Exploration of Coordinated Behaviors in Multi-Agent Deep Reinforcement Learning

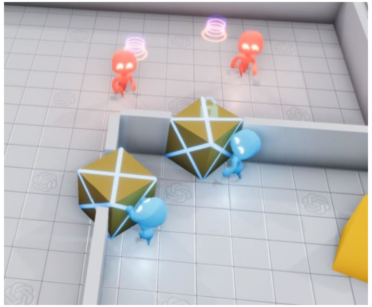
Maxime Toquebiau[1,2], Nicolas Bredeche[2], Faïz Ben Amar[2], Jae Yun Jun Kim[1]

[1]ECE Paris
[2]ISIR, Sorbonne Universités

March 14th, 2023

Multi-agent Deep Reinforcement Learning

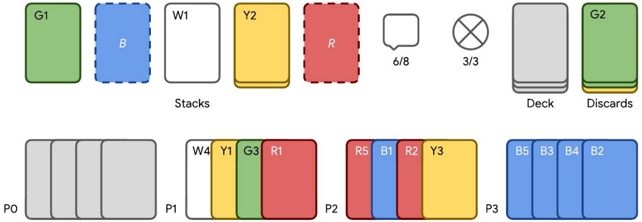
OpenAI Five (2019)[(1)]



Hide-and-seek (2019)[(2)]

Google Research Football (2019)[(5)]

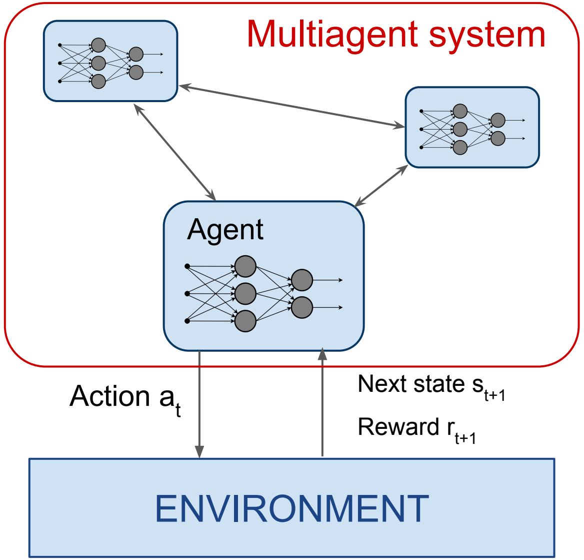
Starcraft Multi-Agent Challenge (2019)[(3)]



Hanabi (2019)[(4)]

[(1)]OpenAI et al., *Dota 2 with Large Scale Deep Reinforcement Learning*, 2019
[(2)]Baker et al., *Emergent Tool Use From Multi-Agent Autocurricula*, 2019
[(3)]Samvelyan et al., *The StarCraft Multi-Agent Challenge*, 2019
[(4)]Bard et al., *The Hanabi challenge: A new frontier for AI research*, 2020
[(5)]Kurach et al., *Google Research Football: A Novel Reinforcement Learning Environment*, 2019

Multi-agent Deep Reinforcement Learning (MADRL)



Multi-agent System

- Multiple agents interacting and learning concurrently

Deep Reinforcement Learning

- Agents try to maximise a global expected discounted return:

$$V_\pi(s_{t'}) = E_\pi\left[\sum_{t=t'}^{T} \gamma^t r_t\right]$$

- We use deep neural networks to model the agents' policy π and value V

$$\pi(s) = argmax_a V(s)$$

Issues with multi-agent systems

⇒ Credit assignment
⇒ Non-stationarity
⇒ Information sharing
⇒ …

Relative overgeneralization

	A	B	C
A	10	−5	−5
B	−5	7	7
C	−5	7	7

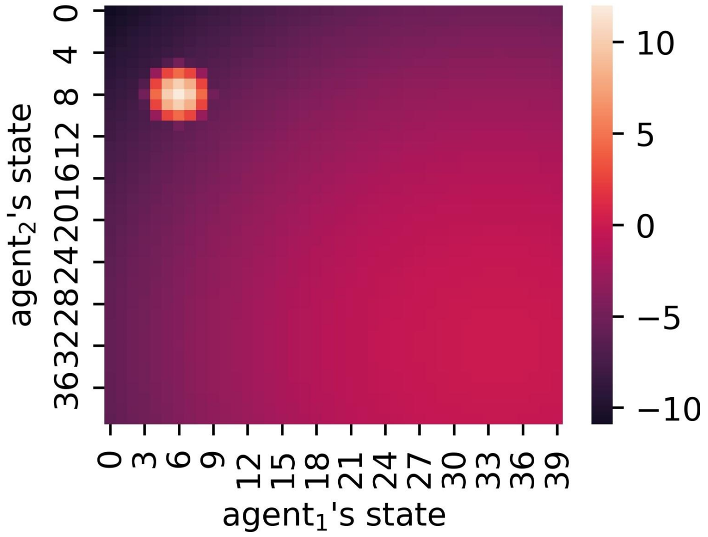
Optimal joint action: (A, A)

