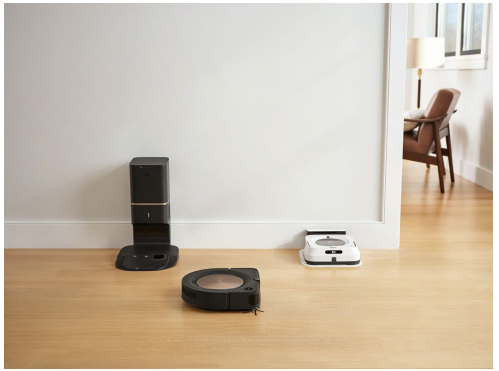
Algorithms and Benchmarks for Robust Multi-Agent Coordination

22/09/25

Multi-Agent Systems & Cooperation

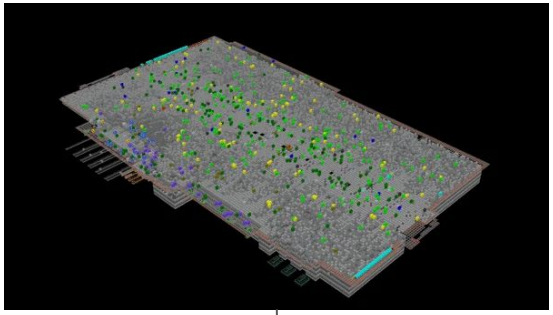


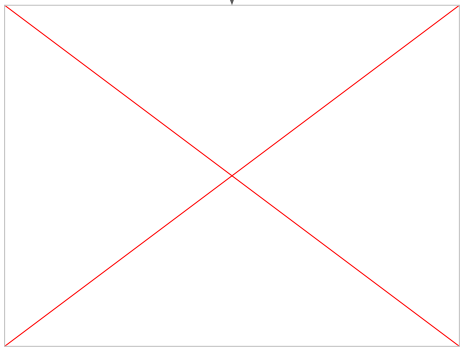
Multi-Agent Systems are everywhere!

You don’t even have to leave your house.

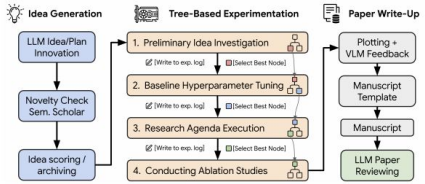
Source: Warehouse Traffic Home Robots

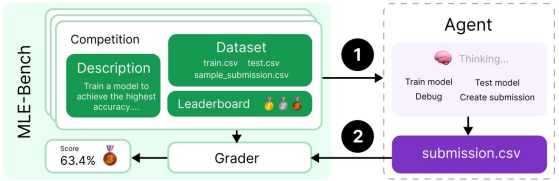
Multi-Agent AI Systems

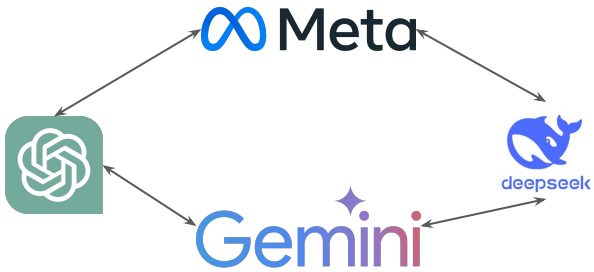




AI Agents - AI Scientist, MLE-bench (kaggle),...







Multi-Agent Coordination

- LLM Debate can improve reasoning [1,3,4]
- Division of labour improves performance on tasks such code generation [5]

Automation -> Competitive (Financial Incentives) -> More agents -> More interactions [2]

Specific Domains - Multi-Agent RL

Agents Powered by LLMs

For MA/MARL to work

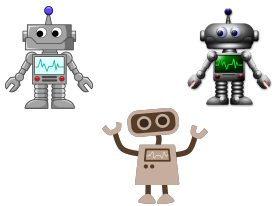
- Right ***algorithms & representations*** 🤖
- Right ***baselines & environments*** 🌍

