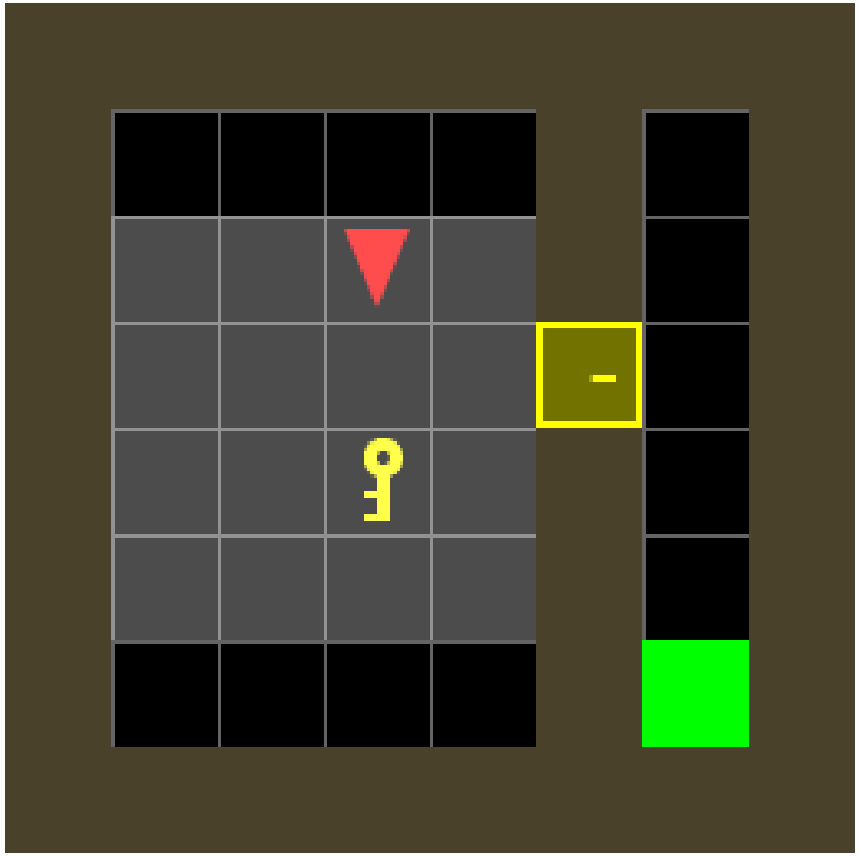
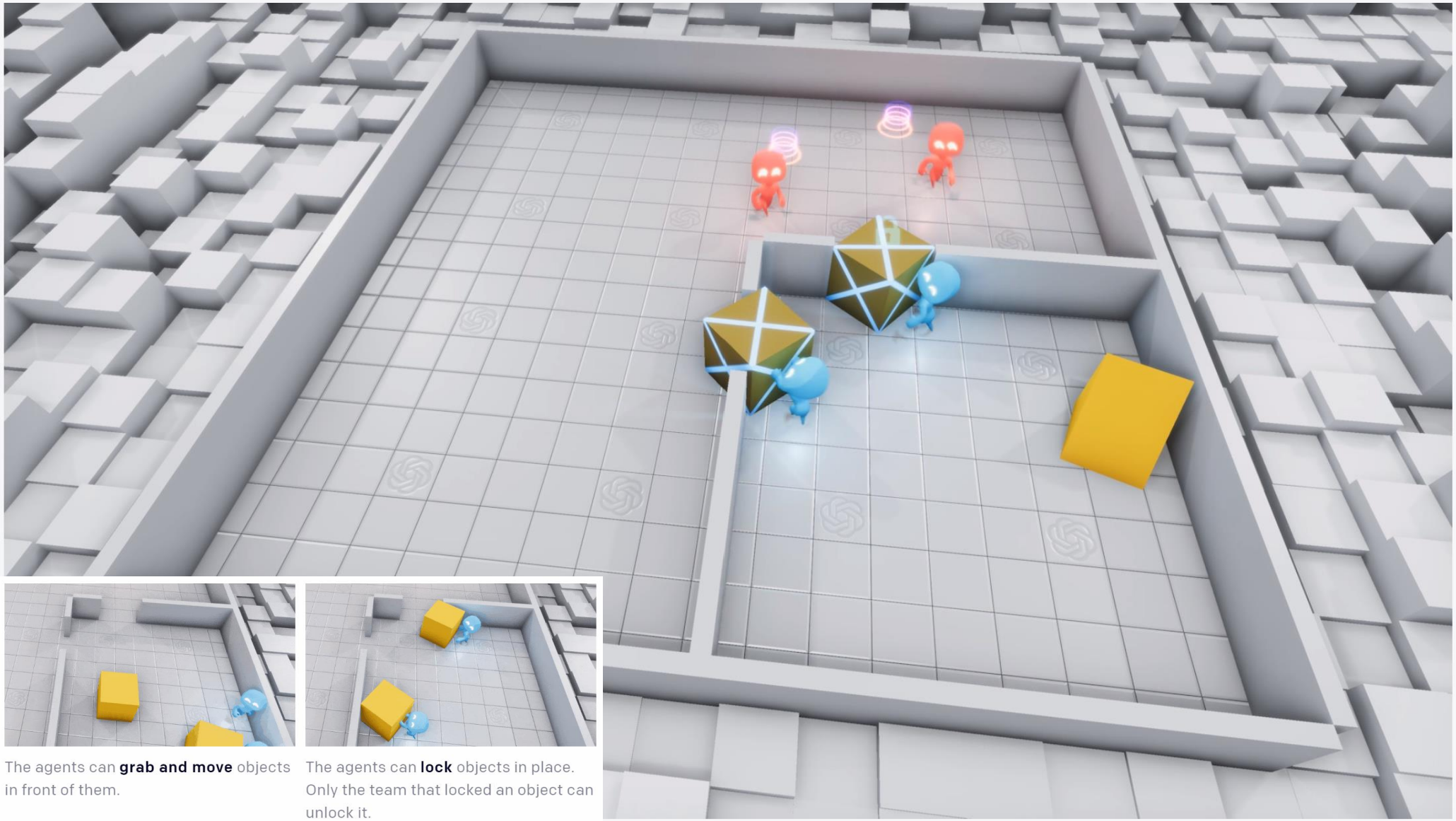


