A Hierarchical Bayesian Approach to Inverse Reinforcement Learning with Symbolic Reward Machines

Weichao Zhou Wenchao Li

Dependable Computing Lab
Department of Electrical and Computer Engineering
Boston University

BOSTON UNIVERSITY

Reward Design in RL Tasks

Aug 14, 2019

Designing agent incentives to avoid reward tampering

By Tom Everitt, Ramana Kumar, and Marcus Hutter From an AI safety perspective, having a clear design principle and a crisp characterization of what problem it solves means that we don't have to guess which agents ar…

Artificial Intelligence 7 min read

ELSEVIER

Reward is enough

David Silver*, Satinder Singh, Doina Precup, Richard S. Sutton

Check for updates

On the Expressivity of Markov Reward

Part of Advances in Neural Information Processing Systems 34 (NeurIPS 2021)

Bibtex Paper Reviews And Public Comment » Supplemental

Authors

David Abel, Will Dabney, Anna Harutyunyan, Mark K. Ho, Michael Littman, Doina Precup, Satinder Singh

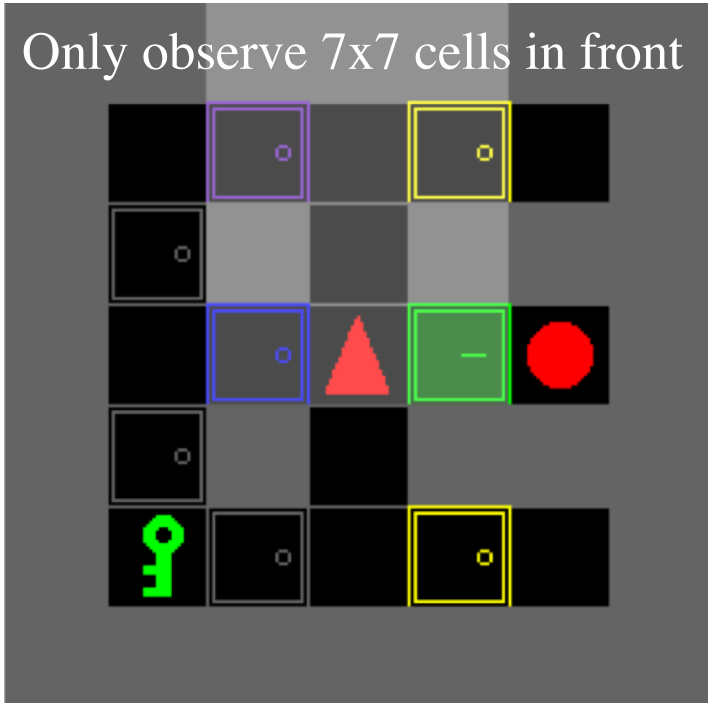
Proceedings of the AAAI Conference

Current Archives About

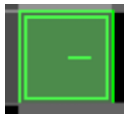
Home / Archives / Vol. 34 No. 04: AAAI-20 Technical Tracks

Reinforcement Learning with Perturbed Rewards

Motivating Example[1]




Agent


Locked door


Unlocked doors


Key


Target

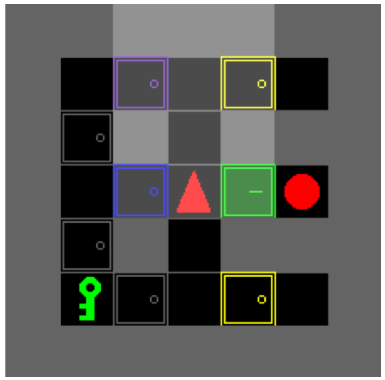
Open doors => Pick up key => Unlock door => Drop the key => (**Goal**) pick up the target