Adaptability 🤖

HyperMARL: Adaptive Hypernetworks for Multi-Agent RL

https://arxiv.org/abs/2412.04233

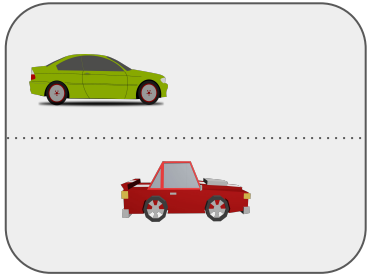
World is Big = Adaptation



Expect a lot -- **adaptability** -- different **behaviours**.

Javed, K. and Sutton, R.S., 2024. The big world hypothesis and its ramifications for artificial intelligence. ***Finding the Frame Workshop at RLC 2024.***

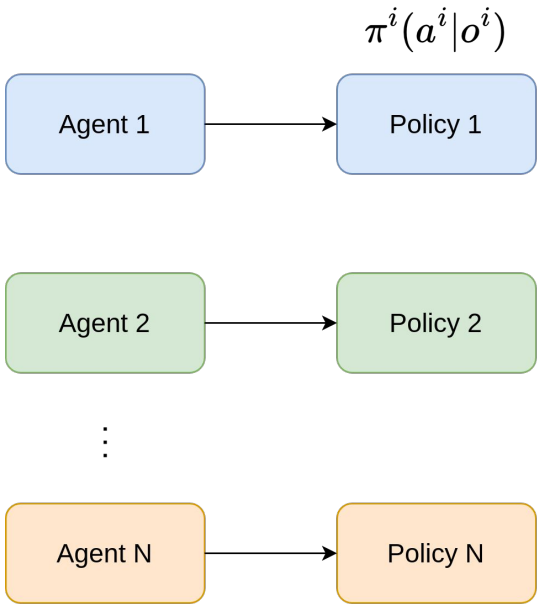
Adaptation: Individual Specialisation vs Shared Behaviours



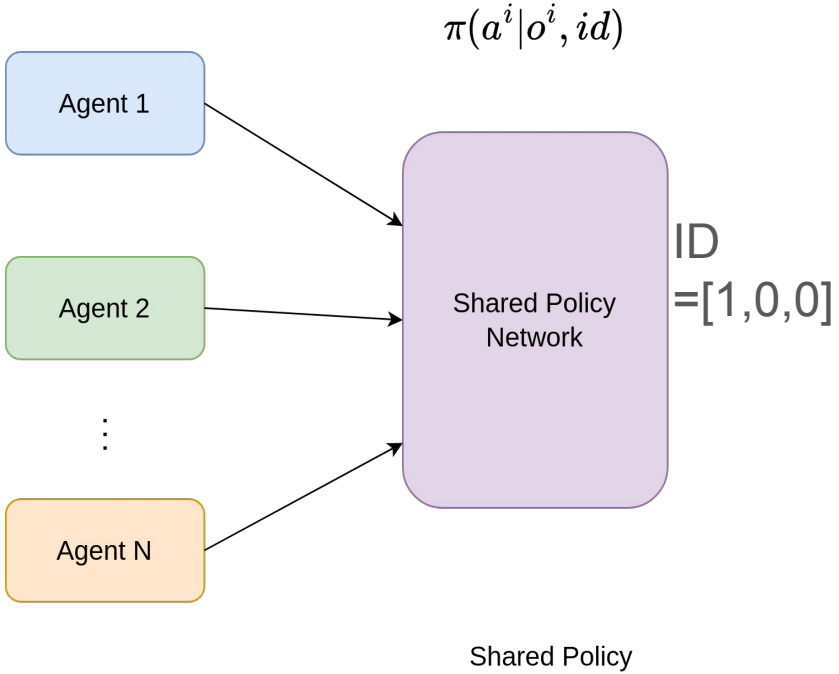
Source: Google Football, Search and Rescue, Robots, Birds, Fish.