Multi Agent Reinforcement Learning

Yseult Héjja-Brichard

TA: Yen-Ling Kuo

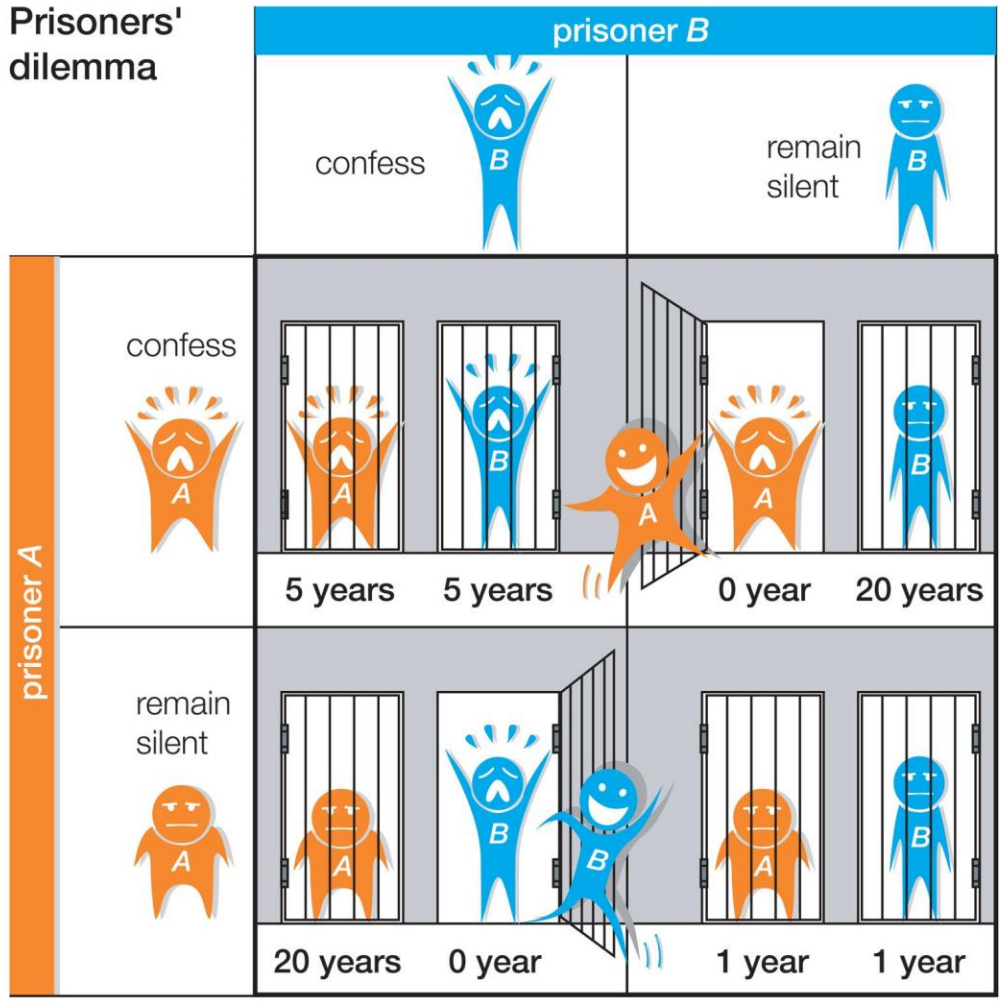


The agents can **grab and move** objects in front of them.

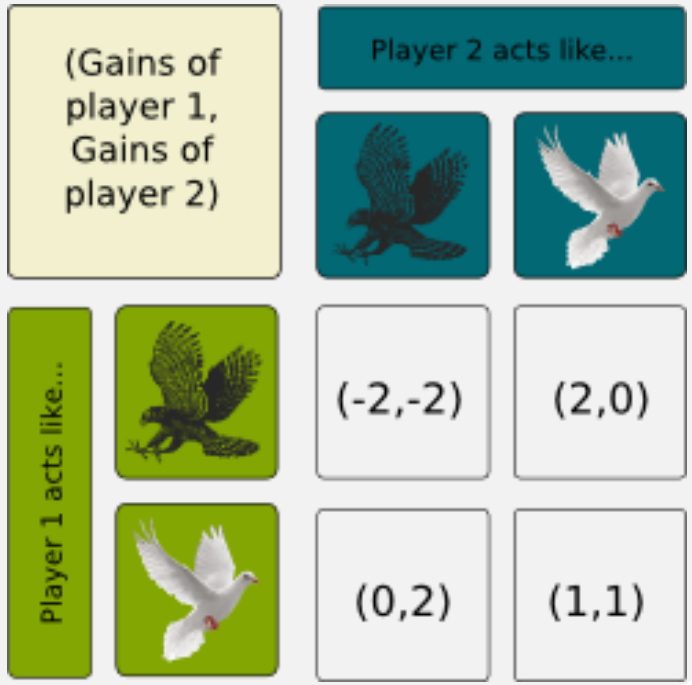
The agents can **lock** objects in place. Only the team that locked an object can unlock it.

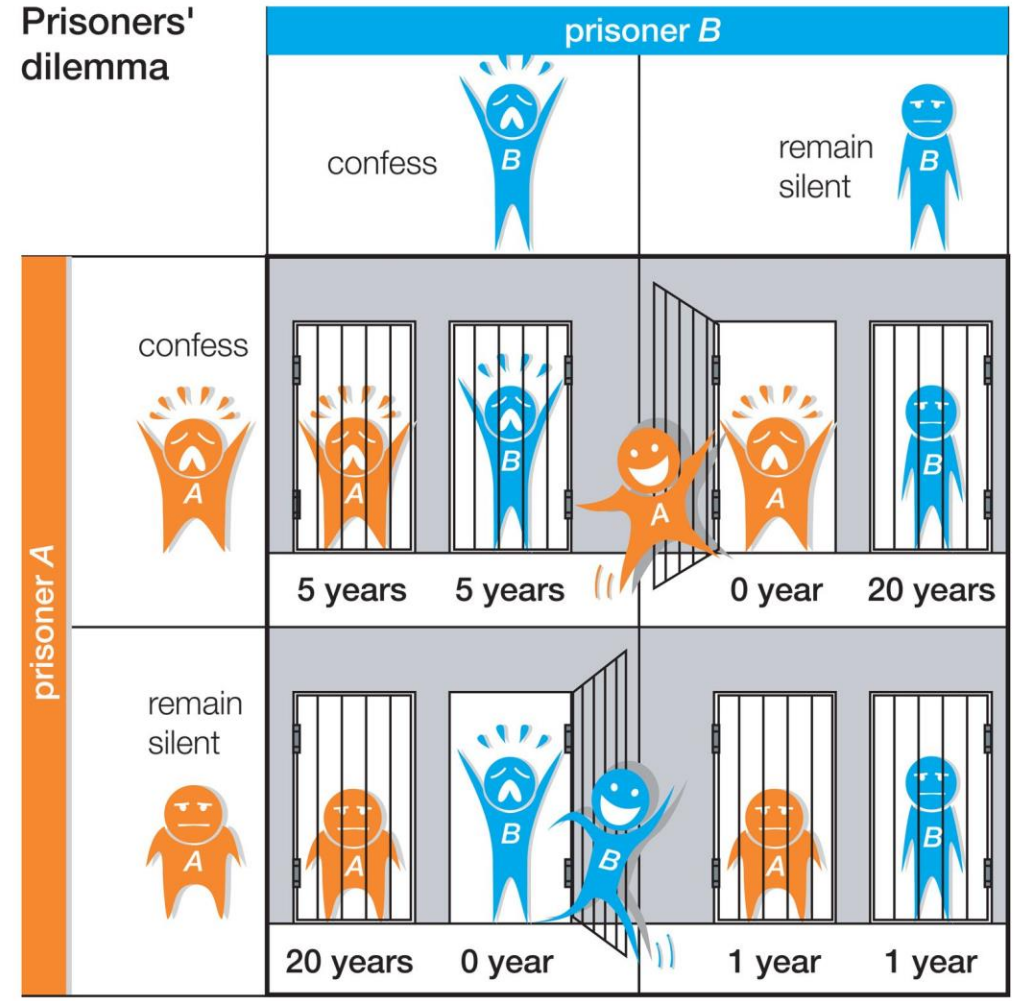
Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020). Emergent Tool Use From Multi-Agent Autocurricula (arXiv:1909.07528). arXiv.