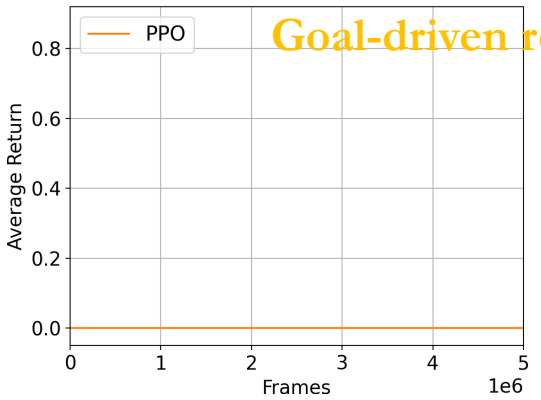
BU

1. https://github.com/Farama-Foundation/gym-minigrid

Motivating Example

reward +1 if picking up target



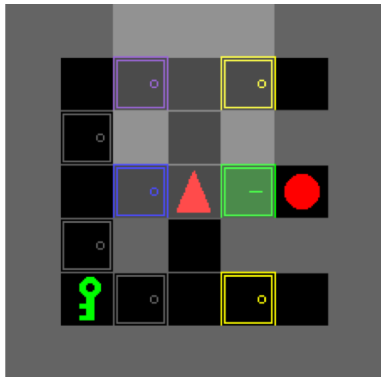


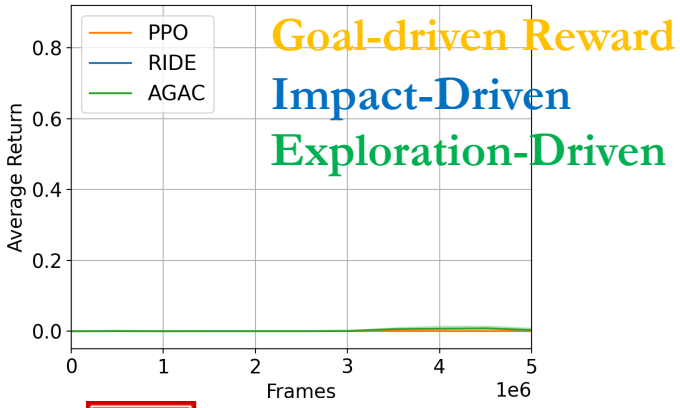
- **Goal-driven reward**
  - RL algorithms fail to solve the task

BU

Motivating Example

reward +1 if picking up target;
intrinsic reward otherwise

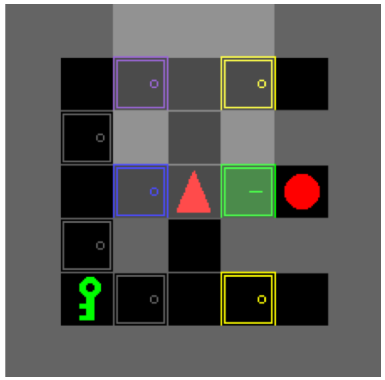


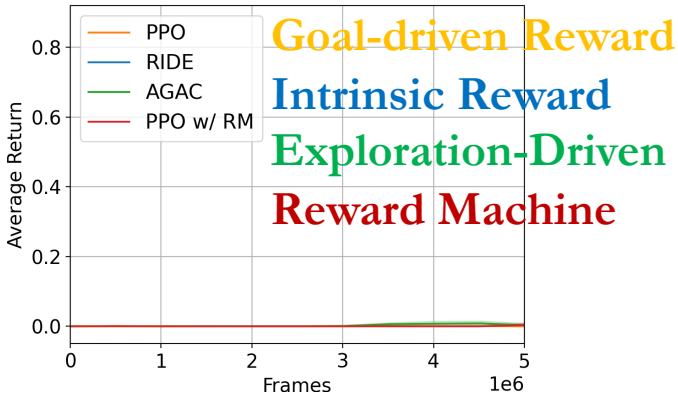


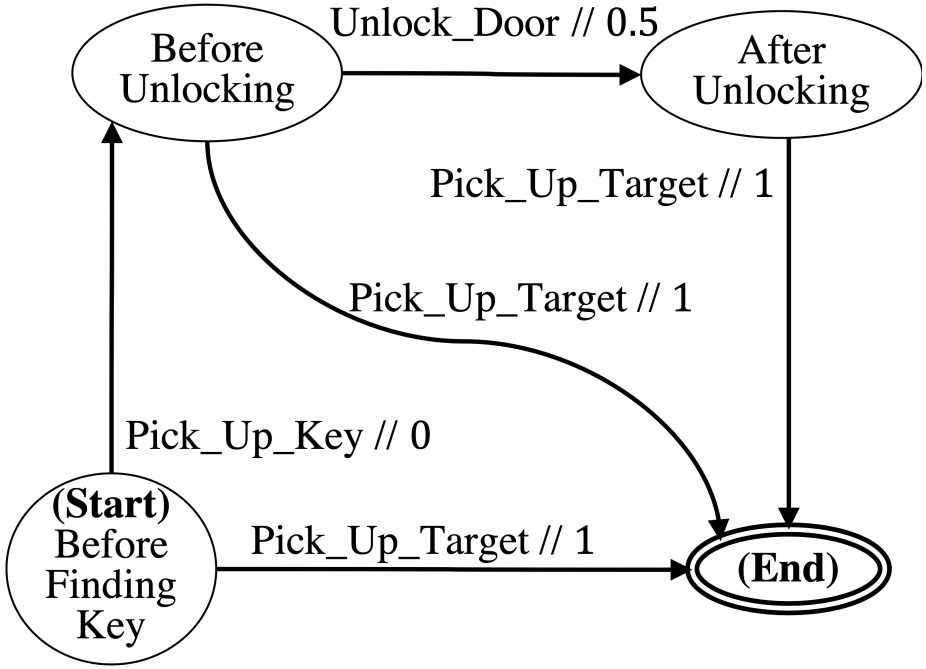
- **Goal driven reward**
  - RL algorithms fail to solve the simple task
- **Intrinsic Reward**
  - Impact-driven [Raileanu, R. et al. 2020] and exploration-driven [Flet-Berliac, Yannis et al. 2021] (SOTA) cannot solve the task efficiently

