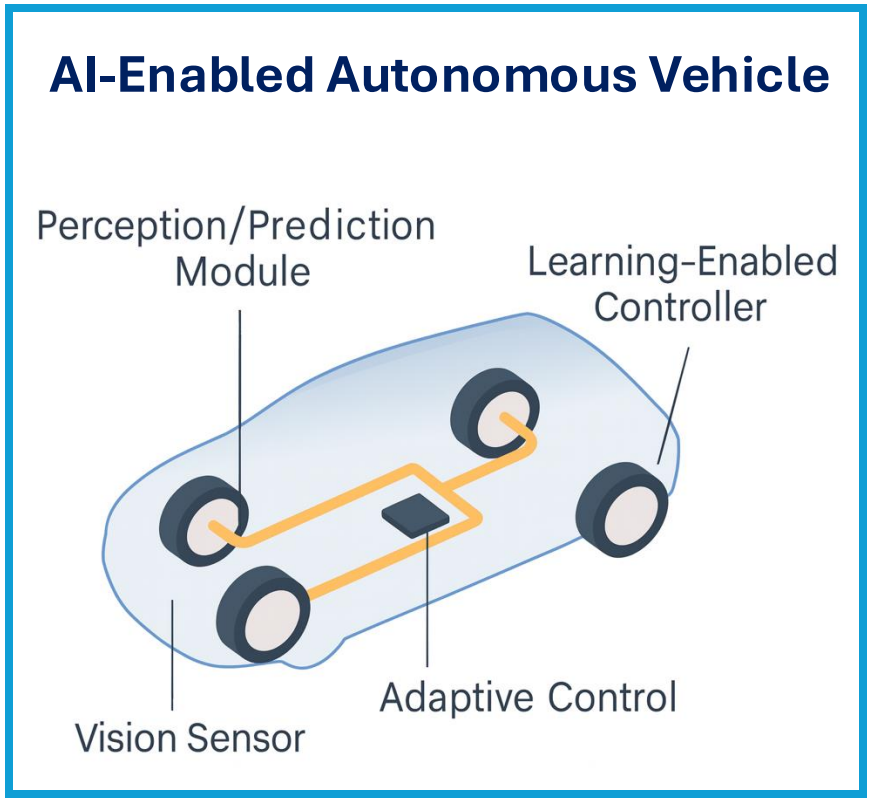


Robotics
→ unsafe actions risk workers

Smart Cities
→ failures disrupt thousands

# AI-Enabled CPS

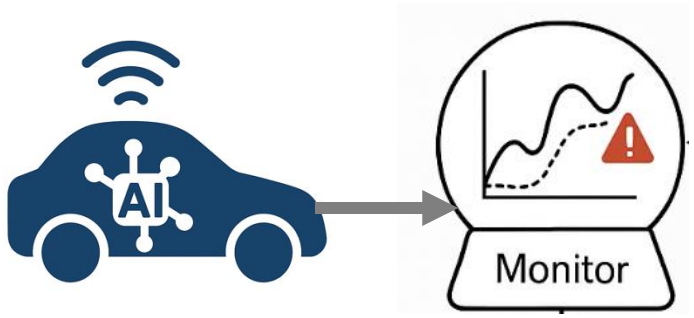
- Increasing use of AI/ML for perception, prediction, and control
- AI components are adaptive, uncertain, and data-driven
- Safety assurance is harder than in traditional CPS
- Need runtime safety mechanisms



# Runtime Safety for AI-Enabled CPS

## Predictive Monitoring:

Anticipate unsafe behavior before it happens




[EMSOFT'21, AAAI'25]

## Shielding:

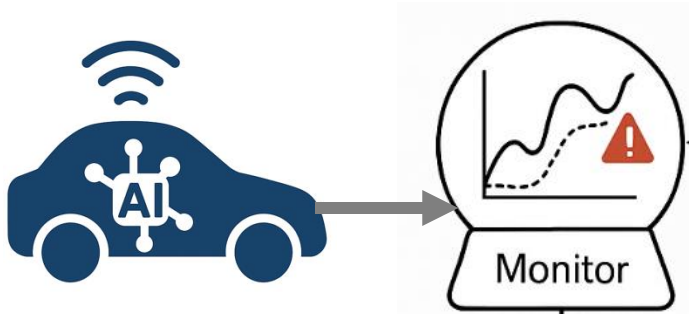
Block unsafe actions at runtime


[AAMAS'21, ICRA'24, RA-L'24]

# Runtime Safety for AI-Enabled CPS

## Predictive Monitoring:

Anticipate unsafe behavior before it happens




[EMSOFT'21, AAAI'25]

## Shielding:

Block unsafe actions at runtime


[AAMAS'21, ICRA'24, RA-L'24]

# Predictive Monitoring with Uncertainty

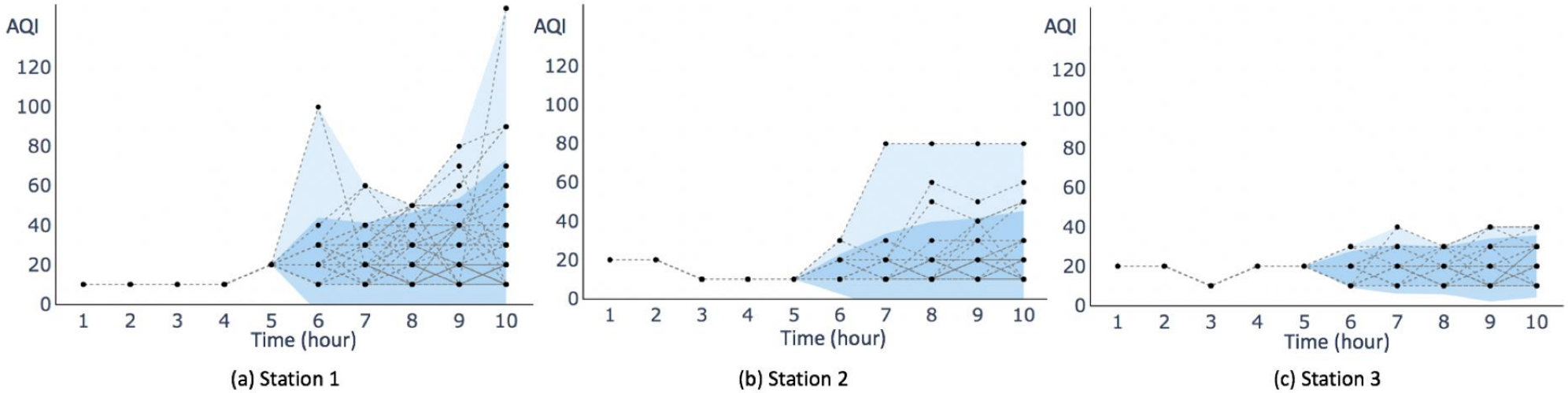
- Enhances CPS decision-making at runtime
- Examples:
  - Adapt traffic signals based on predicted congestion
  - Adjust insulin dosage when hypoglycemia is predicted
- Prior work: monitoring individual predictions only
- Our contribution: monitoring **sequential predictions** to capture CPS uncertainty

Ma, Stankovic, Bartocci, **Feng**. “Predictive monitoring with logic-calibrated uncertainty for CPS”. EMSOFT 2021.