Efficient Adaptability in MARL

No Parameter Sharing (NoPS)



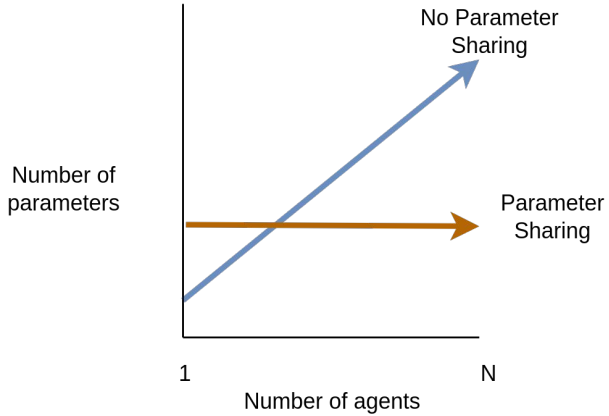
Full ***Parameter Sharing*** (FuPS)



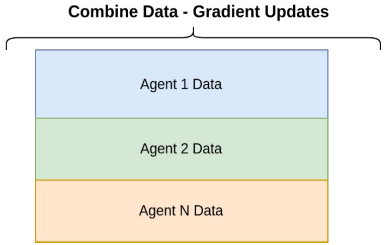
Full Parameter Sharing (FuPS)

Pros

Scalable



Sample Efficient - Parallel Gradient Updates



Cons

Challenges at learning different behaviour for agents – ***specialisation***.

Current Approaches - Specialisation

- **Intrinsic Rewards based on MI** [Li et al., 2021,Jiang and Lu, 2021] encourage diversity in the objective.
  - Influence the **learning objective**.
  - Complicated implementations.
  - Outperformed by FuPS and NoPS [Fu et al., 2022]

Architecture Based Approaches:

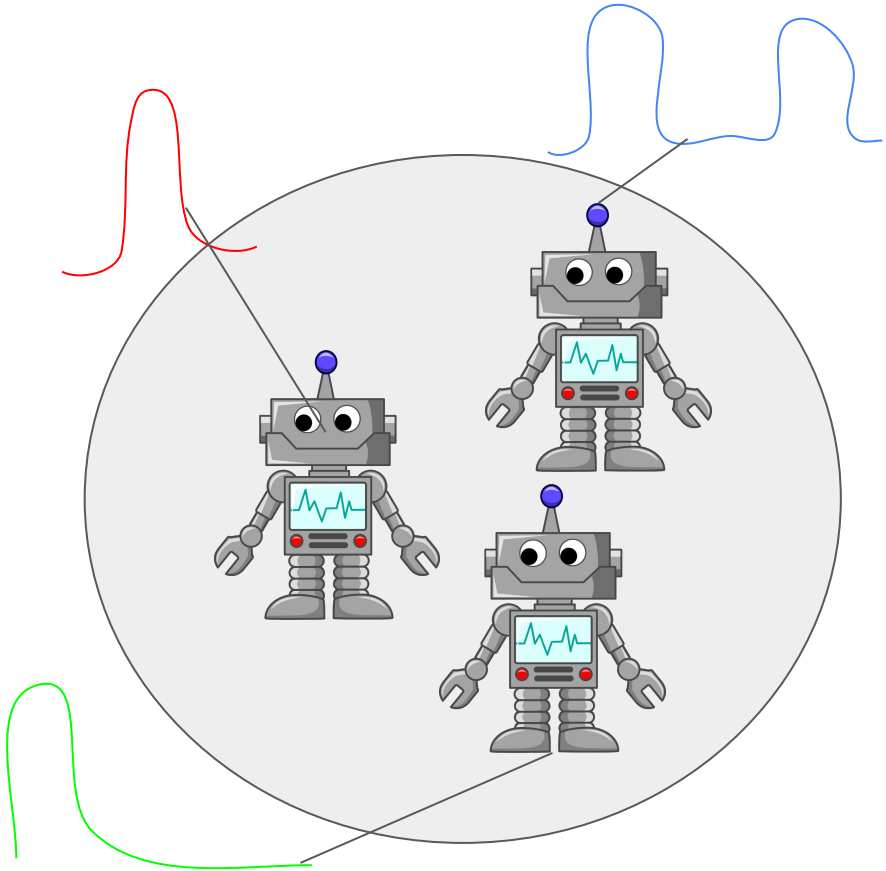
- **Diversity Control (DiCo)** [Bettini et al., 2024] allow control desired diversity levels.
  - Require **prior knowledge of the optimal diversity level**
  - Still need shared and non-shared parameters.