BU

Reward Machine (RM)[1]



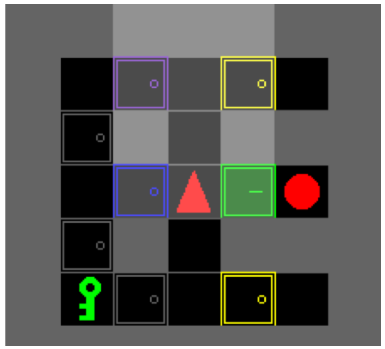


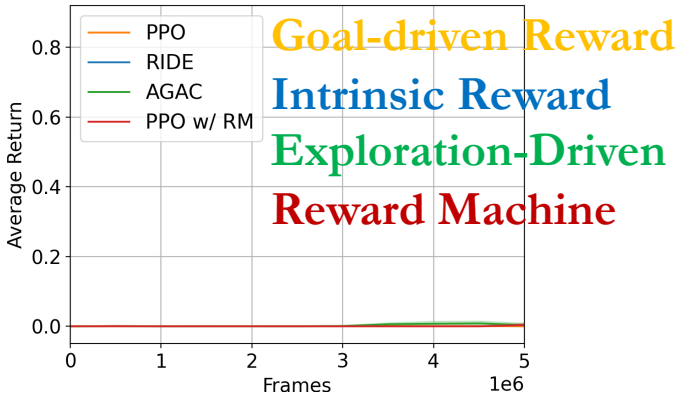


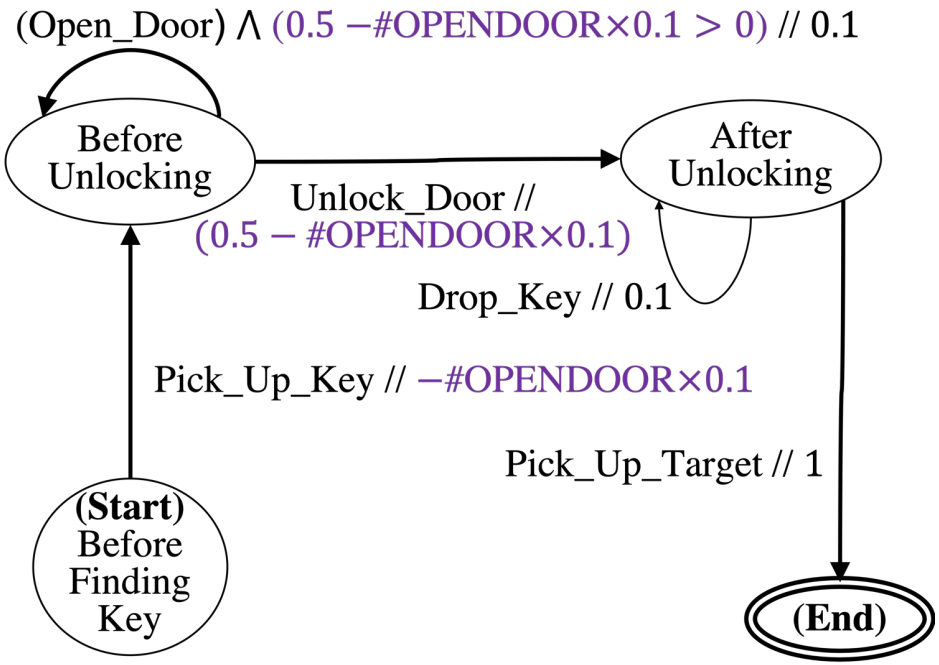
BU

1. Icarte, R. T., Klassen, T. Q., Valenzano, R., and McIlraith, S. A. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

Symbolic Reward Machine (SRM)