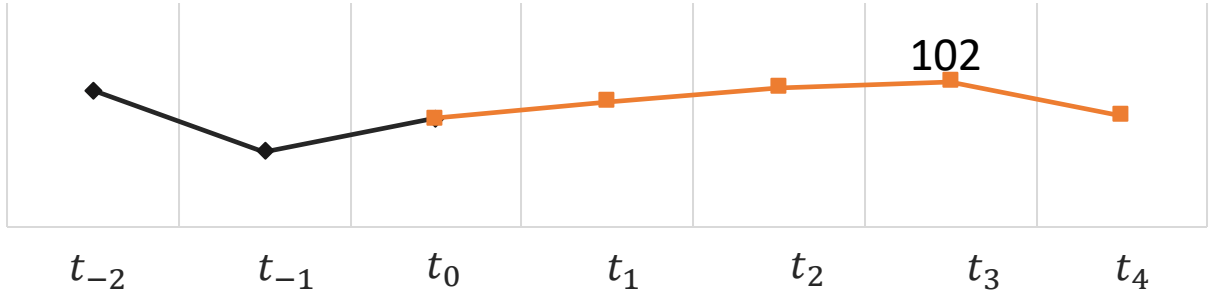
# Insights from Real-World CPS Datasets

- Uncertainty in CPS data
  - Sensing noise
  - Environment
  - Human behavior

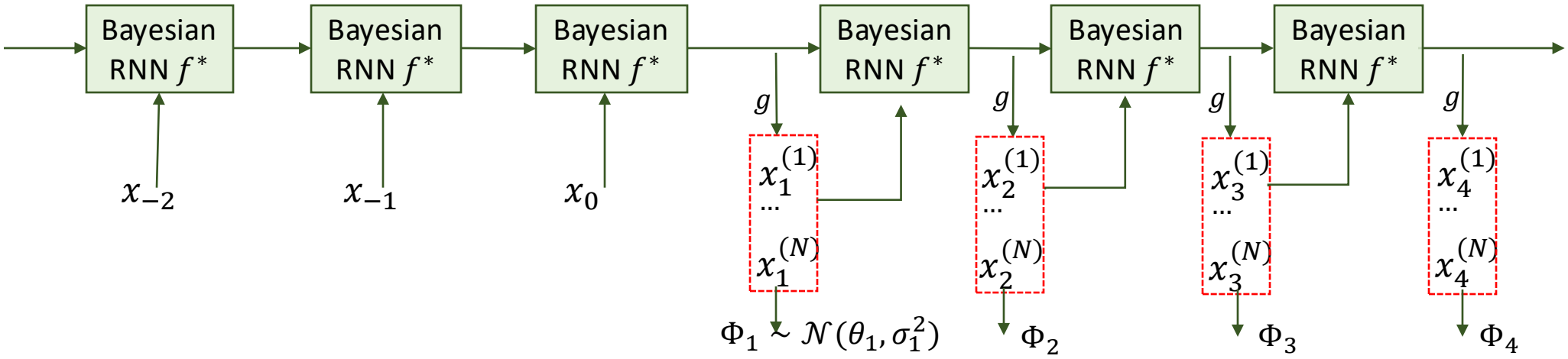
| Dataset | Location | Period | # Records |
|---|---|---|---|
| Air quality | 437 stations | 5/2014-4/2015 | 2,891,393 |
| Traffic volume | 1,490 streets | 9/2014-4/2018 | 514,776 |