- Kaleidoscope [Li et al.,2024] - leanable masks.
  - Still requires **diversity loss/objective**
  - Many hyperparms, complicated, delicate implementation.

Goal: Universal/Adaptive MARL Architecture

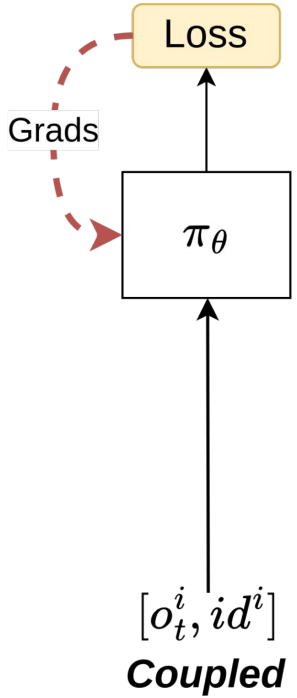
Goal:

- Can we develop a general purpose method that ***adapts*** to both **specialisation and homogenous behaviours**?
- Can we do this without
  - interfering with the **learning objective**?
  - knowing the **optimal diversity level**?
  - requiring **sequential updates**?
- Using a shared architecture.



Problem: Parameter Sharing + Specialisation

Coupling agent-IDs and observations = **higher gradient interference**



Hypernetworks

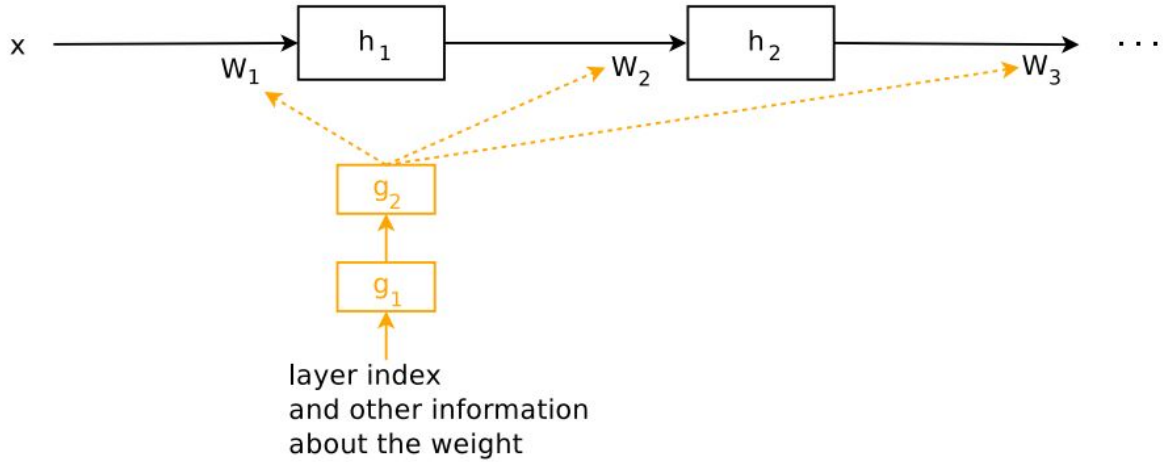
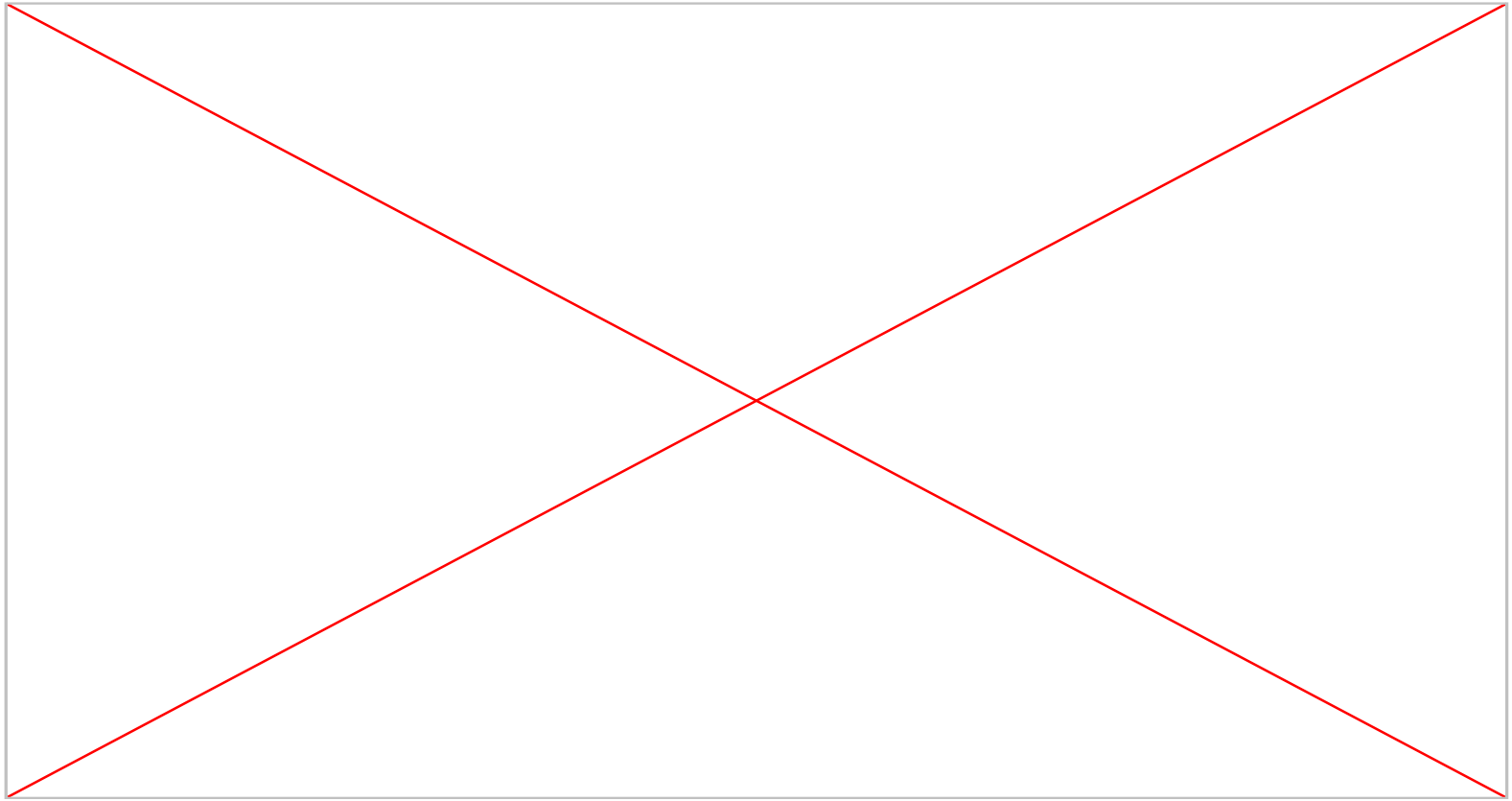


Figure 1: A hypernetwork generates the weights for a feedforward network. Black connections and parameters are associated the main network whereas orange connections and parameters are associated with the hypernetwork.