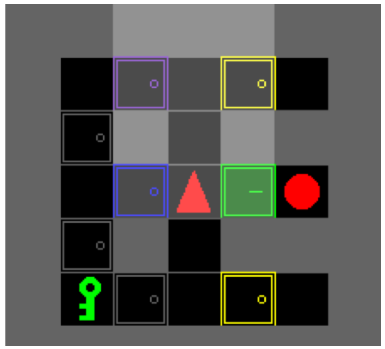


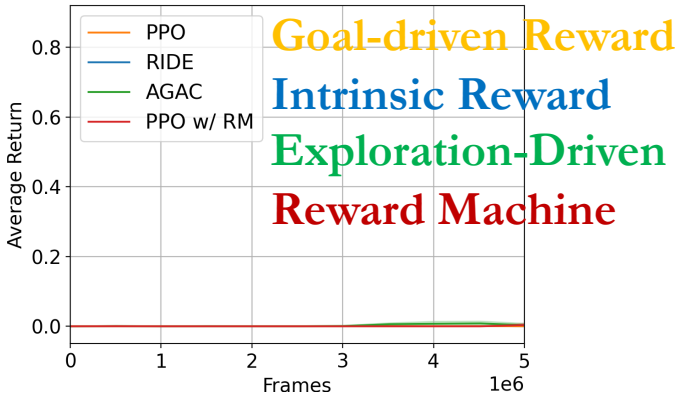


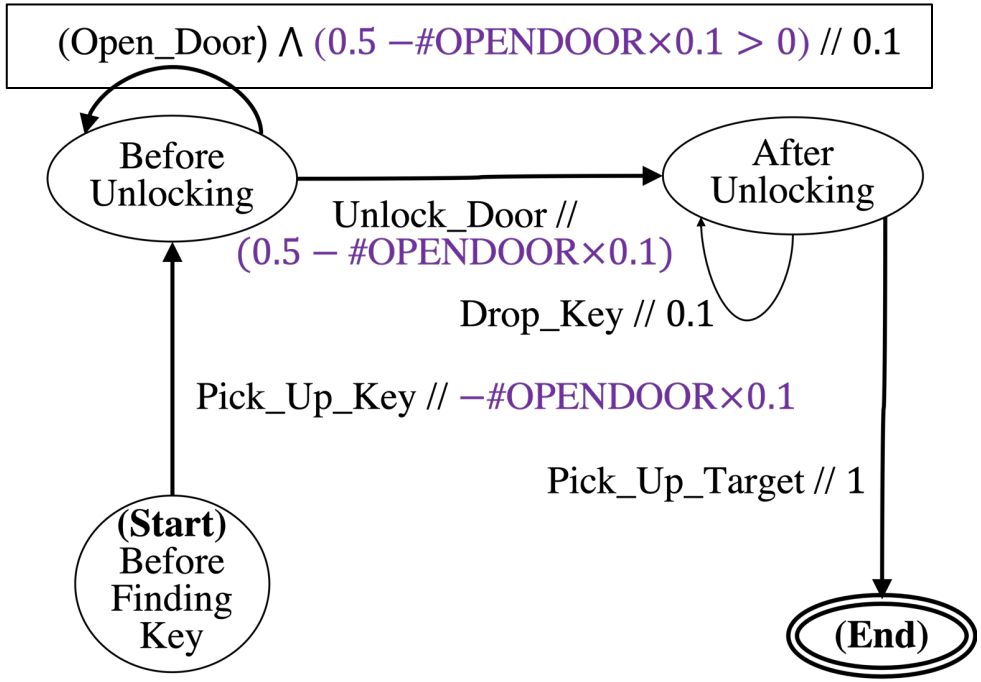


- Output rewards represented as functions
- Leverage the history experience

Symbolic Reward Machine (SRM)



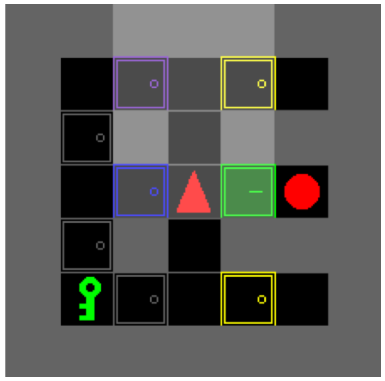


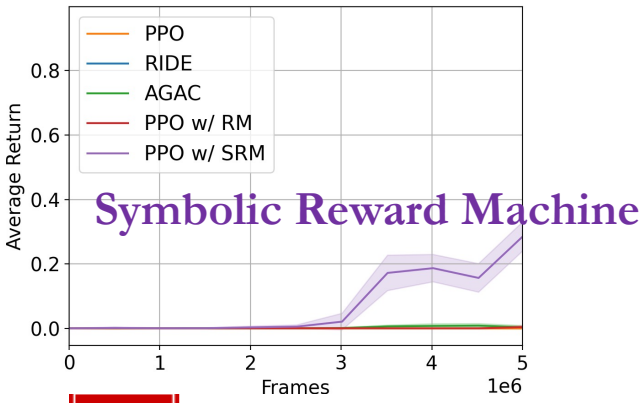


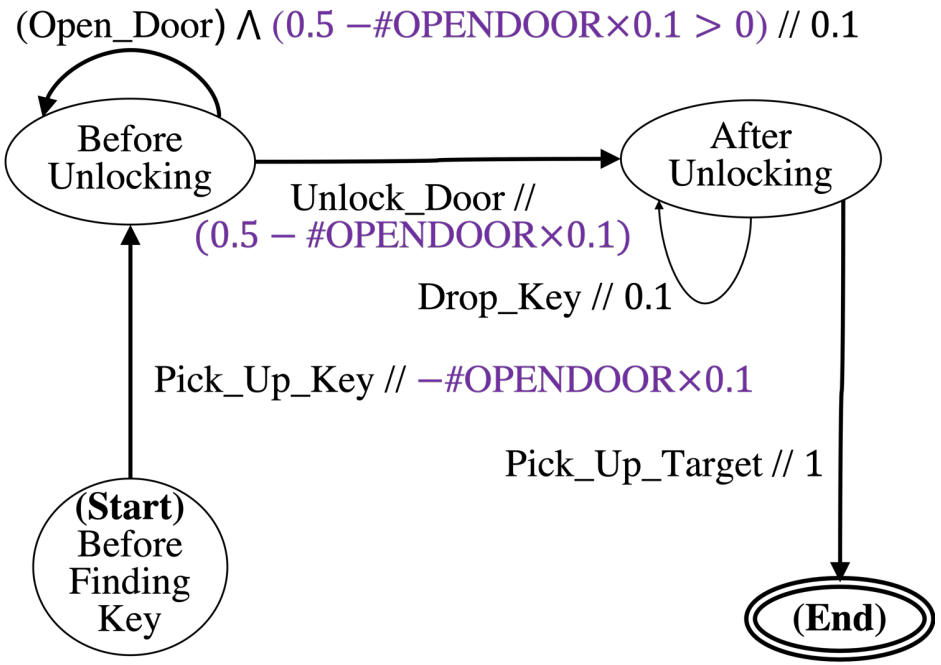
- Output rewards represented as functions
- Leverage the history experience

