Safe Multi-Agent Reinforcement Learning with Shielding

Stavros Tripakis and Chris Amato

Northeastern University
Khoury College of Computer Sciences

Challenge:

- Want to solve cooperative multi-agent systems while being safe
- How can we combine *multi-agent reinforcement learning* (MARL) for high-performance with *formal methods* for guaranteed safety?

Solution:

Broken up into three trusts:

1. Decentralized Shields for Safe Execution of MARL Systems
2. Safety Coaches for Safety-Oriented MARL Training
3. New Environment Abstraction Methods

Have approaches for 1) published at major machine learning conferences (NeurIPS and RLC)

Project info (2319500, Northeastern University, stavros@northeastern.edu, c.amato@northeastern.edu)

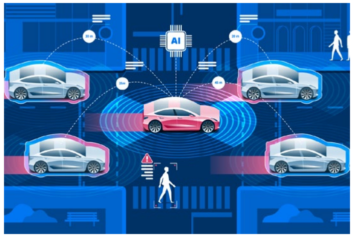
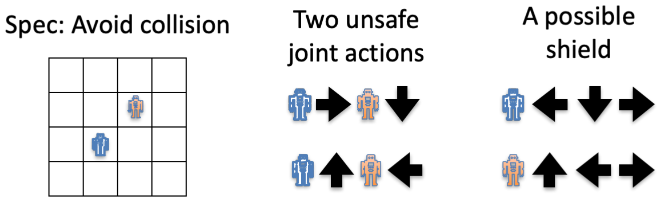


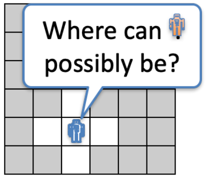
Image from SmartCitiesWorld



Spec: Avoid collision

Two unsafe joint actions

A possible shield



Scientific Impact:

- New formal methods and concepts such as decentralized shield synthesis, partial observability, and safety coaches
- New MARL techniques such as directed exploration and training for safety, hardwiring safe policies), and
- Novel applications of model learning and abstraction refinement
- Combines FM and MARL

Broader Impact and Broader Participation:

- Should allow MARL to be used for the first time in safety-critical systems by providing rigorous safety guarantees: multi-agent systems (e.g., using LLMs) and multi-robot systems (e.g., autonomous cars)
- Broadening participation and undergraduate research: PIs have a history and plans for both

The NSF Formal Methods in the Field PI Meeting (2024 FMitF PI Meeting)
November 12-13, 2024 | The University of Iowa | Iowa City, Iowa