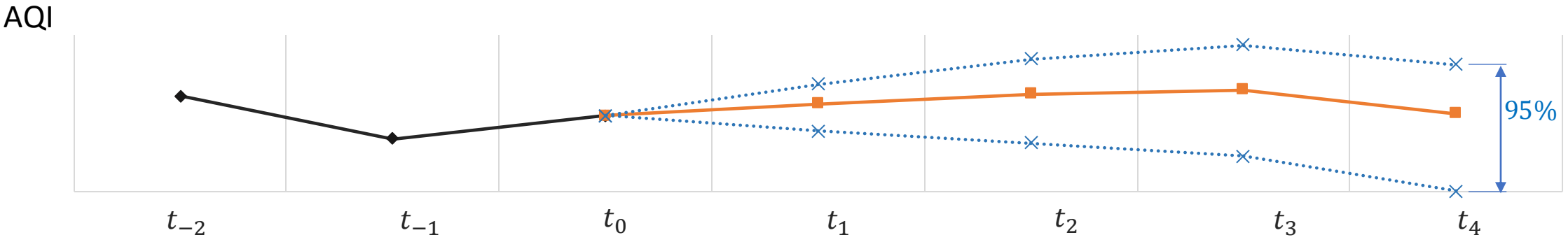


# Example: Air Quality Monitoring

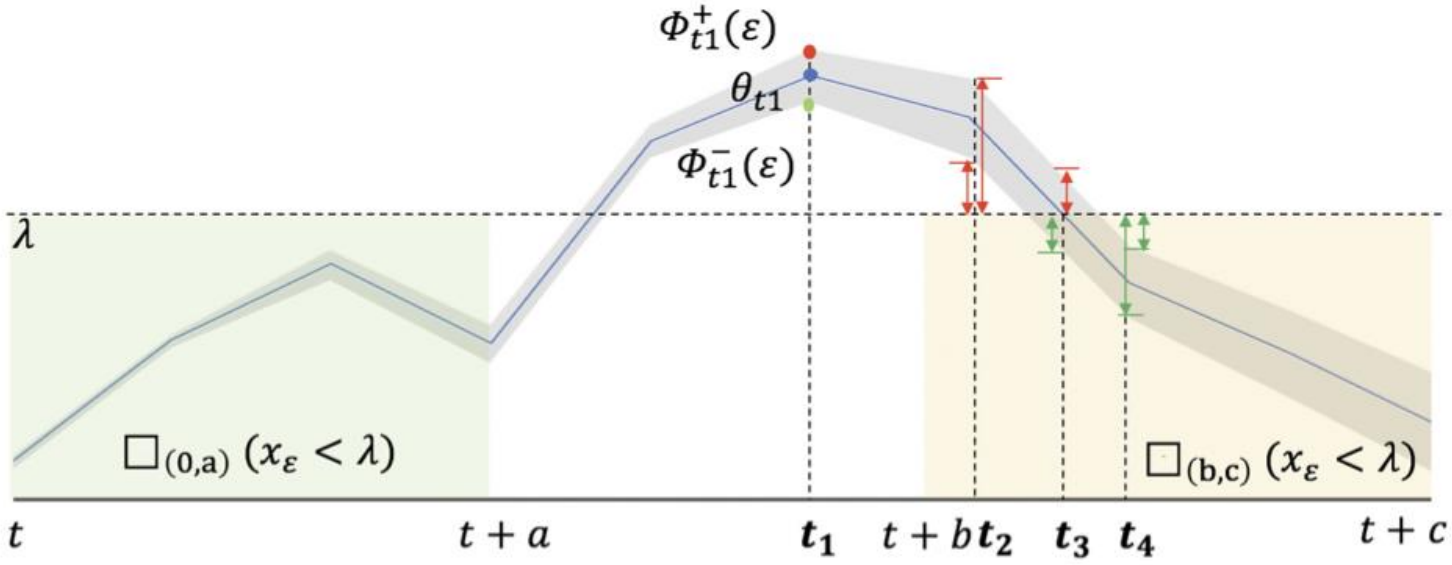



# Uncertainty in Deep Learning

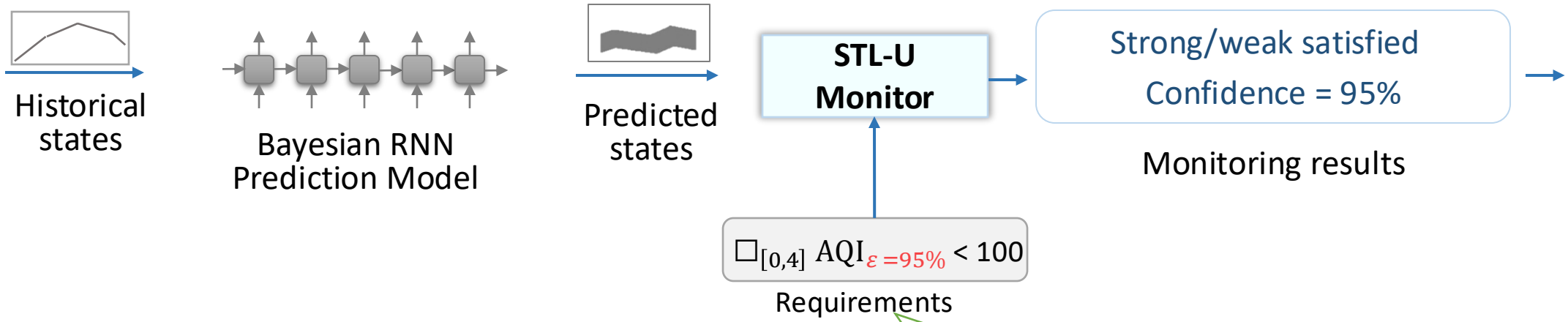


# STL-U: Signal Temporal Logic with Uncertainty



# Predictive Monitoring with STL-U

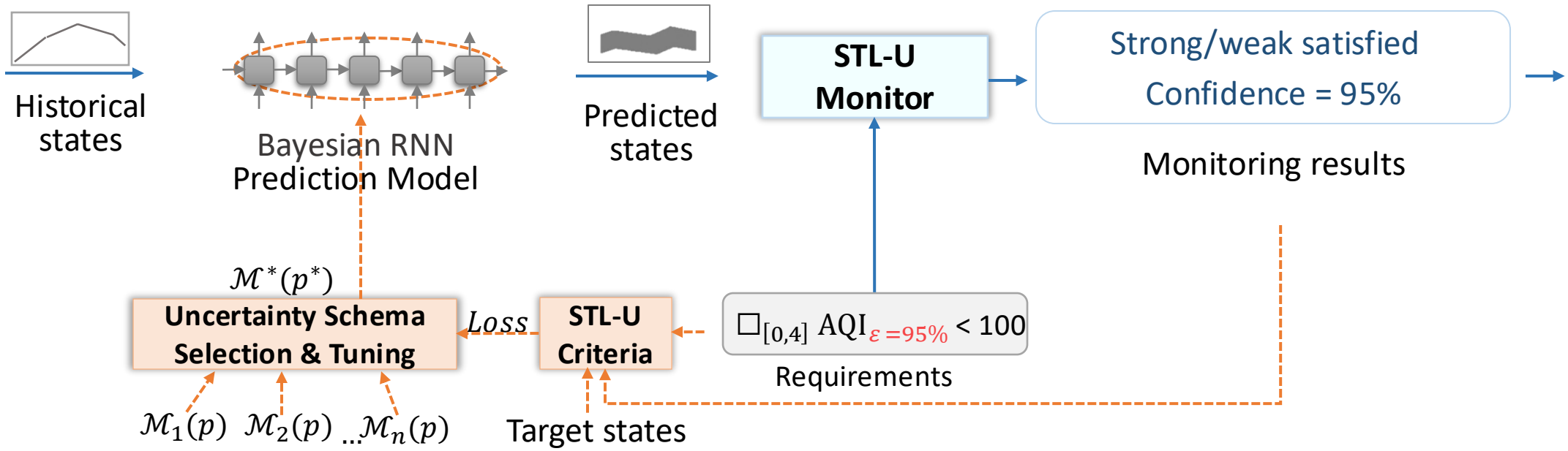



With 95% confidence level, the predicated air quality index in the next 4 hours should always be below 100

- A Novel Specification Logic: STL-U
  - Strong/Weak Semantics
  - Confidence Calculation

What is the confidence level that guarantees the predicated air quality index in the next 4 hours always be below 100