BU

Symbolic Reward Machine (SRM)



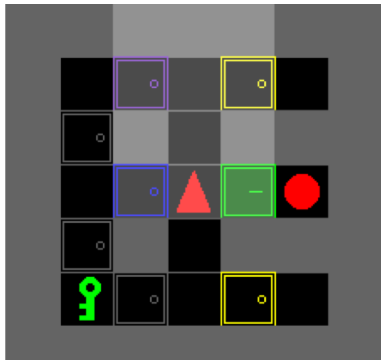


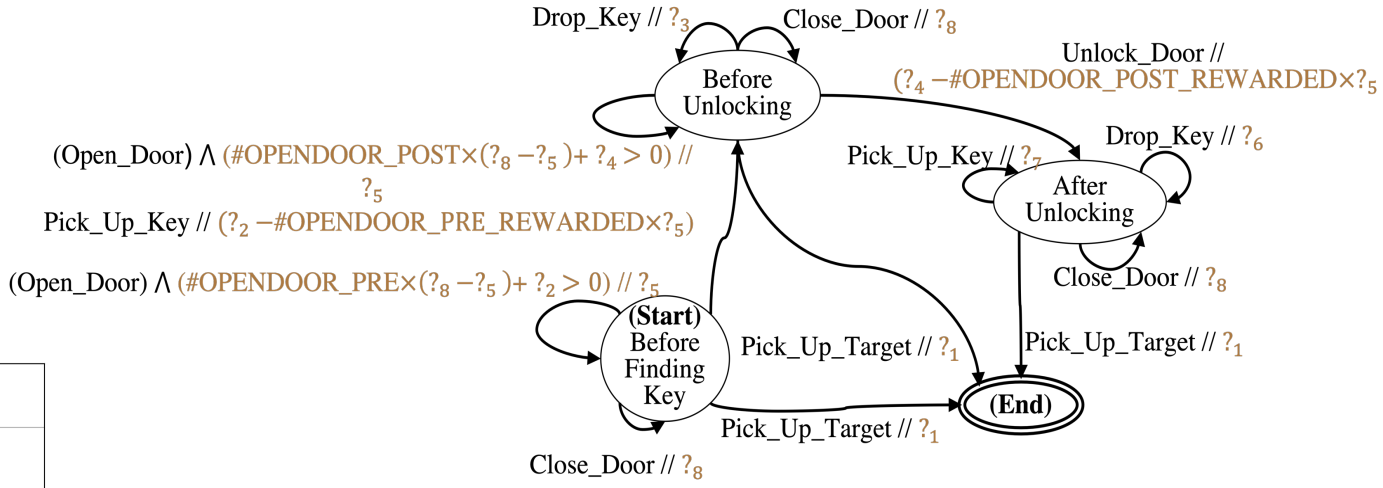


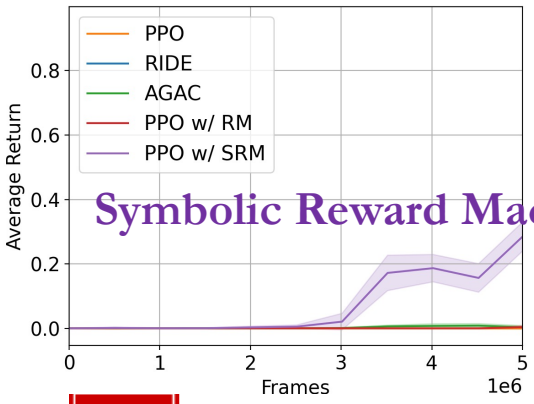
- Have to manually assigned the numerical terms in the rewards

Symbolic Reward Machine (SRM)

- Design SRM with unknown variables $?_1, ?_2, \ldots$
- Design variable constrains, e. g. $\bigwedge_{id=2}^{8} ?_{id} \leq ?_1$