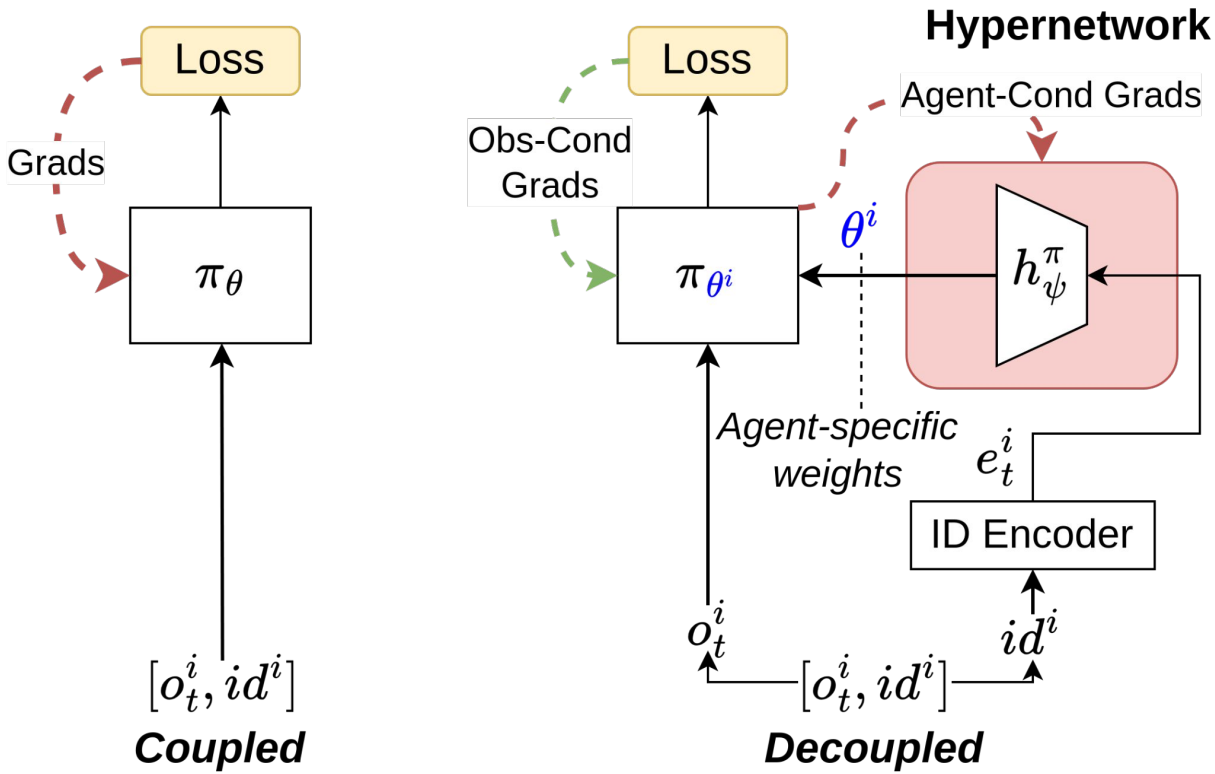
Ha D, Dai AM, Le QV. HyperNetworks. **ICLR 2022**

HyperMARL



Tessera, K.A., Rahman, A., Storkey, A. and Albrecht, S.V., 2025. HyperMARL: Adaptive Hypernetworks for Multi-Agent RL. ***CoCoMARL Workshop at RLC 2025 & NeurIPS 2025.***