# Predictive Monitoring with STL-U

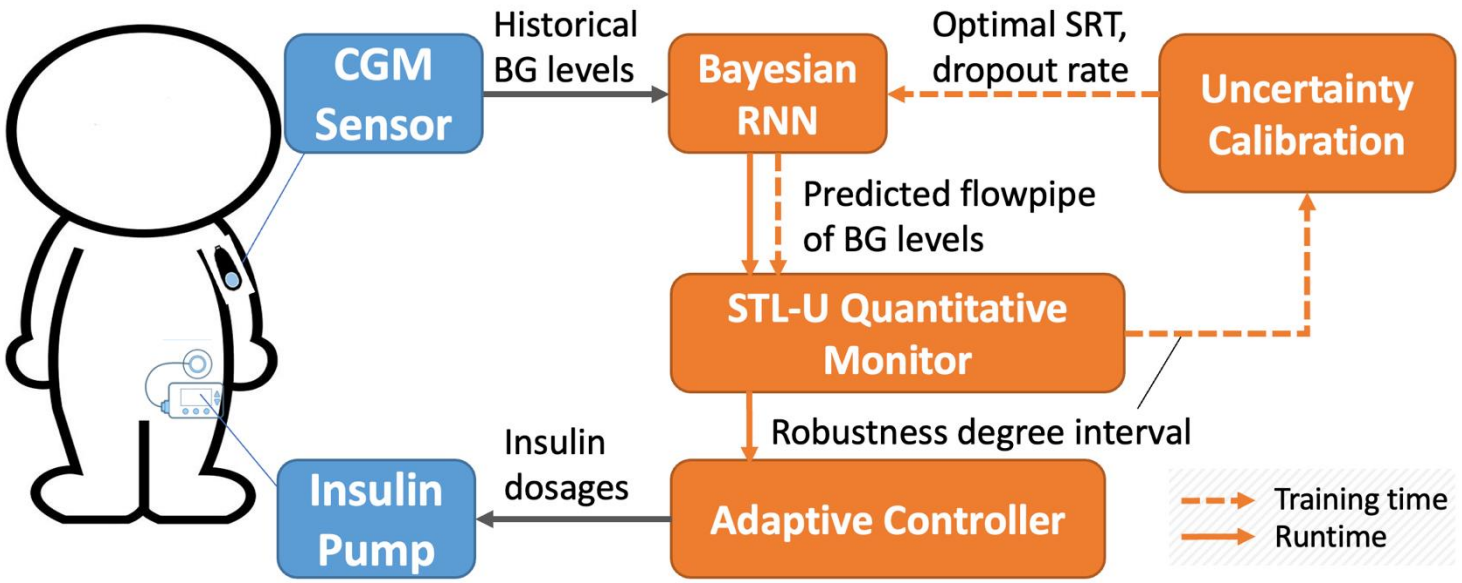


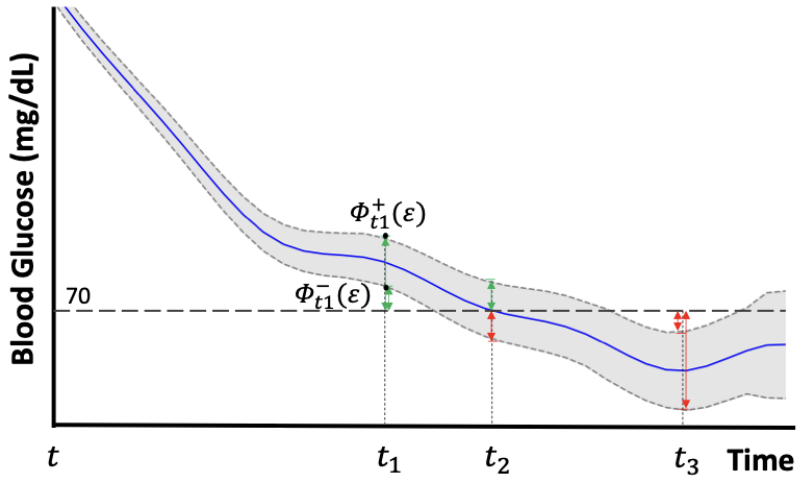
# Evaluation: Smart City

| | No Monitor | STL Monitor | STL-U Monitor | |
|---|---|---|---|---|
| Number of Violation | undetected | 267 | 189 | |
| Air Quality Index | 67.91 | 57.22 | 43.65 | 23.7% |
| Noise (db) | 73.32 | 49.27 | 48.21 | |
| Emergency Waiting Time (s) | 20.32 | 14.87 | 10.65 | 28.3% |
| Vehicle Waiting Number | 22 | 18 | 15 | |
| Pedestrian Waiting Time (s) | 190.2 | 148.9 | 121.1 | |
| Vehicle Waiting Time (s) | 112.12 | 89.77 | 80.31 | |

☞ City safety & performance ⬆

# Application to Diabetes Management




Dong, Ma, Lamp, Elbaum, Dwyer, **Feng**. "Quantitative predictive monitoring and control for safe human-machine interaction." AAAI 2025.

# Evaluation: Diabetes Management

| Model | Hazard | Baseline Time | Baseline F1 | Proposed Time | Proposed F1 |
|---|---|---|---|---|---|
| Adults | Hypo | 0.6 | 0.54 | **23.9** | **0.96** |
| | Hyper | 1.9 | 0.56 | **22.2** | **0.63** |
| | Overall | 1.2 | 0.54 | **23.0** | **0.93** |
| Adolescents | Hypo | 4.2 | **0.57** | **24.9** | 0.48 |
| | Hyper | 10.6 | **0.82** | **22.6** | 0.78 |
| | Overall | 9.2 | **0.78** | **23.1** | 0.71 |
| Children | Hypo | 3.9 | 0.89 | **13.1** | **0.91** |
| | Hyper | 21.2 | 0.71 | **27.7** | **0.75** |
| | Overall | 10.6 | 0.88 | **18.7** | **0.90** |

| Model | Metric | Baseline | Proposed |
|---|---|---|---|
| Adults | Time in range | 90.9% | **96.3%** |
| | Hypo time | 4.8% | **2.4%** |
| | Hyper time | 4.3% | **1.3%** |
| Adolescents | Time in range | 80.9% | **85.2%** |
| | Hypo time | 4.2% | **3.3%** |
| | Hyper time | 14.9% | **11.5%** |
| Children | Time in range | 63.6% | **74.8%** |
| | Hypo time | 29.4% | **19.3%** |
| | Hyper time | 7.0% | **5.9%** |

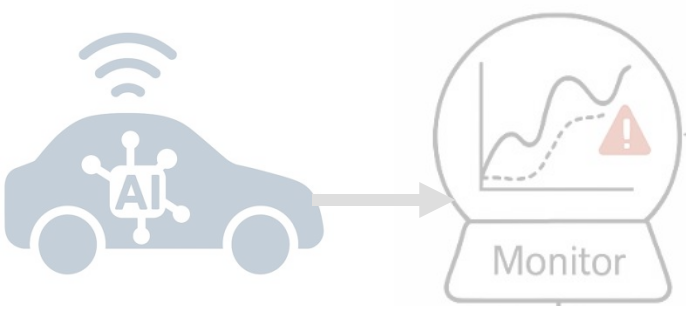
Predictive monitor enables **early, accurate detection** of impending safety hazards.

Closed-loop simulations show predictive monitoring and control can **improve safety** in diabetes management.

# Runtime Safety for AI-Enabled CPS

## Predictive Monitoring:

Anticipate unsafe behavior before it happens




[EMSOFT'21, AAAI'25]

## Shielding:

Block unsafe actions at runtime


[AAMAS'21, ICRA'24, RA-L'24]

# Safe Multi-Agent Reinforcement Learning via Shielding

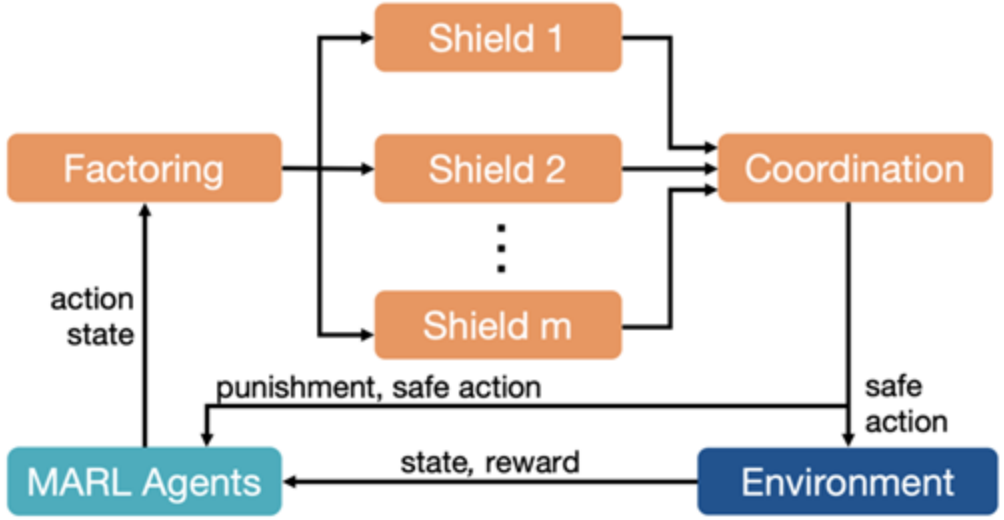
- MARL: widely applied in real-world CPS
- Traditional MARL: optimize returns, but allow unsafe actions
- Our methods: provide **safety guarantees** during both learning and execution

ElSayed-Aly, Bharadwaj, Amato, Ehlers, Topcu, **Feng**. “Safe multi-agent reinforcement learning via shielding”. AAMAS 2021.