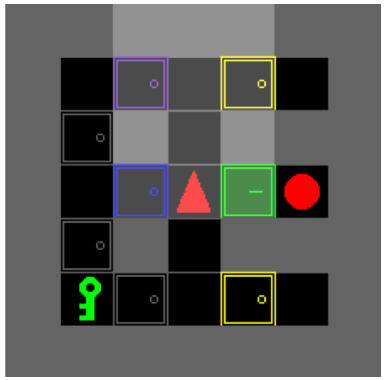


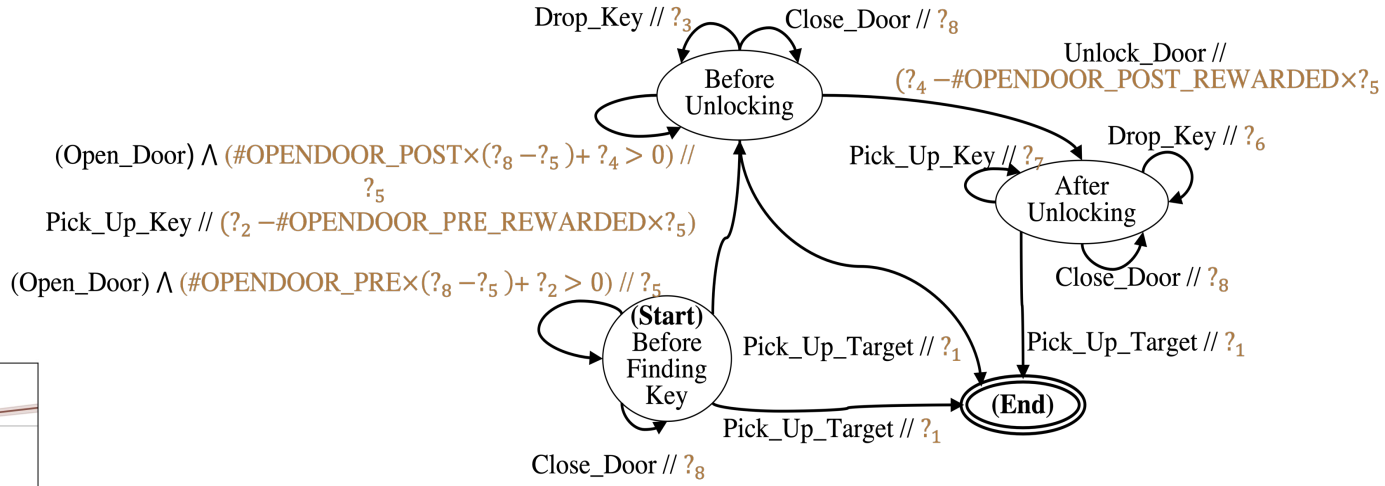


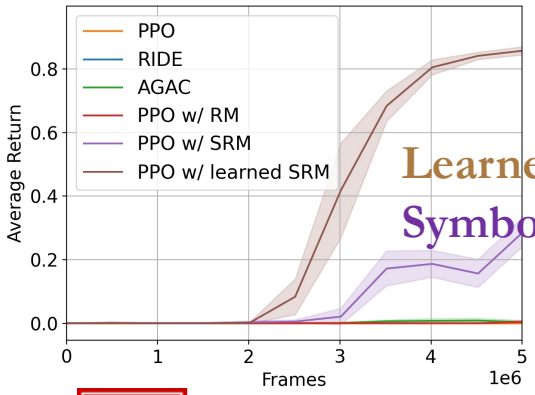
- **Concretization**: learn $?_1, ?_2, \ldots$ from human demonstrations

Symbolic Reward Machine (SRM)

- Design SRM with unknown variables $?_1, ?_2, \ldots$
- Design variable constrains, e. g. $\bigwedge_{id=2}^{8} ?_{id} \leq ?_1$







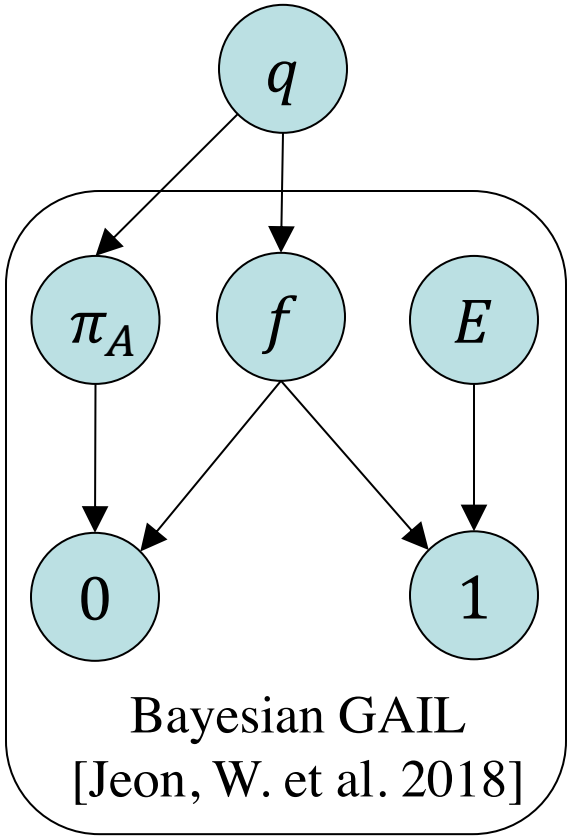
Learned Symbolic Reward Machine

Symbolic Reward Machine

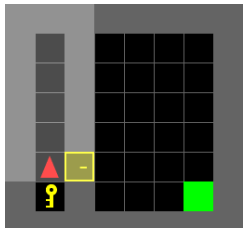
- **Concretization**: learn $?_1, ?_2, \ldots$ from human demonstrations

