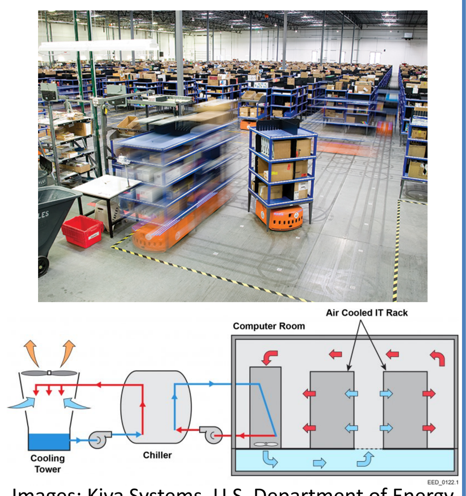
Safe Multi-Agent Reinforcement Learning with Shielding

Stavros Tripakis and Christopher Amato

Northeastern University, Boston, MA

Overview



Images: Kiva Systems, U.S. Department of Energy

- Even if agents have individual goals, they must collaborate to enforce safety.
  - Collision avoidance (warehouse robots, cars, planes), datacenter temperature control (multiple zones), etc.
- Agents may have limited communication with each other (either by default, or as a failure state).
- Two different approaches to solving tasks:

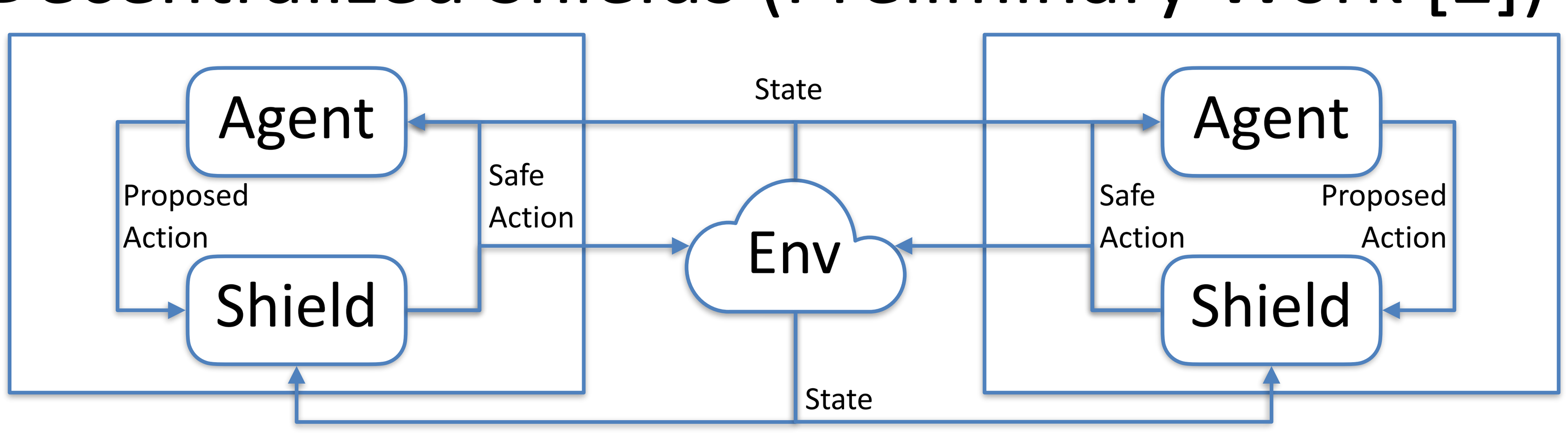
Formal Methods	**Reinforcement Learning**
• Able to guarantee safety and correctness • Difficult to scale to large environments with many agents	• Cannot guarantee safety or correctness • Can often solve problems in complex multi-agent domains

- Goal: Combine the best parts of FM and RL to provide rigorous safety guarantees that scale to large environments, while also solving agent-specific tasks.