# Safety Shielding for MARL

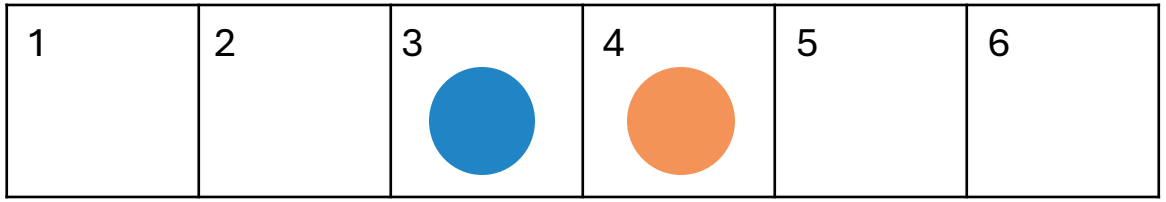
*Centralized Shielding*

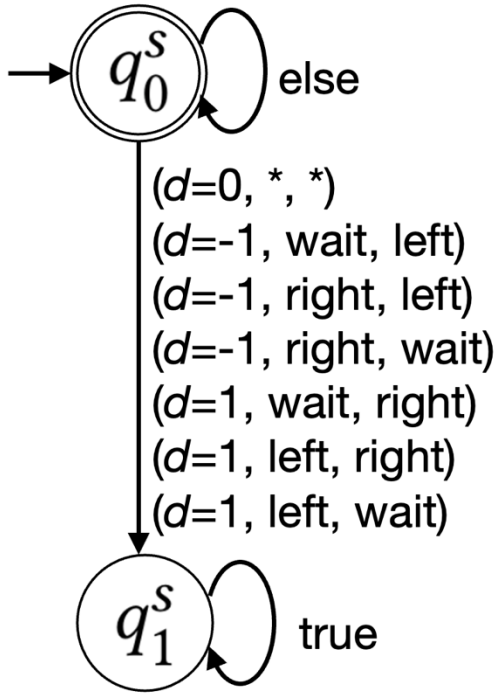

*Factored Shielding*



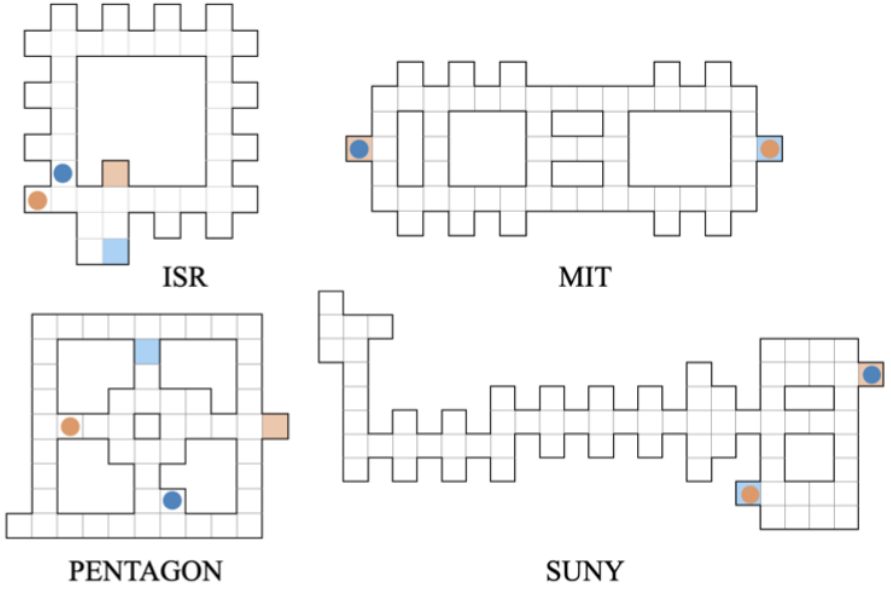
# Shield Construction

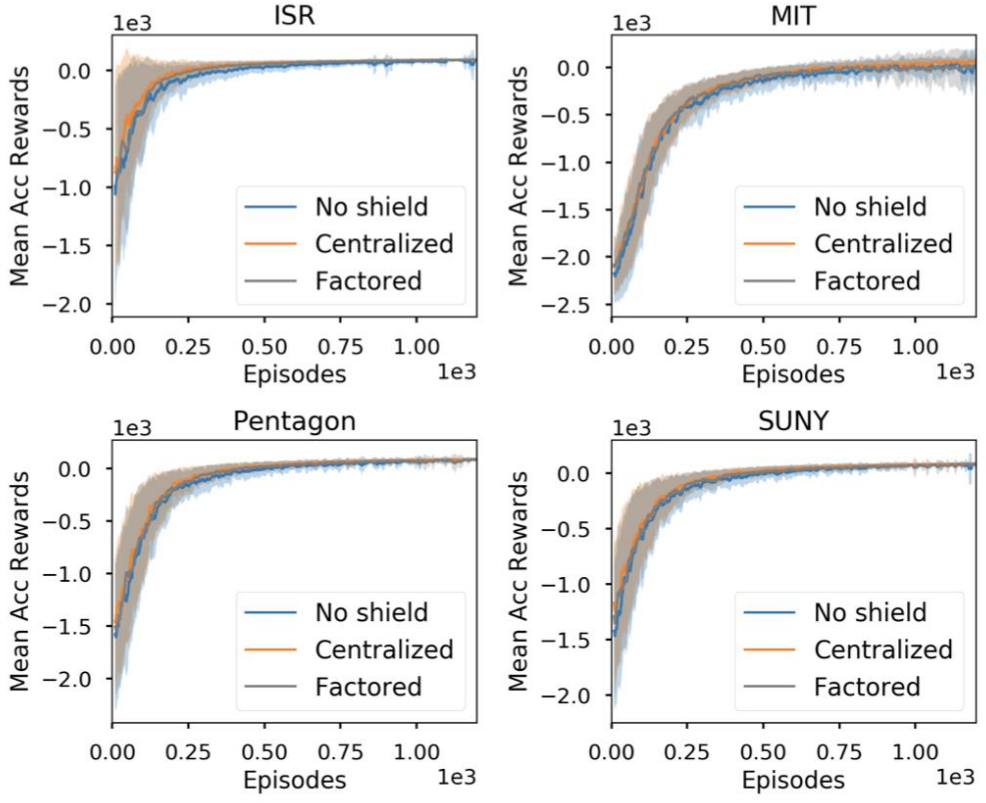
- Safety specification in Linear Temporal Logic
- Synthesizing shields by solving two-player safety games



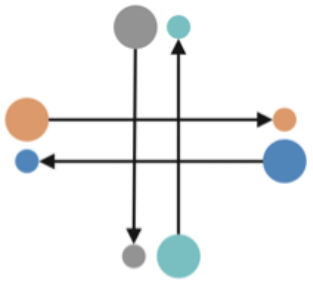
# Evaluation on Discrete Environments



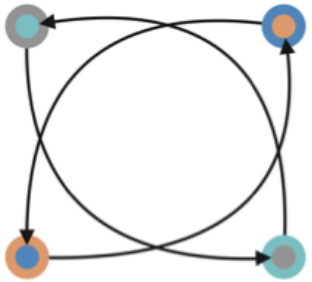




| | | IQL | | | CQ | | | CQ with centralized shield | | | CQ with factored shield | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Maps | Optimal Steps | Steps | Reward | Collisions | Steps | Reward | Collisions | Steps | Reward | Collisions | Steps | Reward | Collisions |
| ISR | 5 | 30.35 | -10.20 | 20.30 | 8.66 | 89.53 | 0.40 | 7.03 | 93.85 | **0.00** | 7.31 | 93.74 | **0.00** |
| Pentagon | 10 | 46.58 | -19.17 | 11.60 | 10.96 | 88.96 | 0.20 | 12.08 | 88.44 | **0.00** | 13.20 | 84.88 | **0.00** |
| MIT | 18 | 20.84 | 77.33 | 0.00 | 42.93 | 30.38 | 0.90 | 28.38 | 73.94 | **0.00** | 29.96 | 37.96 | **0.00** |
| SUNY | 10 | 34.80 | -160.175 | 72.60 | 13.97 | 84.78 | 0.30 | 11.97 | 88.44 | **0.00** | 14.02 | 83.77 | **0.00** |