(Evolutionary) Game Theory



Evolutionary Game Theory



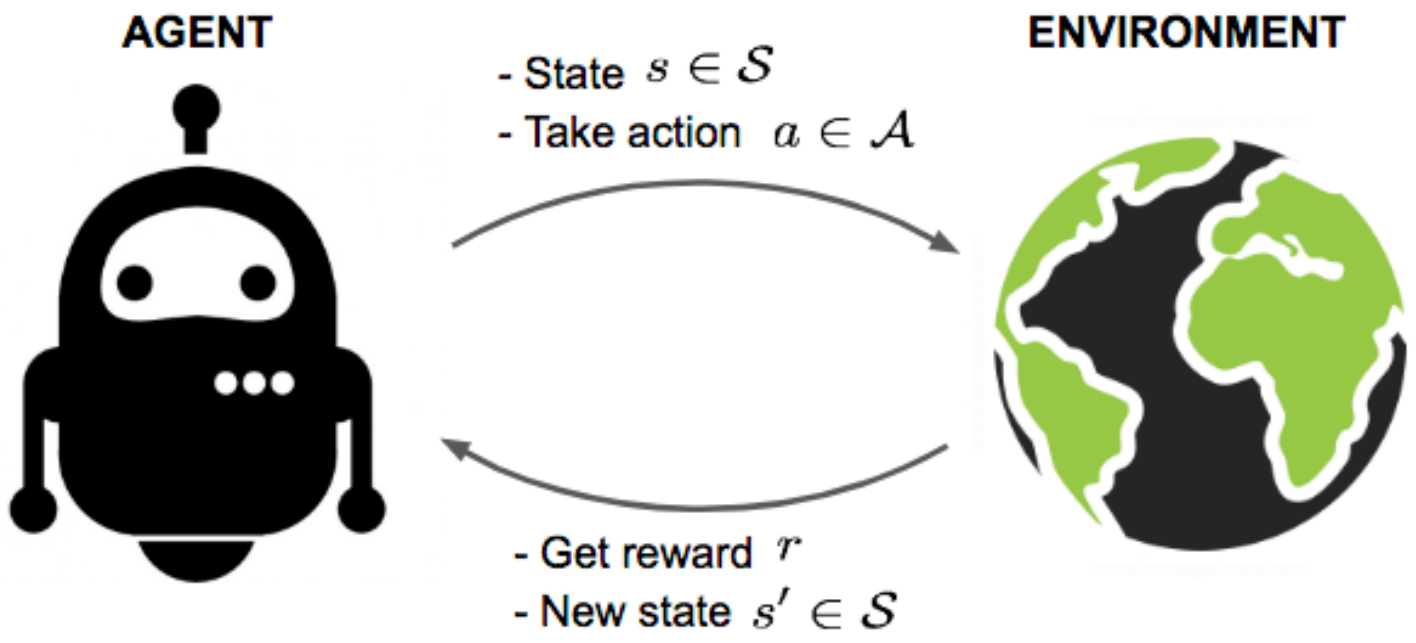


Practical motivations

Learning how to implement RL in a multi-agent configuration

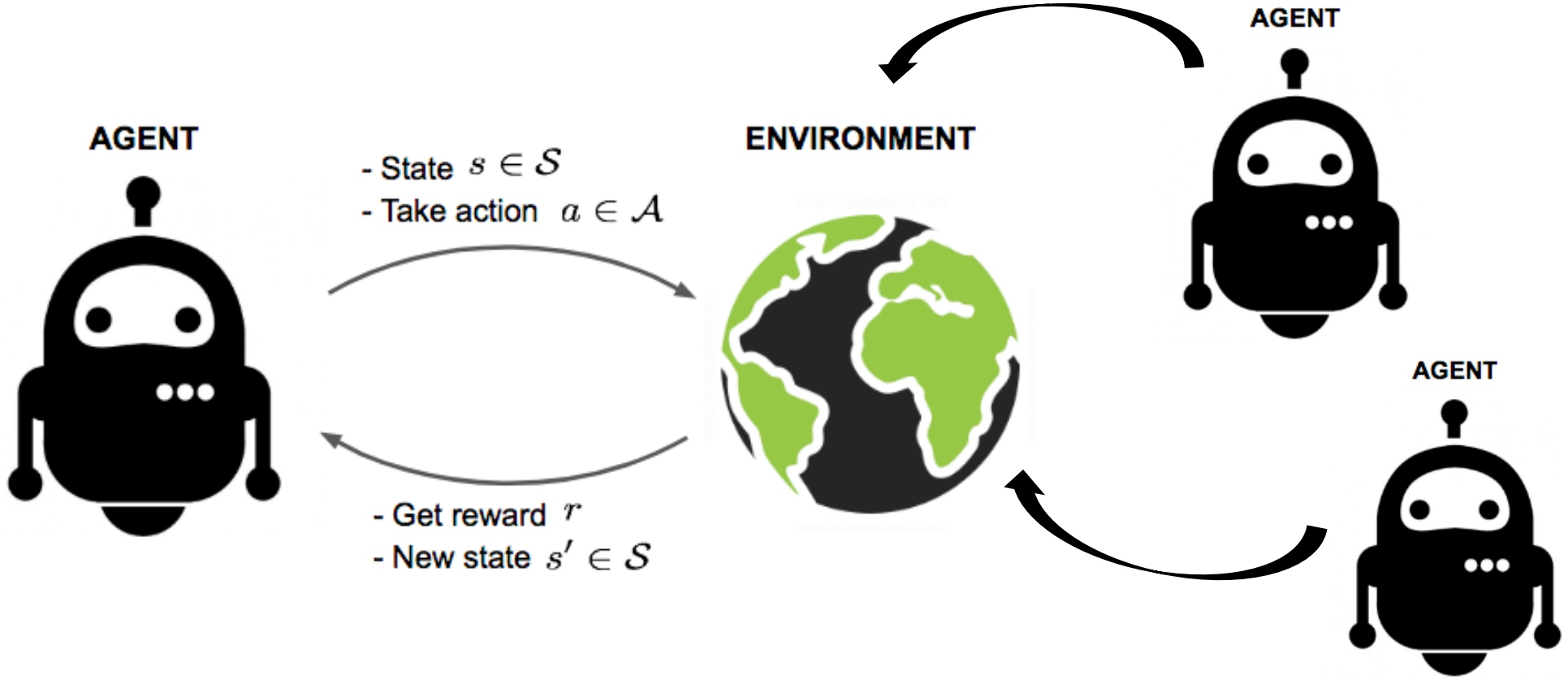
Interest in how to model different types of social interactions

Reinforcement Learning



Markov Decision Process (MDP)

Reinforcement Learning with multiple agents



(Partially Observable) Markov Decision Process (POMDP)

Interacting with other agents

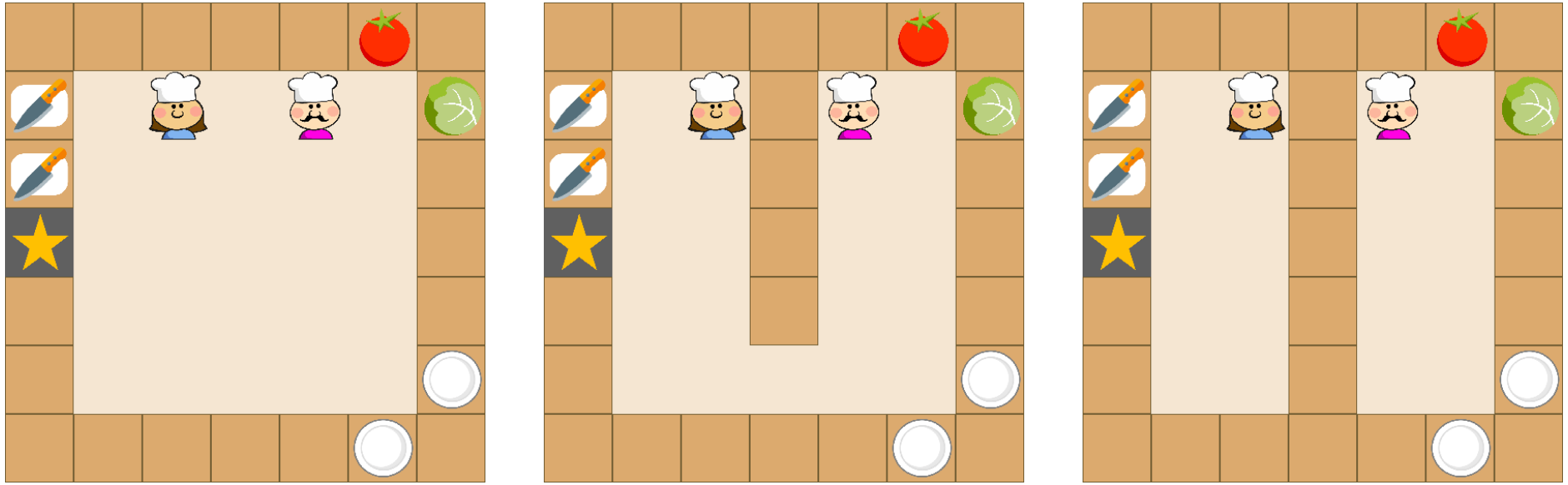
Strategies: ToM-based models for goal inference, Bayesian inverse planning, learning the reward functions of other agents, imitation