Expected return of local actions:
- Action **A**: 0
- Action **B**: 3
- Action **C**: 3

Issue: MADRL agents often prefer suboptimal local actions when they produce better expected returns than the optimal action.

Relative overgeneralization as an exploration problem

	A	B	C
A	10	-5	-5
B	-5	7	7
C	-5	7	7



Hypothesis: Exploring the *space of joint states* (i.e., (a_1, a_2)) thoroughly will enable agents to find the optimal reward spike more consistently.
BUT exploring local states will not help as it does not insure that agents visit the optimal reward spike.

Reward for exploring the joint-state space

Intrinsic motivation

Add an intrinsically generated reward to reward exploration of the environment:

$$r_t = r_t^e + \beta r_t^i$$

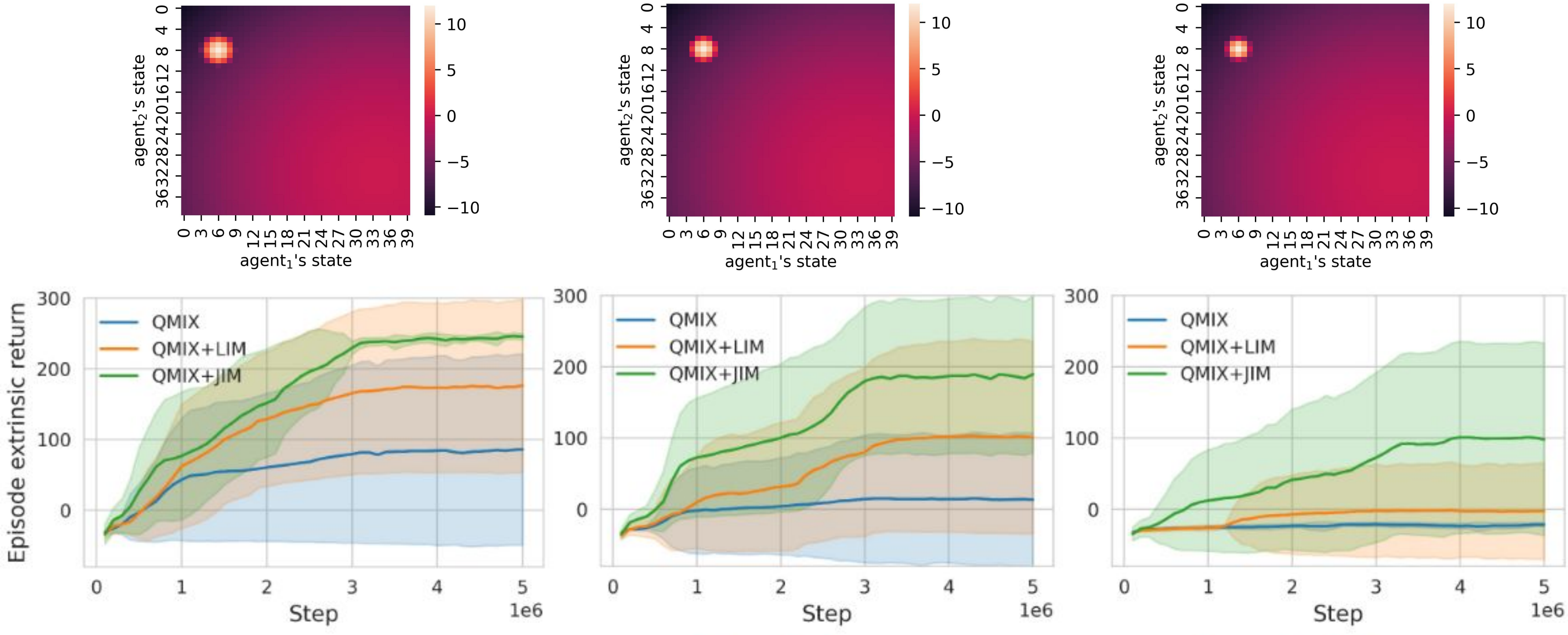
Joint Intrinsic Motivation

Reward for discovering new joint observations (i.e., concatenation of local observations):

$$r_t^i = r_t^{JIM}(\mathbf{o}_t, \mathbf{a}_t, \mathbf{o}_{t+1}) = N_l(\mathbf{o}_t, \mathbf{o}_{t+1}) \times \sqrt{2b(\mathbf{o}_{t+1})}$$