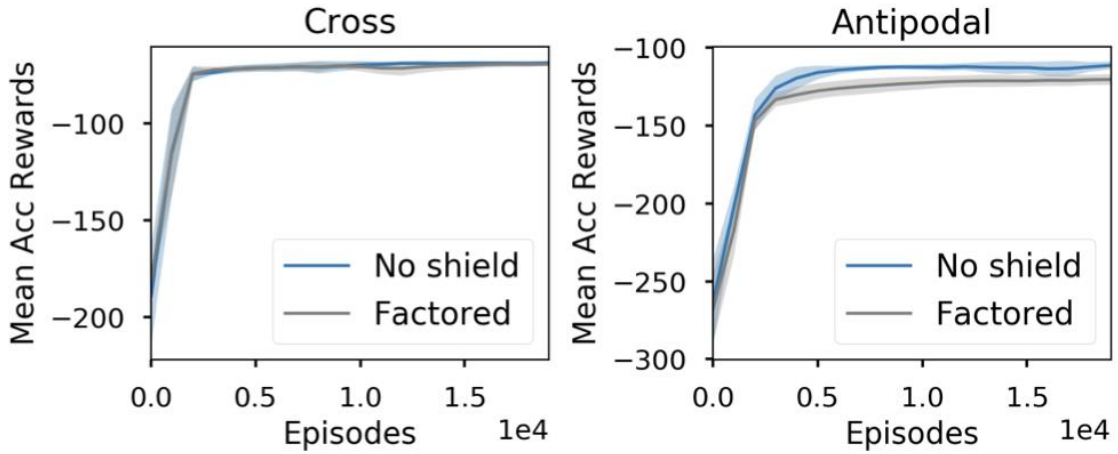
# Evaluation on Continuous Environments



Cross



Antipodal

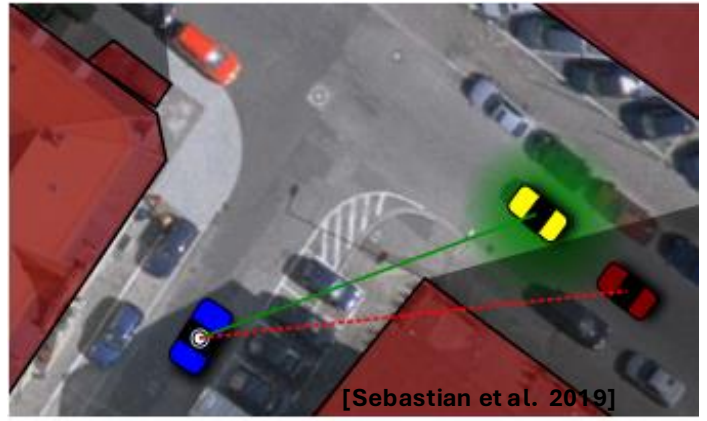



| | MADDPG | MADDPG with Shield |
|---|---|---|
| Cross | 207.20 | **0.00** |
| Antipodal | 14,419.20 | **0.00** |

# Safe POMDP Online Planning via Shielding

- POMDP: general modeling framework for decision-making under uncertainty

- POMDP online planning
  - Interleave policy computation and execution
  - Can scale up to solve very large POMDPs than offline planning

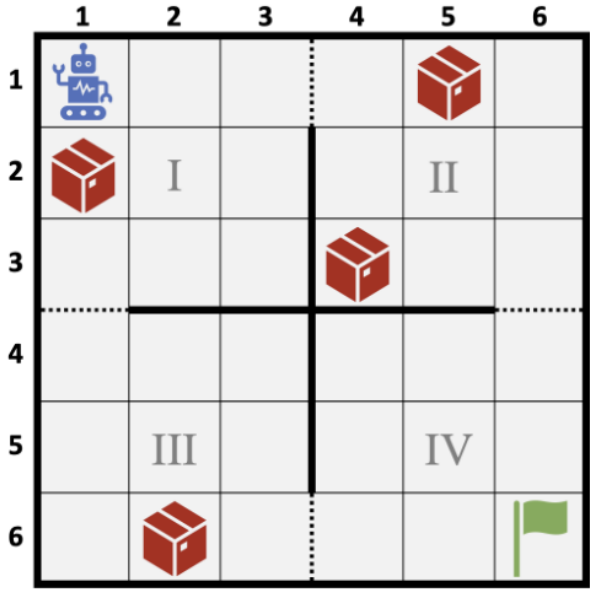




Sheng, Parker, **Feng**. Safe POMDP online planning via shielding. ICRA 2024.

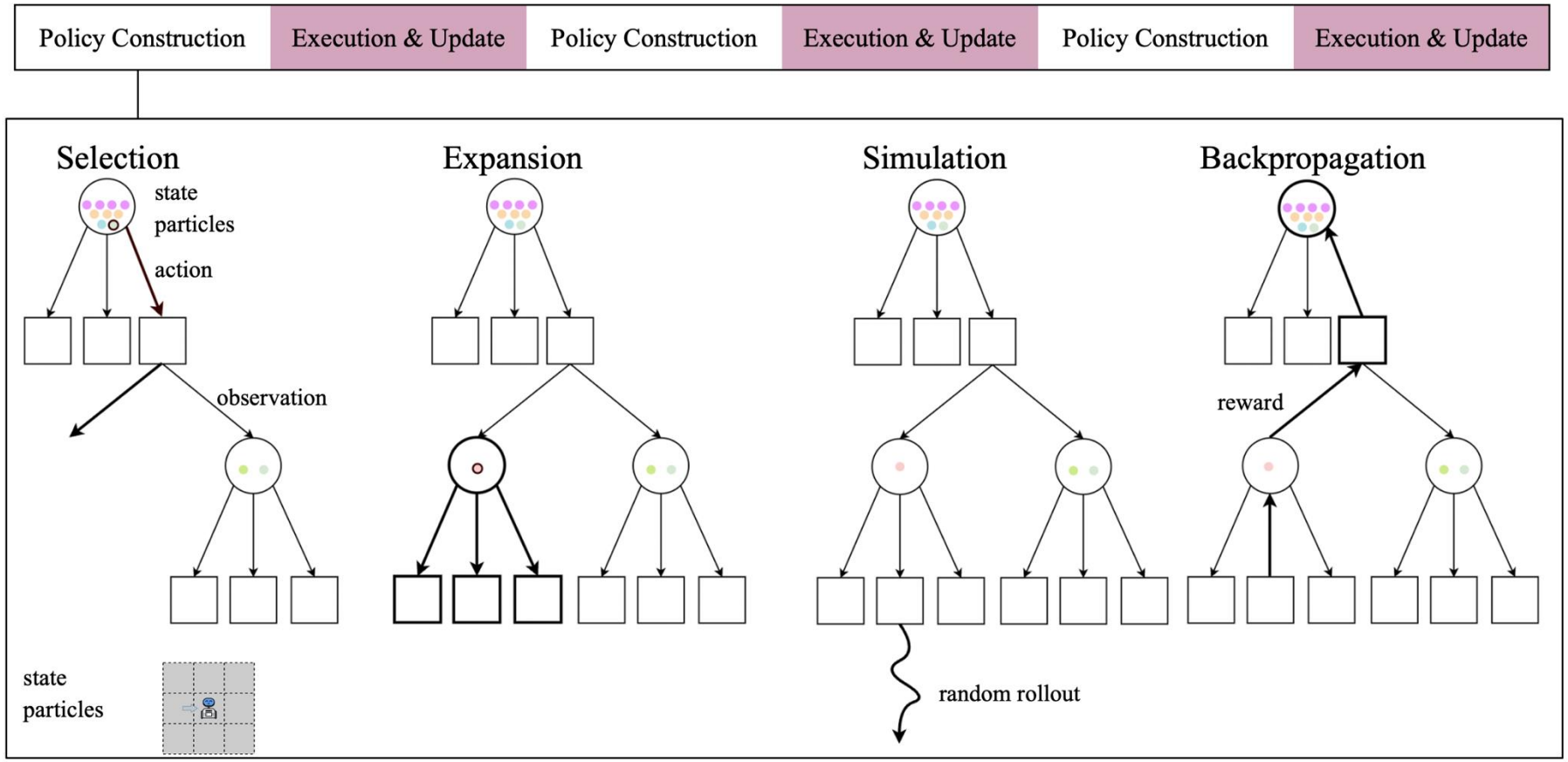
# Safe POMDP Online Planning via Shielding