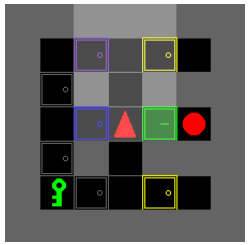
A Hierarchical Bayesian Approach

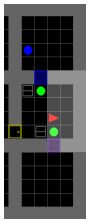
- Input:
  - Human demonstrations E
  - An SRM with unknown variables $?_1, ?_2, \ldots$
- Initialization:
  - A distribution q of possible values of $?_1, ?_2, \ldots$
  - A latent reward function f conditioned on SRM's outputs
  - An agent policy π_A
- Iteration:
  1. Train f to discriminate E from the trajectories of π_A
  2. Optimize q to match f
  3. Train π_A with most likely symbolic reward function w.r.t q
- Output:
  - Trained agent policy π_A
  - Distribution q concentrating on the optimal values of $?_1, ?_2, \ldots$

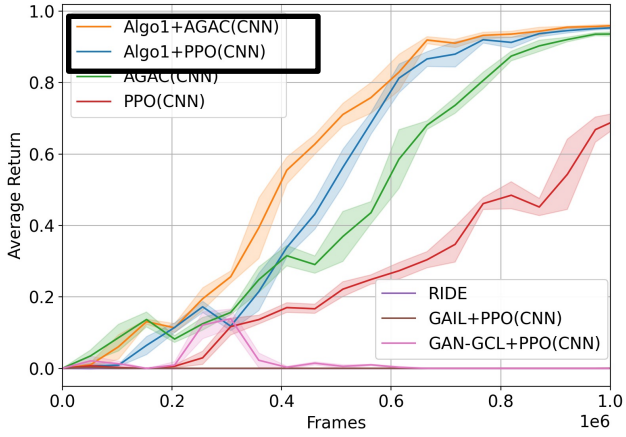


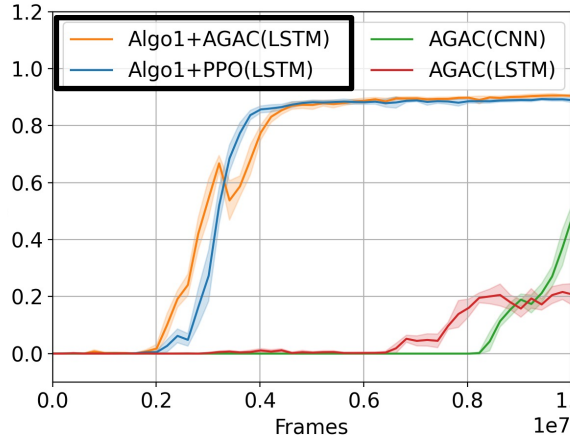
Results: evaluate agent policy π_A

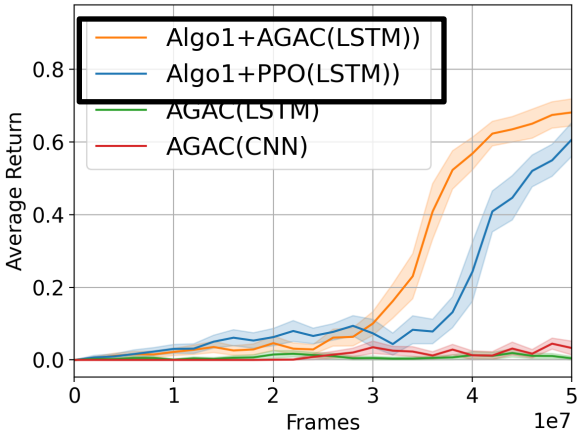








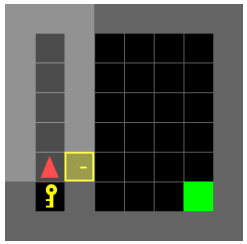




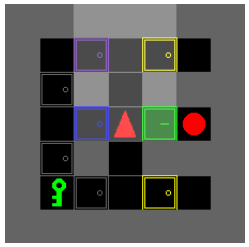
8x8 DoorKey 7x7 KeyCorridor Obstructedmaze-2Dlhb

- **GAIL, GAN-GCL fail while our algorithm (Algo1) performs well in all 3 tasks**
- **SRMs are relatively sparse while baselines use/generate dense rewards**