(1)Zhang et al., *NovelD: A Simple yet Effective Exploration Criterion*, 2021
(2)Henaff et al., *Exploration via Elliptical Episodic Bonuses*, 2022

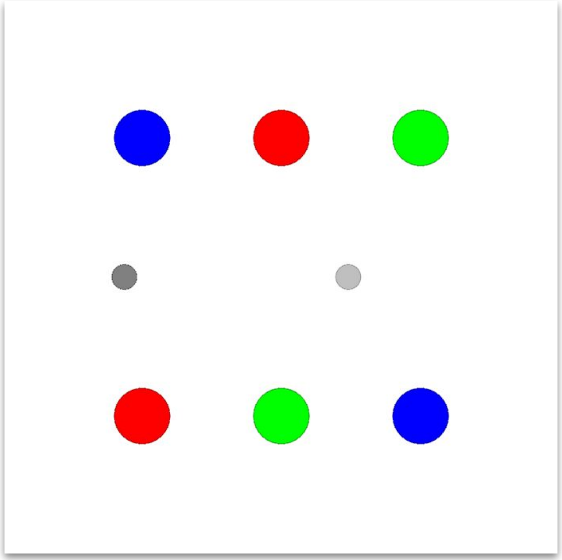
Results

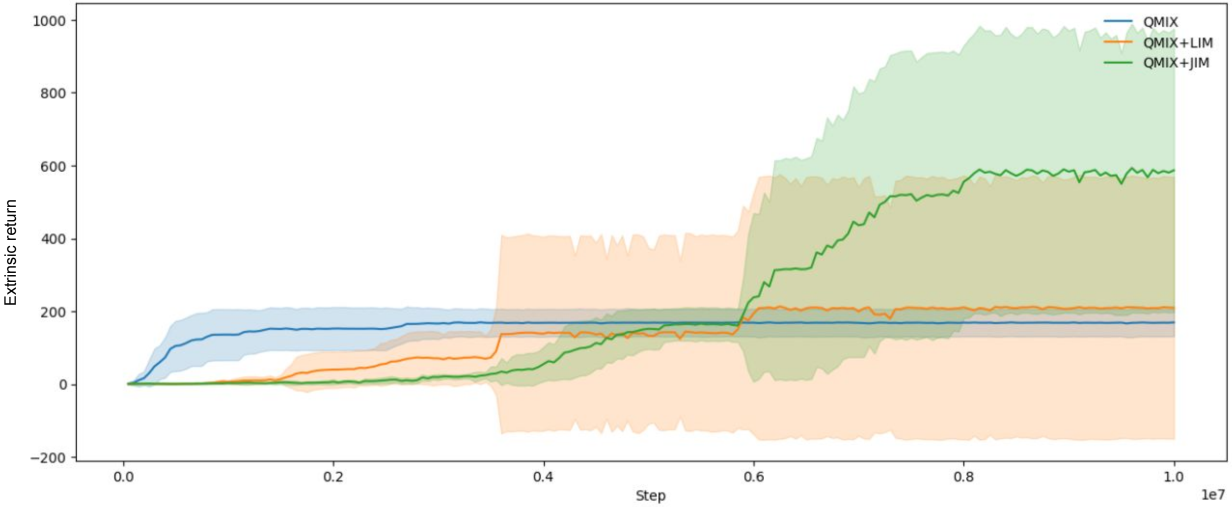


(a) easy (δ = 30) **(b) hard (δ = 40)** **(c) very hard (δ = 50)**

[1]Rashid et al., *QMIX: Monotonic Value Function for Deep Multi-Agent Reinforcement Learning*, 2018

Results





Rewards:

- Both on RED: 10
- Both on BLUE: 2
- Both on GREEN: 1
- One on BLUE/GREEN: 1
- Otherwise: 0

Next steps

- Paper under review in AAMAS workshop
- Work on paper to be submitted to ECAI:
    - Another setup with more agents
    - Ablation studies

Thank you for you attention !

Questions ?