- Existing methods: cost-constrained, chance-constrained
- Our focus: stricter safety — almost-sure reach-avoid specifications
- Approach:
  - Shield synthesis via SAT-based maximal winning regions (Junges et al. 2021)
  - New method: factored shields by decomposing a POMDP into sub-models



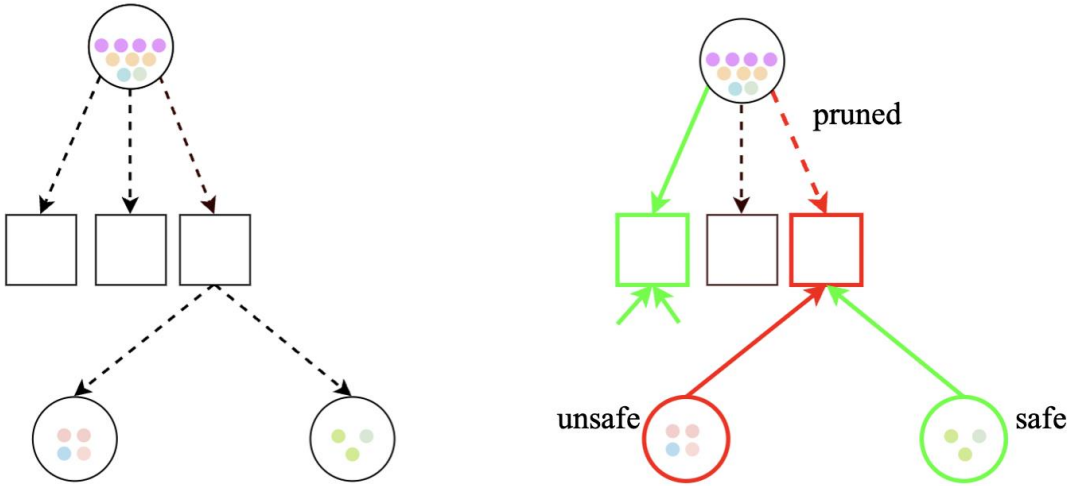
# Partially Observable Monte-Carlo Planning (POMCP)

- A widely used POMDP online planning algorithm (Silver et al. 2008)



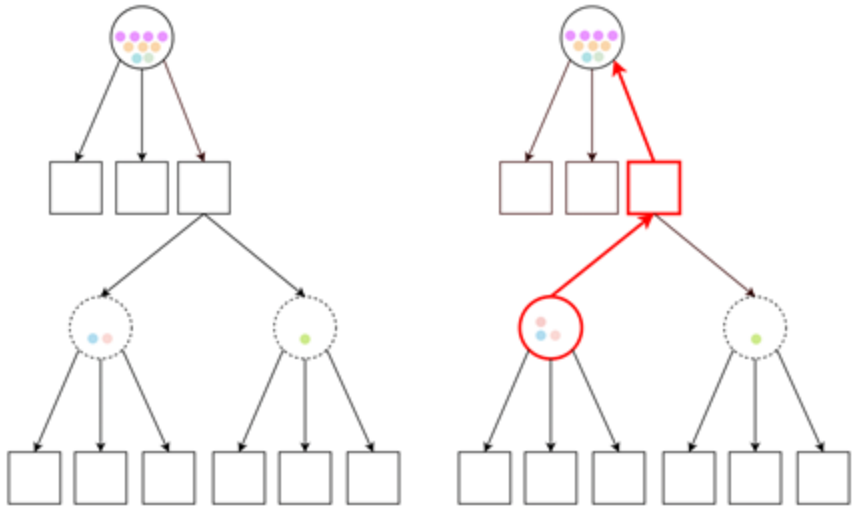
# Prior Pruning

- At each time step *t*:
  - Identify actions disallowed by the shield
  - Prune corresponding branches from the root node in the POMCP tree



# On-the-Fly Backtracking

- During simulation:
  - For each updated particle set along the path
  - Check containment in shield's winning region
  - If not contained → prune the branch