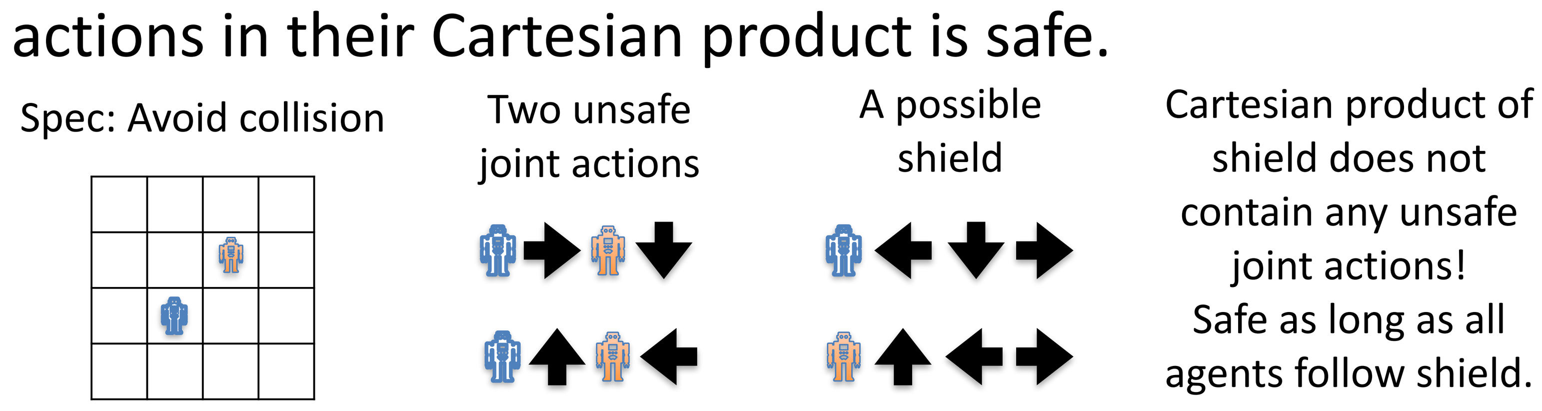
Key Problems

- In no-communication environments, can local observations be relied upon to enforce safety?
  - This problem is **undecidable** in general [1]. Can we find a solution for a useful subset of tasks?
- Is a complete manually-constructed model of the environment necessary, or can a sufficient model be learned through interaction?
- How can hard constraints be implemented without negatively affecting, or even potentially improving, reinforcement learning's ability to optimize for reward?

Decentralized Shields (Preliminary Work [2])



- Given: Model of an environment, safety specification.
  - Environment is assumed to be fully observable.
  - Safety specification is defined as set of unsafe states.
- Synthesize: Decentralized shield such that each agent can independently decide if an individual action is safe.
- Choose sets of individual actions such that all joint actions in their Cartesian product is safe.

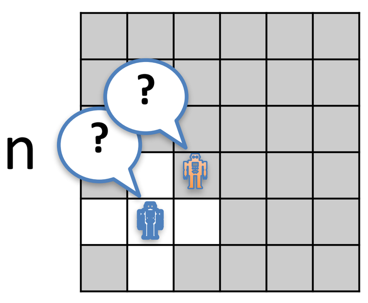


- Experiments: successfully enforces safety in gridworld collision-avoidance domain, including momentum task.
  - 0 collisions during training or evaluation, compared to thousands during training without shield.
- Next step: can we remove input assumptions?

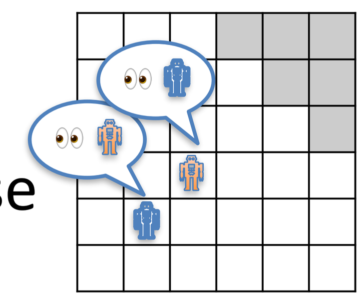
Partial Observability (RLC 2024 [3])

- Problem: most domains are not fully observable! Remove this assumption (but include observability in model).
- Example environment: Flashlight