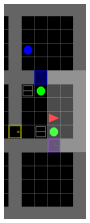
Results: evaluate learned SRMs



8x8 DoorKey

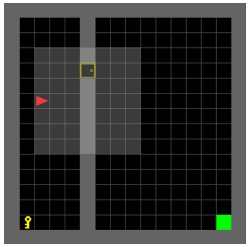


7x7 KeyCorridor

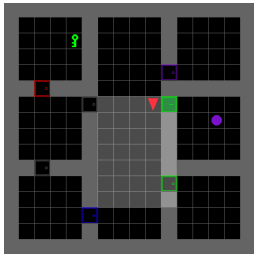


Obstructedmaze-2Dlhb

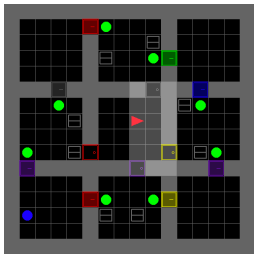
Learned SRMs in small environments
Train RL agents in larger environments



16x16 DoorKey

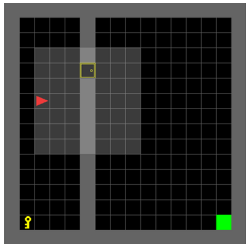


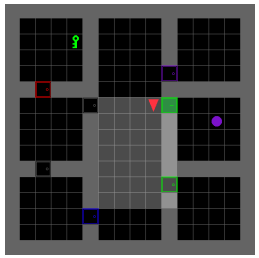
16x16 KeyCorridor

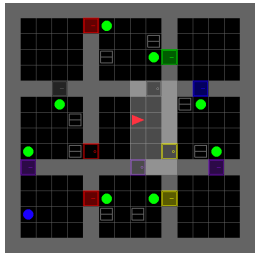


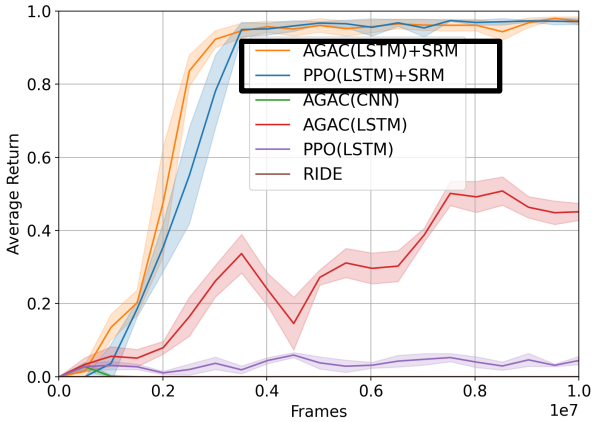
Obstructedmaze-Full

Results: evaluate learned SRMs

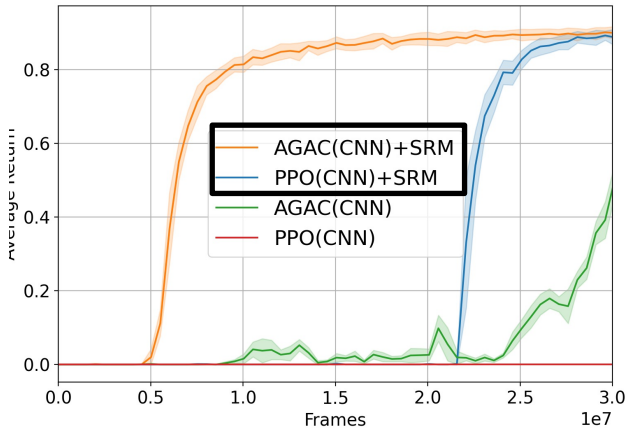




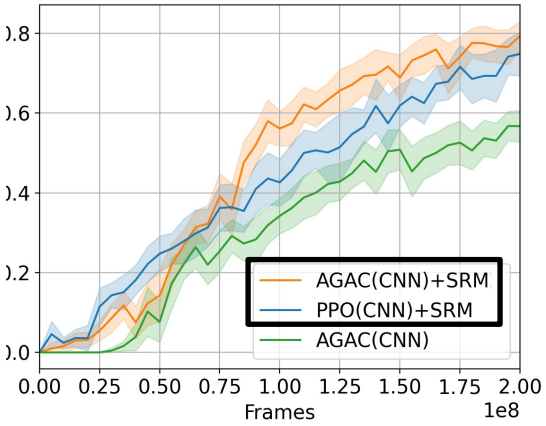




16x16 DoorKey



16x16 KeyCorridor



Obstructedmaze-Full

- **The learned SRMs generalize well in all 3 tasks**

Takeaways

- Propose **Symbolic Reward Machine (SRM)**, a structured reward function
- Propose an algorithm that concretizes SRMs by **learning from expert demonstrations**.
- Our algorithm achieves better RL policy **training efficiency** in challenging benchmarks
- Our algorithm learns SRMs that are **generalizable** in differently configured environments