LeCTR (Dec-POMDP) *Omidshafiei et al., 2019* - Learning to teach in coop MARL

MA-POMDP *Ndousse et al., 2021* - Emergent social learning via MARL

Social MDP *Tejwani, Kuo et al., 2022* - Extended social MDP

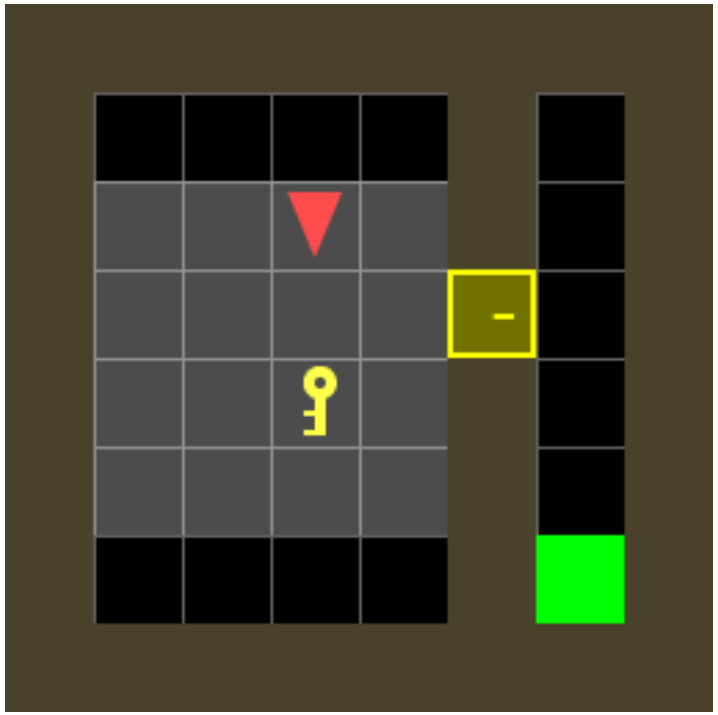
How it started: Cute env but hard to decipher



Wang, R. et al. (2020) Too many cooks: Bayesian inference for coordinating multi-agent collaboration

Bayesian Delegation: inverse planning and inference on subtasks

Where the project landed

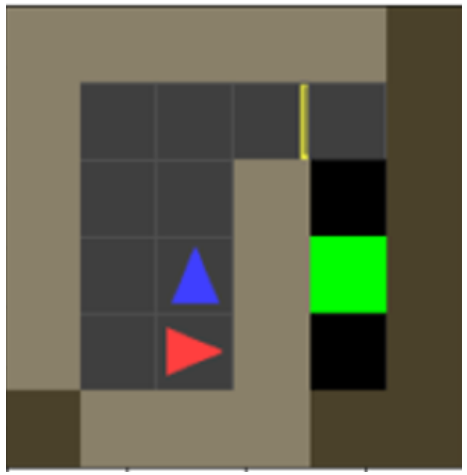


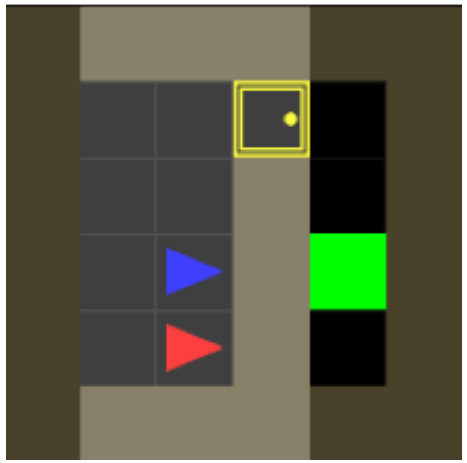
Environment: MarlGrid (Kamal Ndousse), a multi-agent variant of MiniGrid (Gym)

Grid size: 6x6 with a wall

Objects: Open door, closed door (toggle but no key needed)

Where the project landed





Environment: MarlGrid (Kamal Ndousse), a multi-agent variant of MiniGrid (Gym)

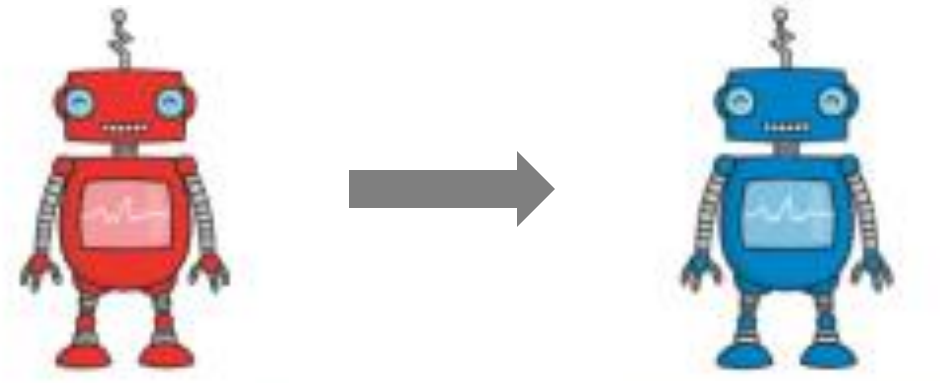
Grid size: 6x6 with a wall

Objects: Open door, closed door (toggle but no key needed)

State = agent[1] * 6 * 4 * 2 + agent[0] * 4 * 2 + agent_dir * 2 + door_state

Agents

- 2 agents: a teacher and a student
- Learning algorithm: Q-learning > Teacher is greedy (optimal)
- Student: behavioral cloning



Teacher Student

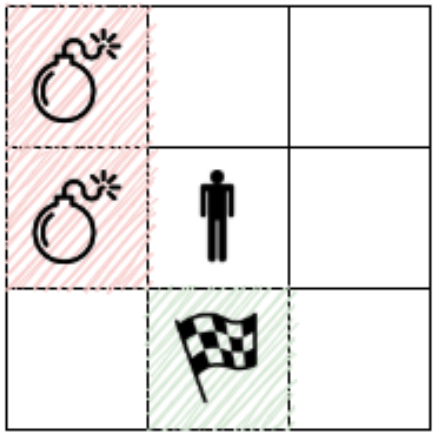
Raymond's rule:

```
if action_student != action_teacher:
     reward = -1
else:
     reward = 1
```