HyperMARL - Agent Conditioned Hypernetworks



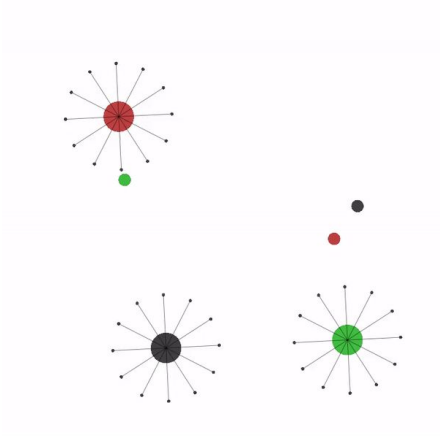
HyperMARL: Gradient Decoupling

Weights of hnets

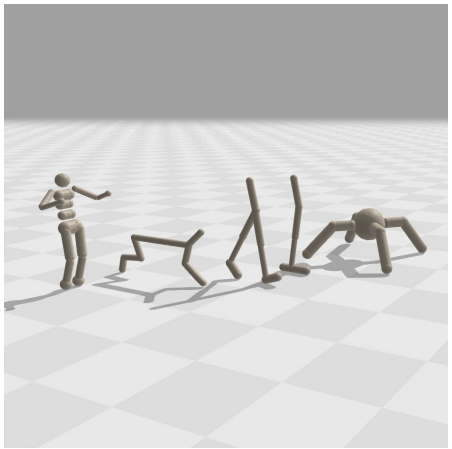
$$\nabla_\psi J(\psi) = \sum_{i=1}^{I} \underbrace{\nabla_\psi h_\psi^\pi(e^i)}_{\mathbf{J}_i \text{ (agent-conditioned)}} \underbrace{\mathbb{E}_{\mathbf{h}_t, \mathbf{a}_t \sim \boldsymbol{\pi}} \left[A(\mathbf{h}_t, \mathbf{a}_t) \nabla_{\theta^i} \log \pi_{\theta^i}(a_t^i \mid h_t^i) \right]}_{Z_i \text{ (observation-conditioned)}}.$$

- ***Agent-conditioned:*** deterministic wrt to mini-batch samples, separating agent identity from traj noise.
- ***Observation-conditioned***: averages trajectory noise **per agent**.

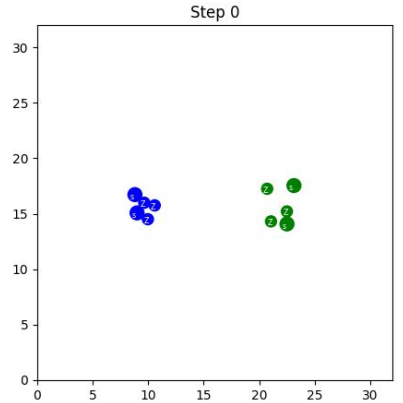
Many Environments - Some Test Specialisation, Homogenous Behaviours, Mixed