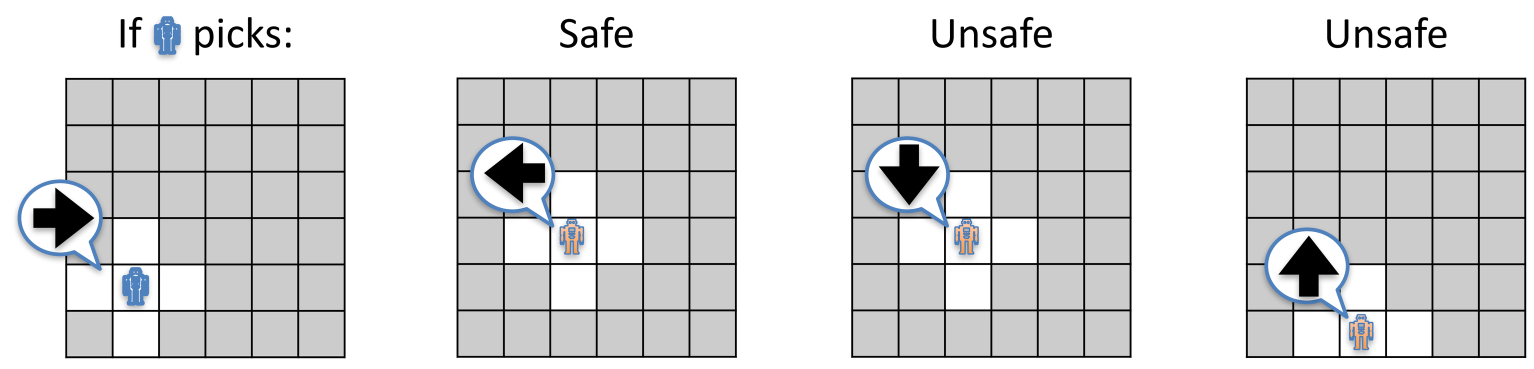
Default Observation Radius 1



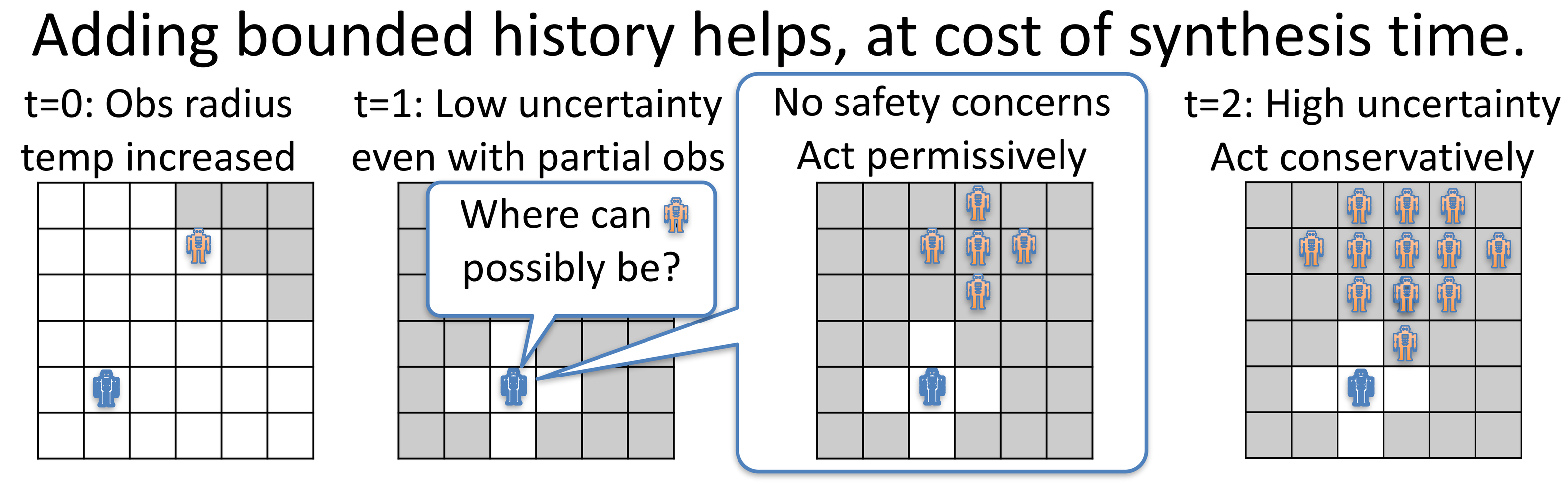
Every *RT* Steps: Action Enabled to Temporarily Increase Obs Radius to 5



- Synthesize: a decentralized shield where each agent operates on local observations, rather than global state.
- At any state, all joint actions in the Cartesian product of the enabled individual actions for that state's observations should be safe at that state.



- Low Observability: Must decide on protocol in advance.
- Encode constraints on action selection as a SAT problem.
- Adding bounded history helps, at cost of synthesis time.



- Resulting shield can be executed independently by agents with no communication in partially observable domains.
- Results: our method prevents safety violations in newly-possible domains. Table: task-specific reward over 10 seeds, with safety violations in parentheses, if any.

Start	RT	SAT (0 History)	SAT (1 History)	SAT (2 History)	Centralized	No Shield
Fixed	3	65.7 ± 1.5	69.2 ± 1.5	71.1 ± 0.8	78.6 ± 0.0	78.6 ± 0.0
	4	58.7 ± 1.5	66.5 ± 1.9	69.1 ± 1.7	78.1 ± 0.5	78.6 ± 0.0
	5	-41.6 ± 60.5	52.0 ± 11.6	65.1 ± 2.5	78.6 ± 0.0	78.6 ± 0.0
	6	-40.6 ± 41.9	16.4 ± 13.9	50.5 ± 12.3	78.6 ± 0.0	77.7 ± 0.9 (10.0)
Rand	3	65.4 ± 1.1	74.9 ± 0.3	74.4 ± 0.4	84.7 ± 0.3	83.4 ± 0.2 (7.2)
	4	53.1 ± 1.2	68.5 ± 0.4	72.0 ± 0.5	83.7 ± 0.2	82.7 ± 0.7 (5.1)
	5	-20.6 ± 14.0	56.0 ± 3.7	67.1 ± 0.7	81.6 ± 0.9	83.5 ± 0.3 (5.2)
	6	-23.8 ± 15.8	30.0 ± 12.7	62.9 ± 1.7	76.9 ± 7.0	83.5 ± 0.3 (5.5)

Learning Through Interaction (Ongoing)

- Environment does not have any model available—assume training environment allows limited safety violations.
- Learn centralized model of safety with a neural network that is structurally constrained to allow decentralization.

Broader Impacts

- Potential to allow MARL to be used for the first time in safety-critical systems by providing rigorous safety guarantees, transforming the way these systems are developed and deployed.
- PIs have a history and plans to broaden participation in research and involve undergraduate students.

References

1. Tripakis, Stavros. "Undecidable problems of decentralized observation and control on regular languages." *Information Processing Letters* 90.1 (2004): 21-28.
2. Melcer, Daniel, Christopher Amato, and Stavros Tripakis. "Shield decentralization for safe multi-agent reinforcement learning." *Advances in Neural Information Processing Systems* 35 (2022): 13367-13379.
3. Melcer, Daniel, Christopher Amato, and Stavros Tripakis. "Shield Decentralization for Safe Reinforcement Learning in General Partially Observable Multi-Agent Environments." *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2024.