Teacher's Q-Learning

Teacher

State



Action Values

NW -50	N 0	NE 0
W -50	-	E 0
SW 0	S 100	SE 0

Actions

	A_W	A_{NW}	A_N	A_{NE}	A_E	A_{SE}	A_S	A_{SW}
S_1	-50	-50	0	0	0	0	100	0

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

Simulations

Level 1a: Open door and no key

Level 1b: Open door at a new position

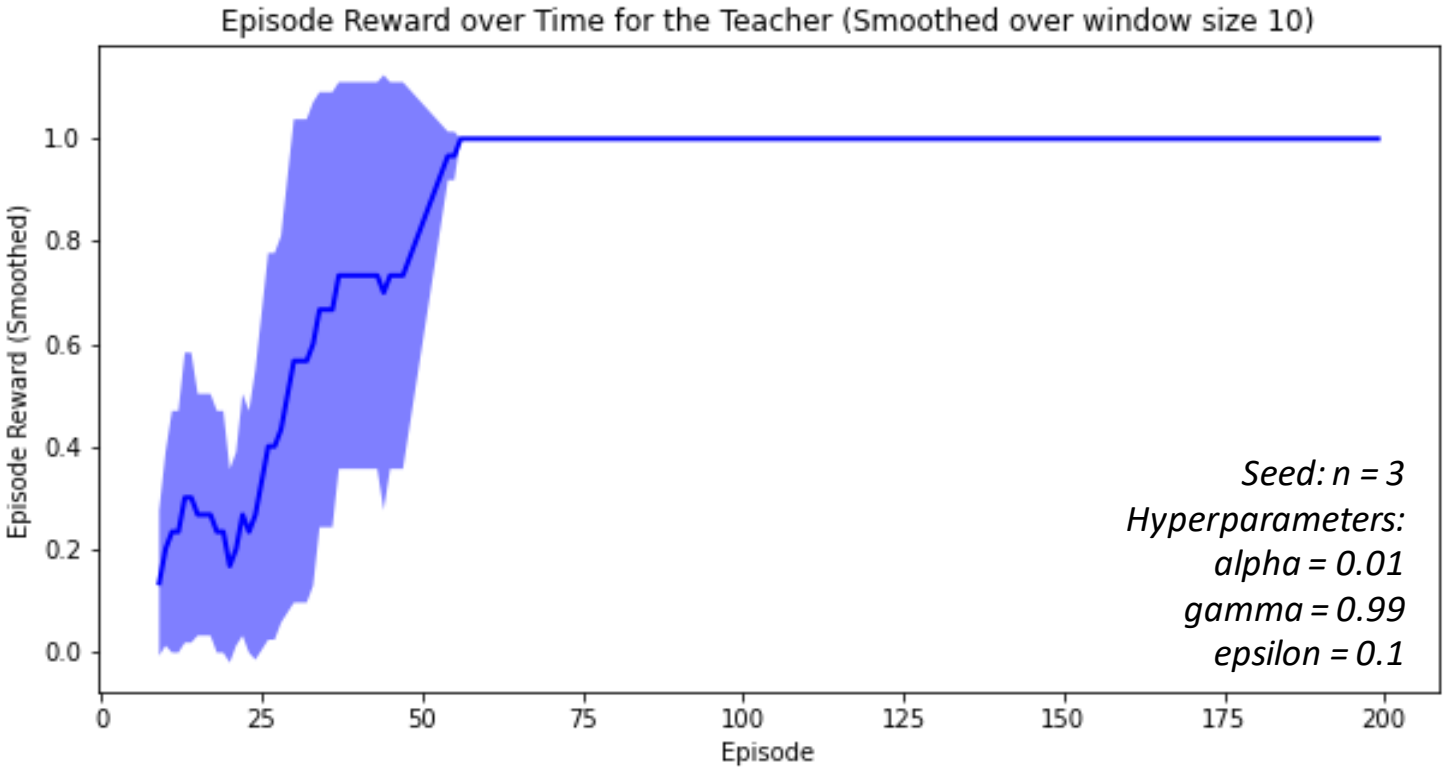
Level 2a: Closed door and no key

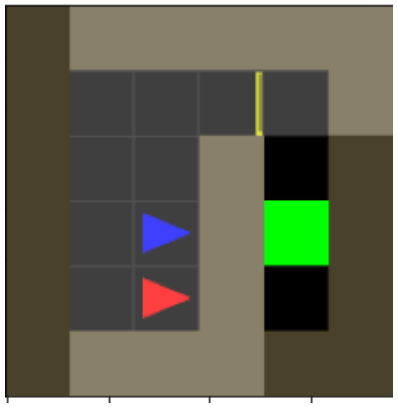
Level 2b: Sparse learning

Level 3: Locked door + key

Environment complexity: Level 1 (open door)

Teacher



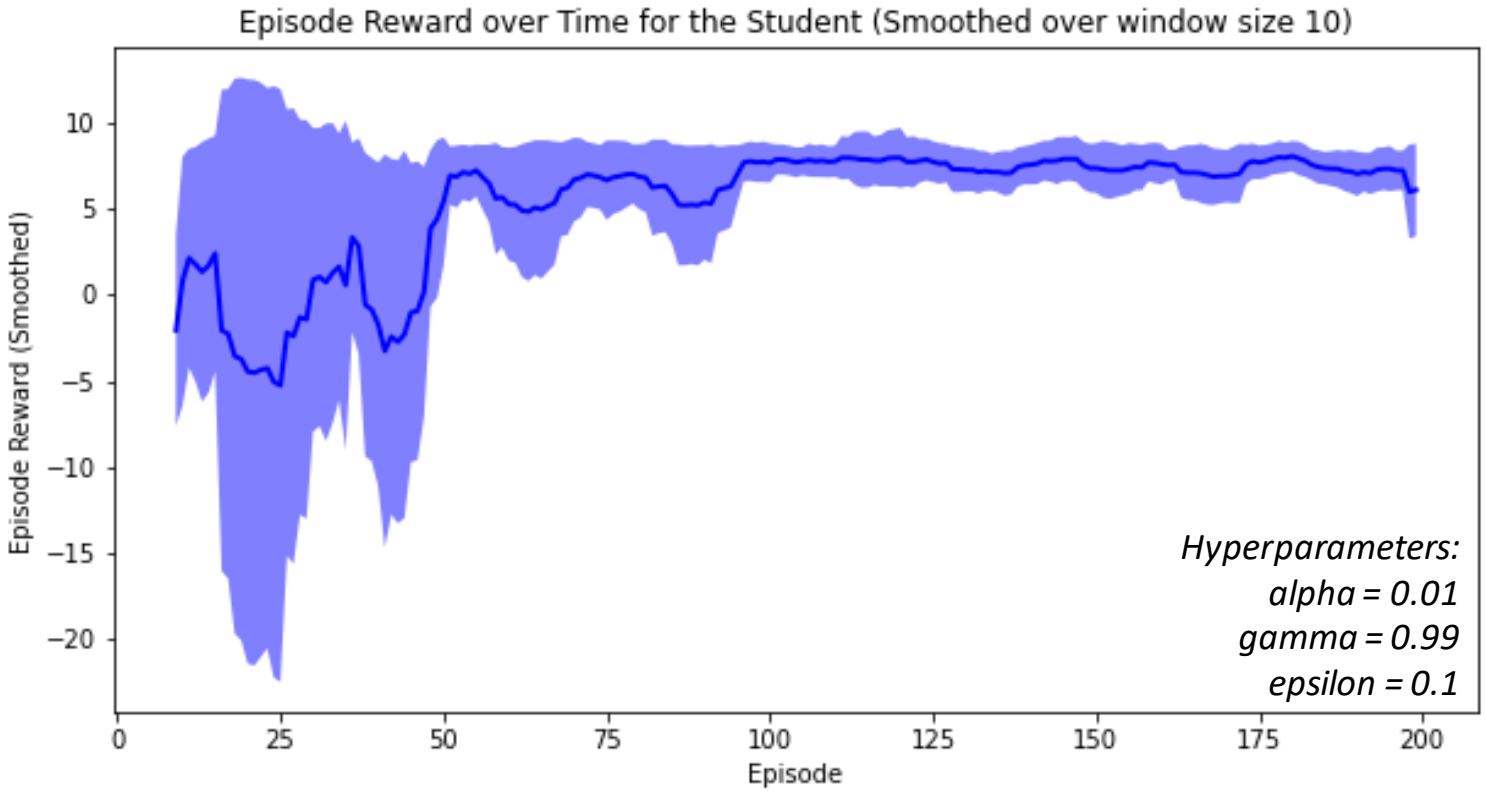


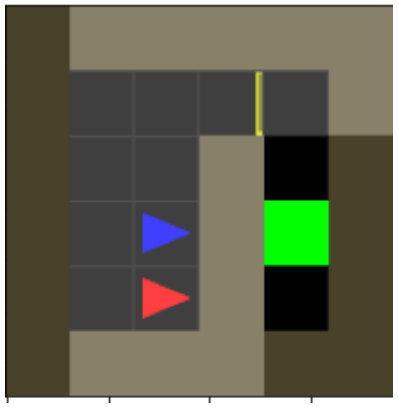
Training: 200 episodes
Test: 100 episodes

Results after 100 episodes:
Average time steps per episode: 9.3
Average reward per episode: 1.0

Env complexity: Level 1a (open door)

