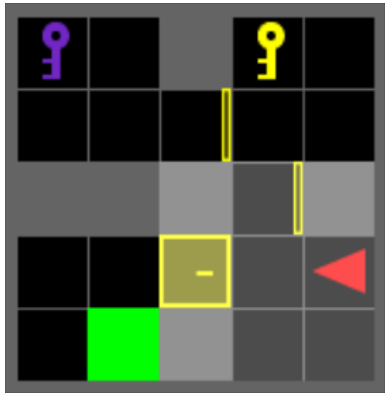
Partially Observable Hierarchical Reinforcement Learning with AI Planning

Brandon Rozek+, Junkyu Lee*, Harsha Kokel*, Michael Katz*, Shirin Sohrabi*

Guide reinforcement learning agents with AI Planning under uncertainty by encoding how an agent might *discover* unknown information.

Example Domain: MiniGrid



Goal:
Get to green square.

Find key (🔑) to

unlock door (▨)

How? *Discover keys by moving to new rooms!*

Approach

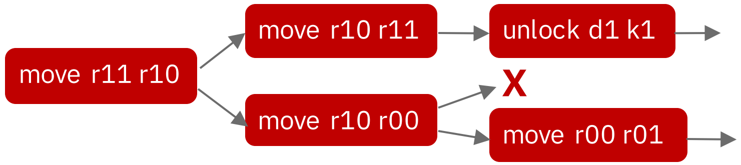
Model discovery as Non-Deterministic Effect:

```
(:action move-room
  :parameters (?d - door ?r1 - room ?r2 - room)
  :precondition (and (at-agent ?r1) (unlocked ?d)
(CONNECTED-ROOMS ?r1 ?r2) (LINK ?d ?r1 ?r2) )

:effect (and
   (not (at-agent ?r1)) (at-agent ?r2)
   (forall (?k -key)

   (when (not (entered-room ?r2))          (1)

   (when (not (discovered ?k))             (2)

   (oneof
    ; Yellow Key Present
    (and (at ?k ?r2) (color ?k yellow) (discovered ?k)
(entered-room ?r2))
    ; Purple Key Present
    (and (at ?k ?r2) (color ?k purple) (discovered ?k)
(entered-room ?r2))
    ; Key not present
    (entered-room ?r2)
))))))
```

Constraints on discovery effects needed to:

1. Prevent cyclic policies
2. Avoid inconsistent policies

Generate High-Level Policy from FOND AI Planner:



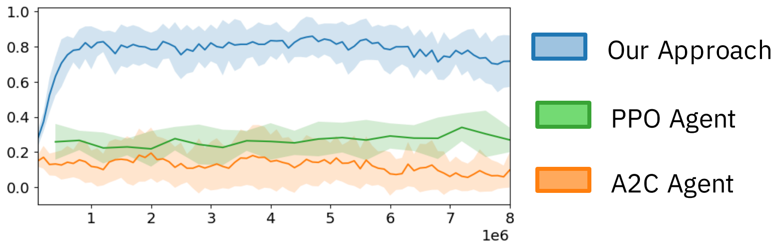
Train a RL-PPO Agent on each High-level Action:

Considerations

1. Penalize agent for deviating but not for discovering.
2. If no high-level policy is found, perform exploration.

Empirical Evaluation

Success Rate over Number of Training Samples



+ Rensselaer

* IBM Research