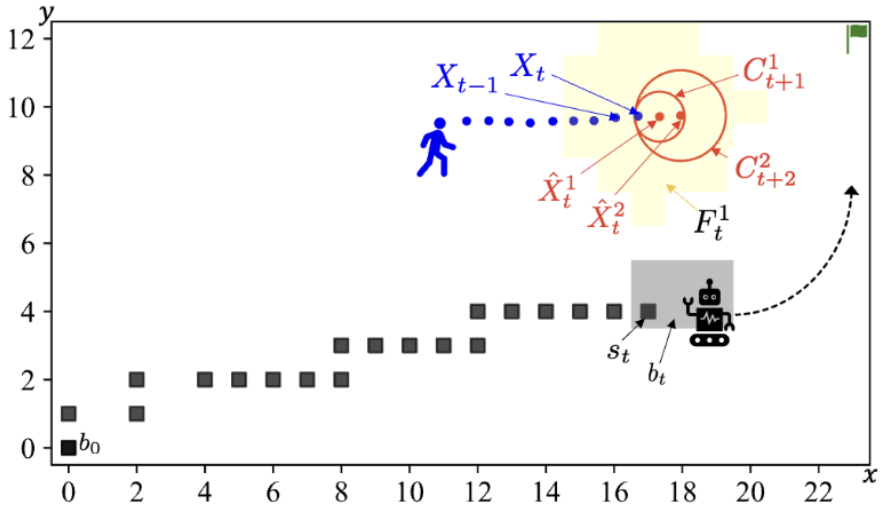
# Experimental Results

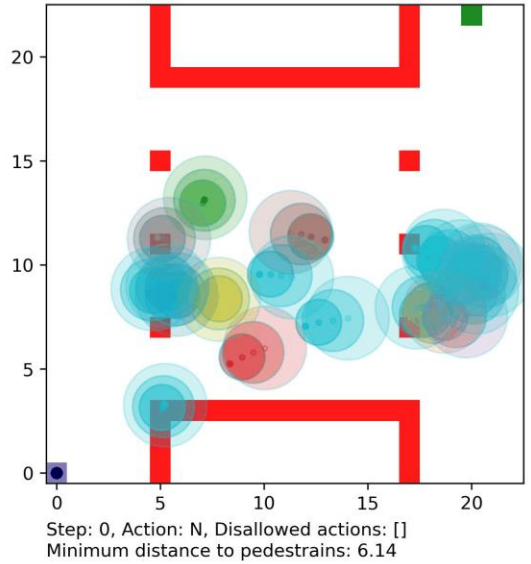
| Case Study | | | | | No Shield | | | Centralized Shield | | | | | | | Factored Shield | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Prior | | | On-the-fly | | | | Prior | | | On-the-fly | |
| Domain | Para. | $\|S\|$ | $\|O\|$ | $\|T\|$ | Time(s) | Return | Unsafe | Time(s | Return | Unsafe | Time(s) | Return | Unsafe | Time(s | Return | Unsafe | Time(s) | Return | Unsafe |
| Obstacle (N) | 6 | 37 | 20 | 204 | 0.09 | 978.9 | 1.0 | 0.12 | 929.9 | 0 | 0.20 | 968.1 | 0 | 0.12 | 921.7 | 0 | 0.16 | 968.7 | 0 |
| | 8 | 65 | 20 | 396 | 0.14 | 972.0 | 2.0 | 0.16 | 977.1 | 0 | 0.19 | 980.1 | 0 | 0.16 | 962.9 | 0 | 0.21 | 979.8 | 0 |
| | 9 | 82 | 39 | 464 | 0.13 | 974.0 | 1.9 | 0.18 | 944.0 | 0 | 0.32 | 962.0 | 0 | 0.18 | 939.3 | 0 | 0.28 | 964.0 | 0 |
| Refuel (N, E) | 6, 8 | 272 | 74 | 1,081 | 0.84 | 861.4 | 20.4 | 0.14 | 795.9 | 0 | 0.53 | 939.9 | 0 | 0.14 | 912.8 | 0 | 0.53 | 917.8 | 0 |
| | 9, 6 | 470 | 151 | 1,848 | 1.19 | 741.9 | 40.0 | 1.36 | -261.6 | 0 | 0.81 | 904.5 | 0 | 1.41 | -259.6 | 0 | 0.81 | 760.8 | 0 |
| | 12, 8 | 1,081 | 180 | 5,003 | 1.54 | -199.0 | 192.1 | - | - | - | - | - | - | 1.15 | -248.4 | 0 | 0.62 | 933.6 | 0 |
| Rocks (N, R) | 6, 3 | 4,157 | 596 | 4.3e**4** | 0.16 | 1,001.1 | 0.7 | 0.33 | 492.6 | 0 | 0.36 | 1,020.9 | 0 | 0.32 | 840.6 | 0 | 0.32 | 1,062.3 | 0 |
| | 8, 4 | 3.7e**4** | 2,036 | 4.7e**5** | 0.54 | 1,013.3 | 0.5 | - | - | - | - | - | - | 0.99 | 406.6 | 0 | 0.73 | 1,091.5 | 0 |
| | 9, 6 | 1.2e**6** | 3.3e**4** | 1.8e**7** | 0.87 | 1,008.0 | 1.9 | - | - | - | - | - | - | 1.33 | -119.0 | 0 | 1.40 | 540.4 | 0 |

- Proposed methods guarantee safety; baseline does not
- Comparable search time per planning step
- On-the-fly backtracking → higher expected return than prior pruning
- Factored shielding → better scalability than centralized shielding

# Adaptive Shielding for POMDP Online Planning

1. Predict dynamic agents' future trajectories and quantify prediction uncertainty with adaptive conformal prediction
2. Online computation of winning regions based on trajectory predictions
3. Shielding POMCP based on the obtained winning regions




Sheng, Yu, Parker, Kwiatkowska, **Feng**. Safe POMDP online planning among dynamic agents via adaptive conformal prediction. IEEE RA-L 2024.

# Experimental Results

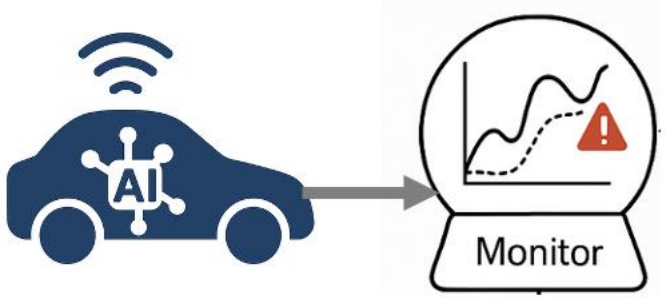
| | ETH | | | | Hotel | | | | GC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $N$ | Safety Rate | Time (s) | Min Distance | $N$ | Safety Rate | Time (s) | Min Distance | $N$ | Safety Rate | Time (s) | Min Distance |
| No Shield | 45 | 0.893 | 21.1 | 0.28±0.19 | 35 | 0.944 | 20.1 | 0.42±0.21 | 160 | 0.91 | 39.3 | 0.22±0.13 |
| Shielding without ACP | | 0.943 | 21.5 | 0.39±0.29 | | 0.969 | 20.3 | 0.54±0.3 | | 0.943 | 67.2 | 0.23±0.13 |
| Shielding with ACP | | **0.974** | 22.1 | 0.51±0.29 | | **0.988** | 20.6 | 0.8±0.49 | | **0.963** | 71.4 | 0.28±0.16 |
| No Shield | 55 | 0.891 | 21.0 | 0.26±0.17 | 45 | 0.931 | 20.1 | 0.38±0.24 | 180 | 0.904 | 39.9 | 0.2±0.1 |
| Shielding without ACP | | 0.951 | 21.8 | 0.41±0.25 | | 0.959 | 20.1 | 0.48±0.24 | | 0.938 | 66.8 | 0.23±0.12 |
| Shielding with ACP | | **0.975** | 22.4 | 0.53±0.37 | | **0.982** | 20.6 | 0.62±0.27 | | **0.953** | 71.1 | 0.24±0.16 |
| No Shield | 65 | 0.872 | 21.2 | 0.24±0.13 | 55 | 0.921 | 20.2 | 0.36±0.18 | 200 | 0.895 | 39.6 | 0.22±0.11 |
| Shielding without ACP | | 0.943 | 21.9 | 0.36±0.2 | | 0.957 | 20.3 | 0.48±0.29 | | 0.931 | 65.7 | 0.2±0.13 |
| Shielding with ACP | | **0.967** | 22.6 | 0.42±0.26 | | **0.982** | 20.3 | 0.6±0.24 | | **0.951** | 74.3 | 0.25±0.15 |

- Proposed ACP-based shielding achieves higher safety rates than baselines
- Comparable travel times in ETH/Hotel; longer in GC due to dense pedestrians
- More conservative: maintains larger minimum robot–pedestrian distances

# Conclusion

- Runtime safety is essential for AI-enabled CPS in safety-critical domains
- Case studies in smart cities, healthcare, and autonomous vehicles demonstrate tangible impact
- **Future challenges**: with generative AI and LLMs embedded into CPS, ensuring safety and trustworthiness becomes even more critical

Predictive Monitoring




Shielding

# Thank you!
# Questions and Comments?

Lu Feng
lu.feng@virginia.edu