Student



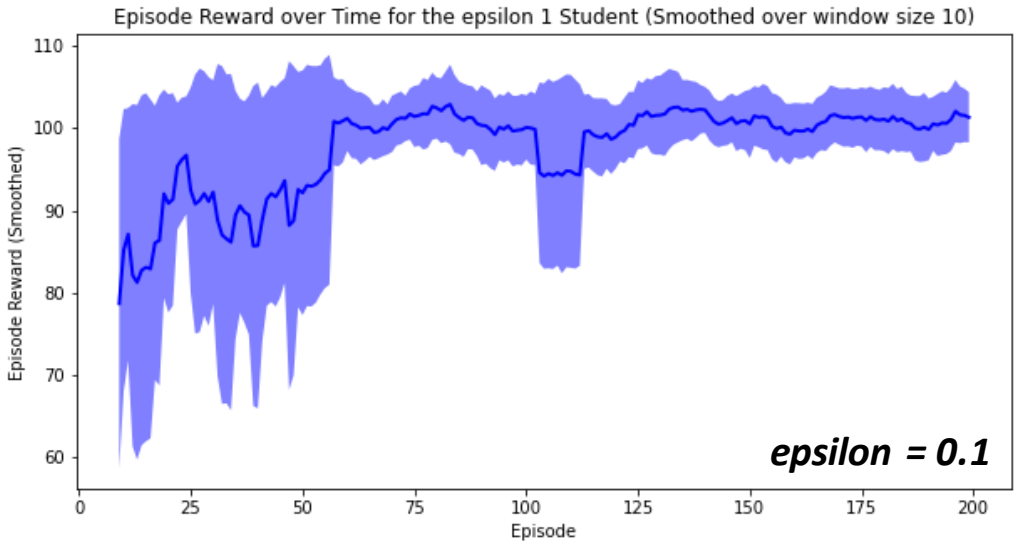


Training: 200 episodes
Test: 100 episodes

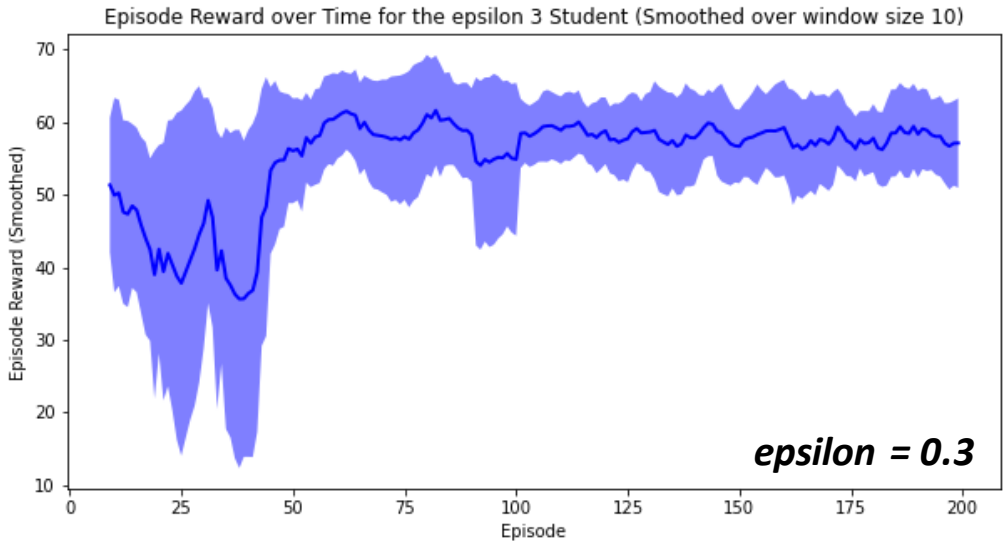
Results after 100 episodes:
Average time steps per episode: 9.3
Average reward per episode: 1.0

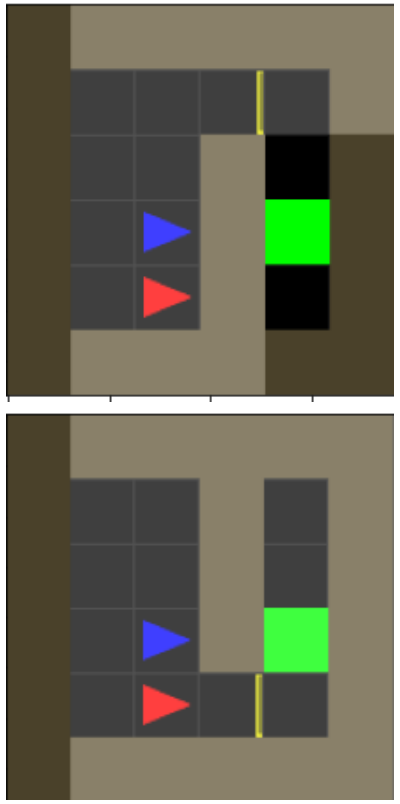
Env complexity: Level 1b (open door, diff loc)

Student 1



Student 2



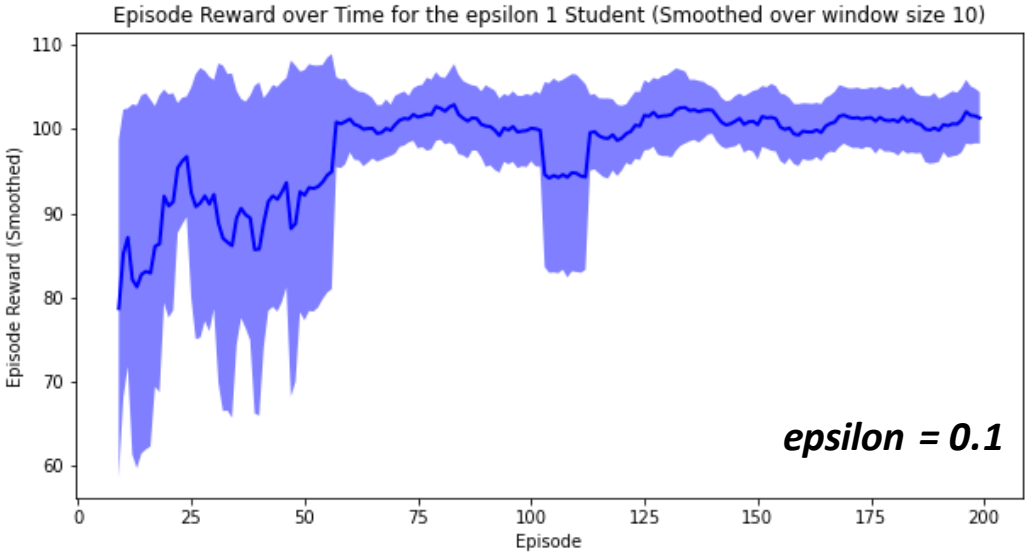


Training: 200 episodes
Test: 100 episodes

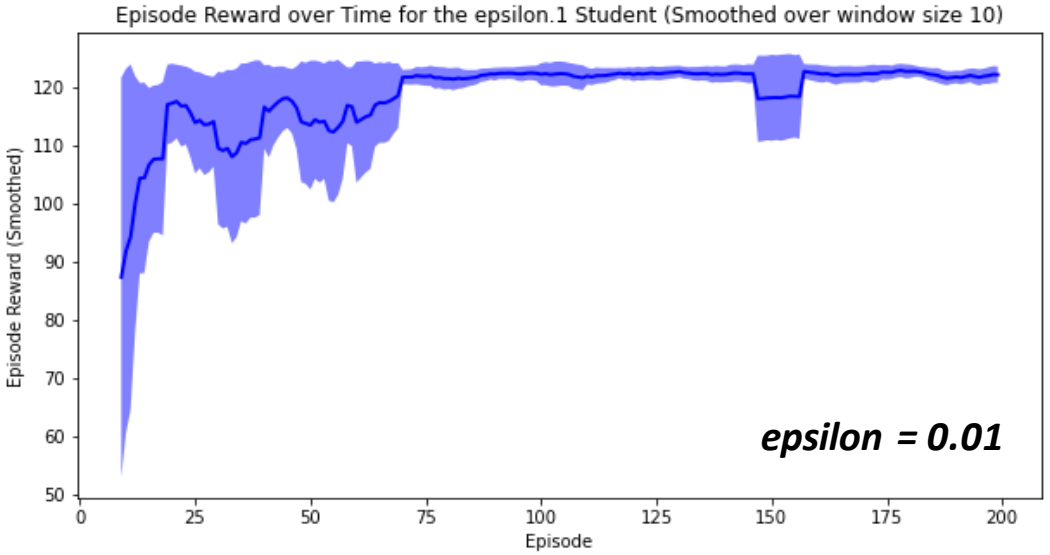
Results after 100 episodes:
Average time steps per episode: 10.0
Average reward per episode: 1.0

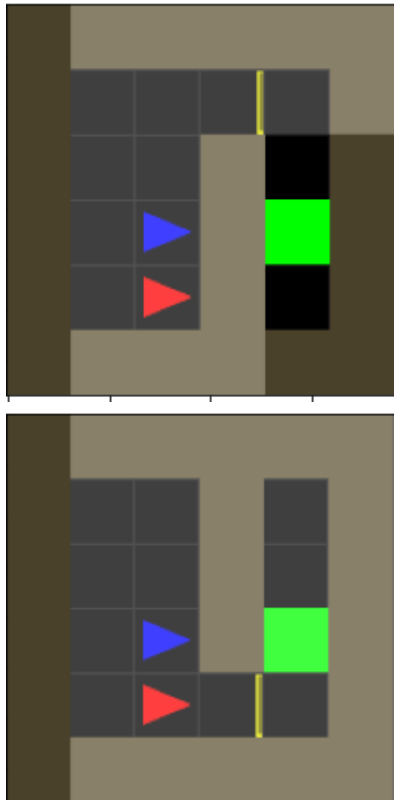
Env complexity: Level 1b (open door, diff loc)

Student 1



Student 3



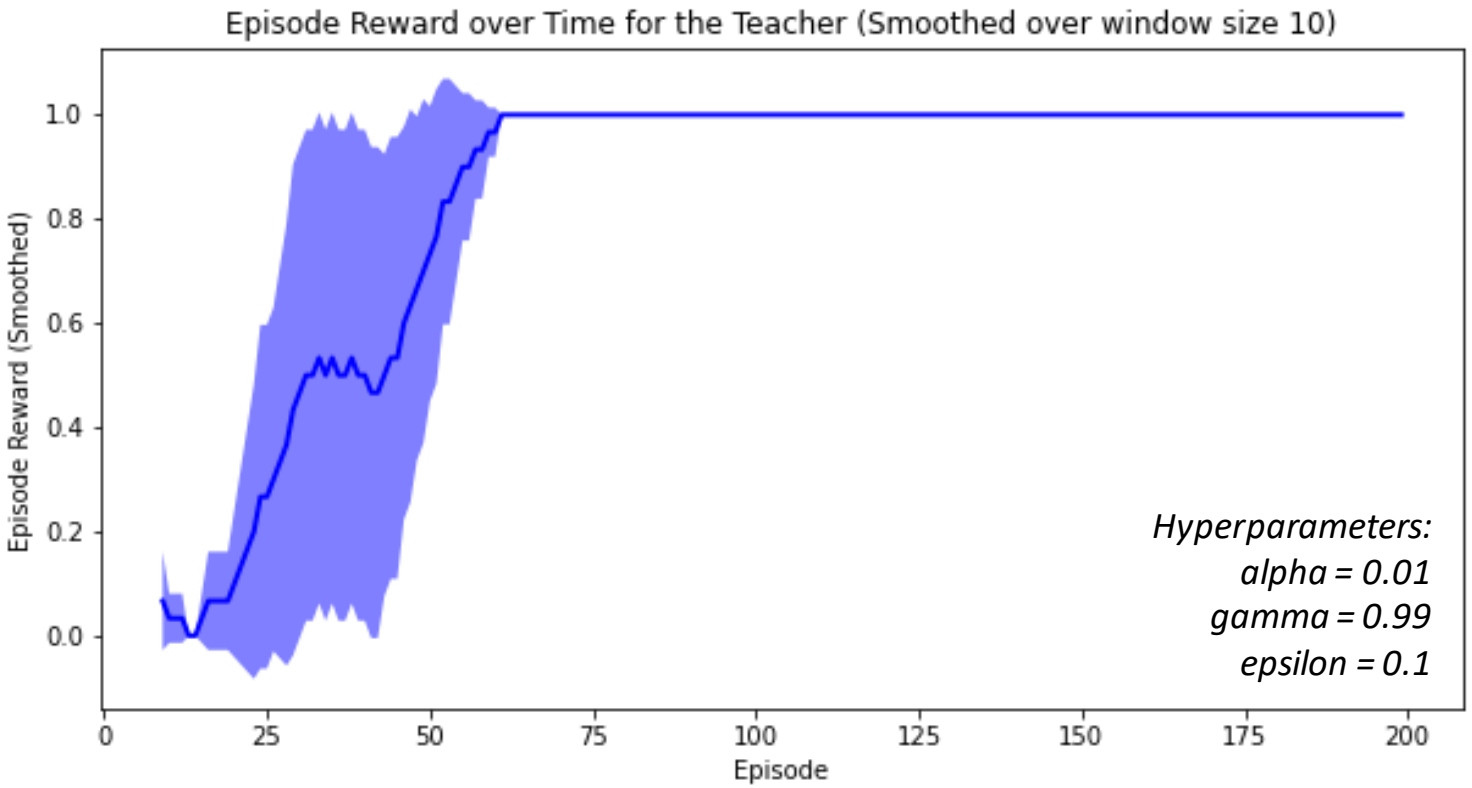


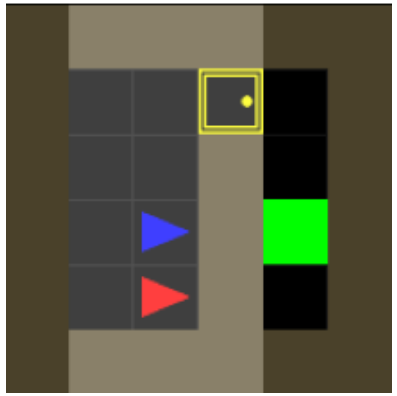
Results after 100 episodes:
Average time steps per episode: 10.0
Average reward per episode: 1.0

Results after 100 episodes:
Average time steps per episode: 250.0
Average reward per episode: 0.0

Env complexity: Level 2 (closed door)

Teacher



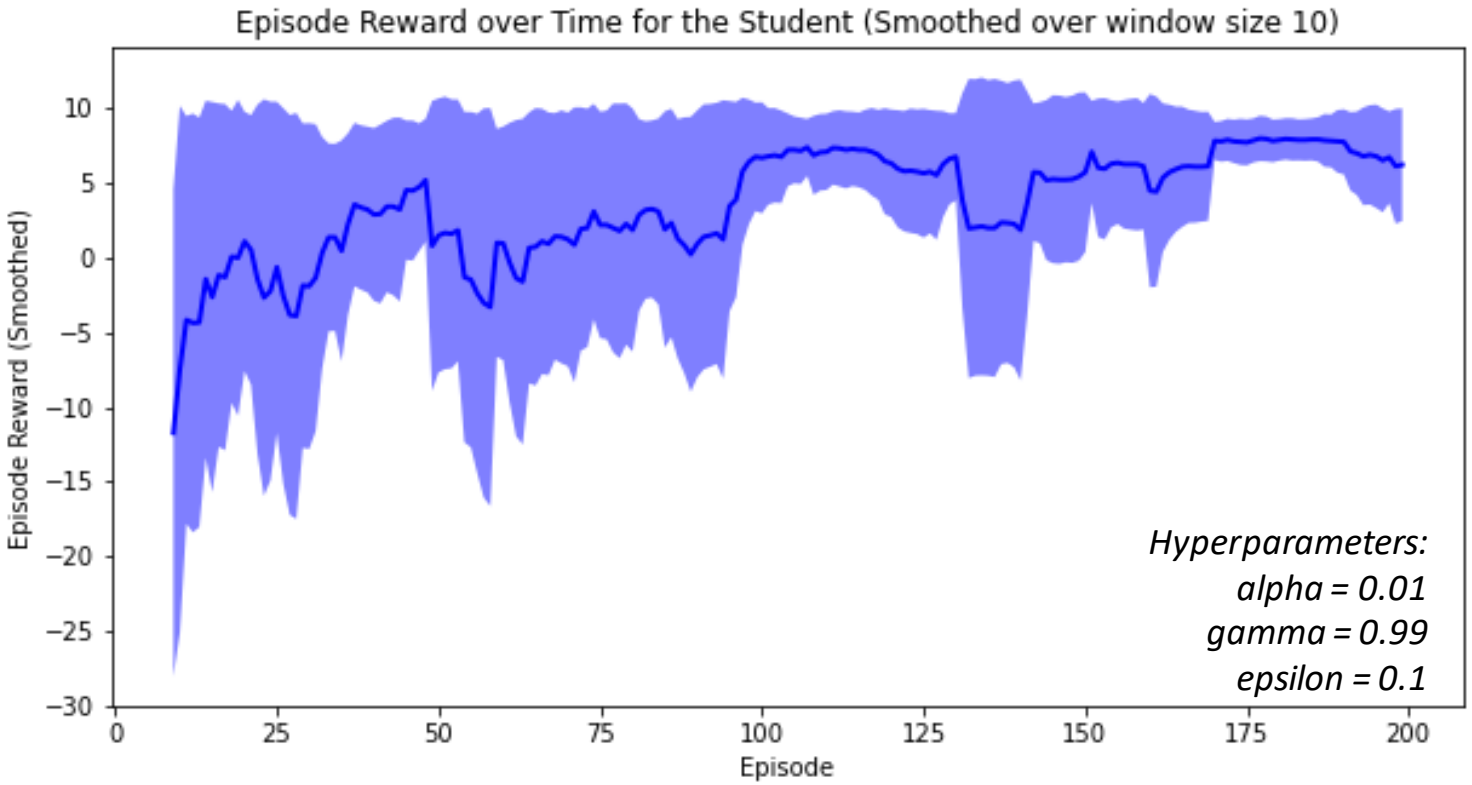


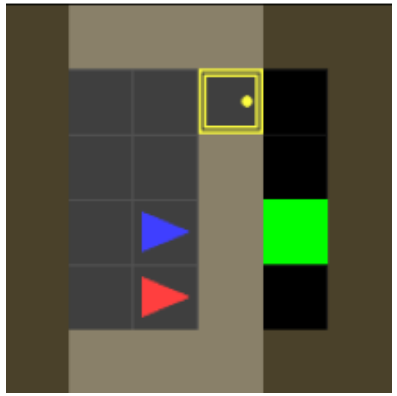
Training: 200 episodes
Test: 100 episodes

Results after 100 episodes:
Average time steps per episode: 10.0
Average reward per episode: 1.0

Env complexity: Level 2 (closed door)

Student



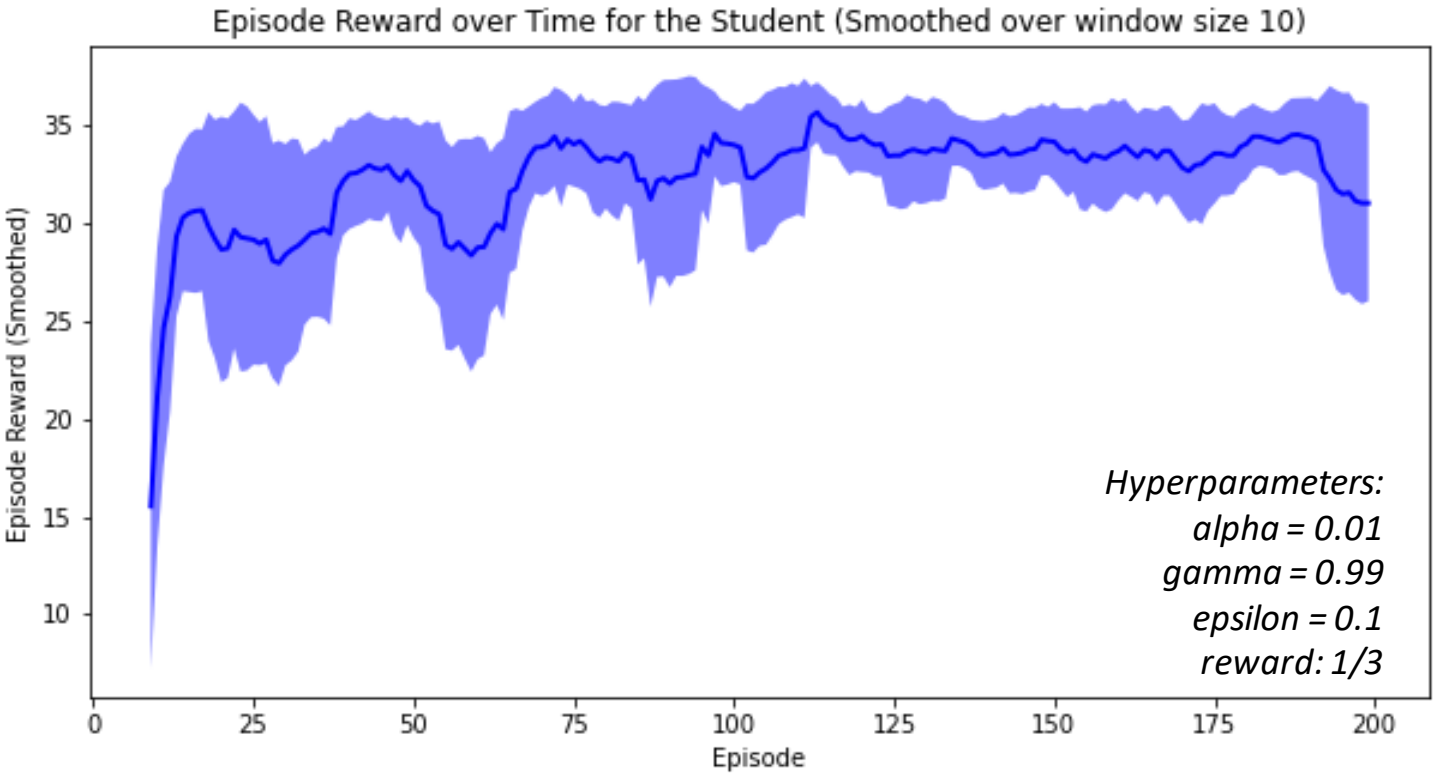


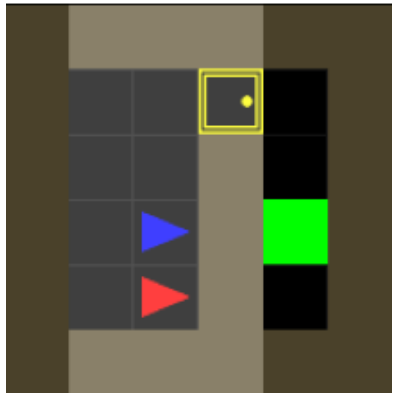
Training: 200 episodes
Test: 100 episodes

Results after 100 episodes:
Average time steps per episode: 10.0
Average reward per episode: 1.0

Env complexity: Level 2 (closed door & sparse learning)

Student



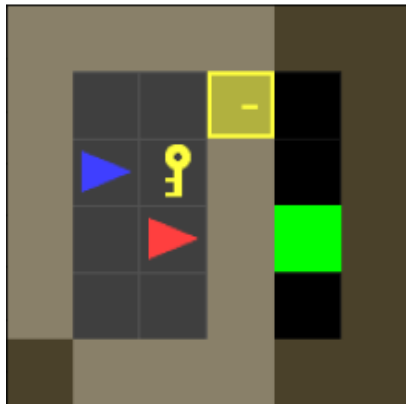


Training: 200 episodes
Test: 100 episodes

Results after 100 episodes:
Average time steps per episode: 200.43
Average reward per episode: 0.26

Environment complexity: Level 3

Harder for the agents, but also for the computer...



Final thought

It was fun to play around …

… but the road for me
to apply MARL to my research
questions is still long!

Thank you!

Q-learning for the Teacher