Navigation - Up to 8 agents

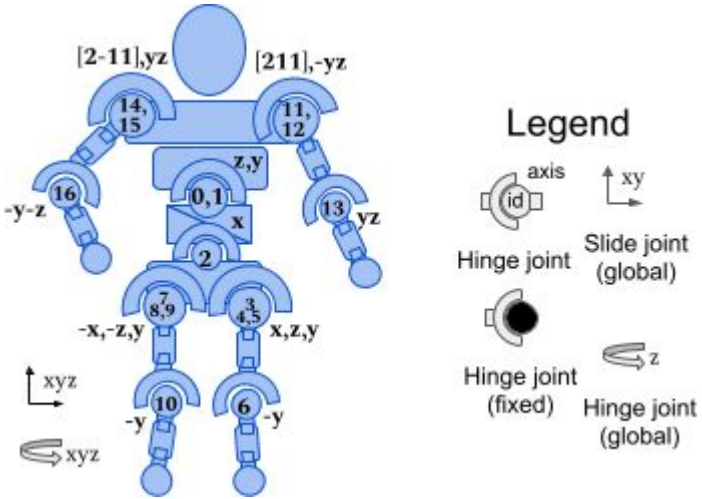


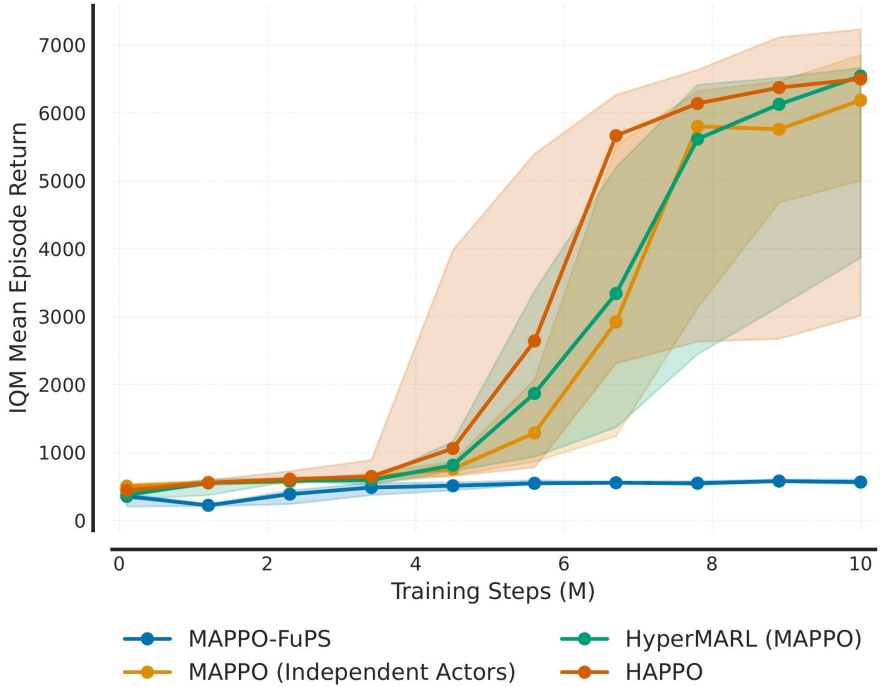
Multi-Agent Mujoco - Up to 17 agents



SMAX (Starcraft) - Up to 20 agents

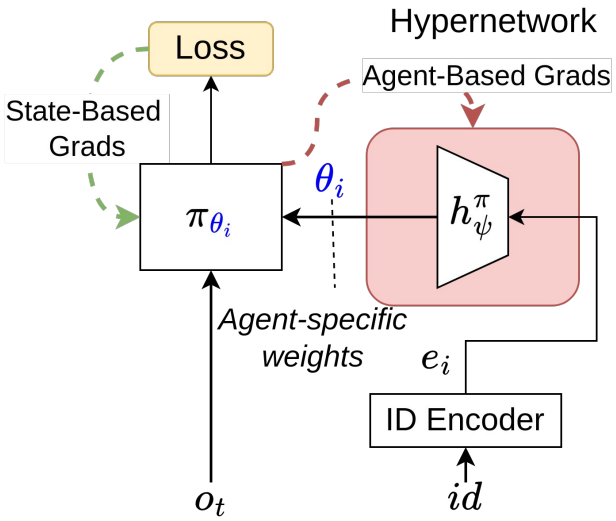
Humanoid-v2 17x1



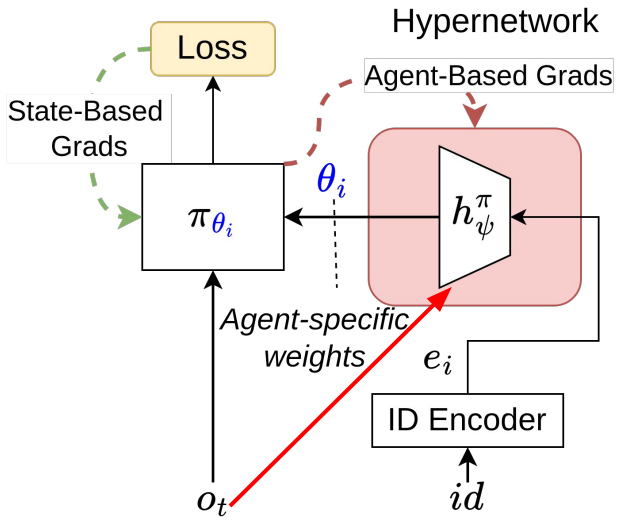


*__Large confidence interval,__ but shows that a __shared representation (HyperMARL)__ is comparable to __non-shared methods__ (HAPPO and Independent Actors).

Ablations - Decoupled Grads

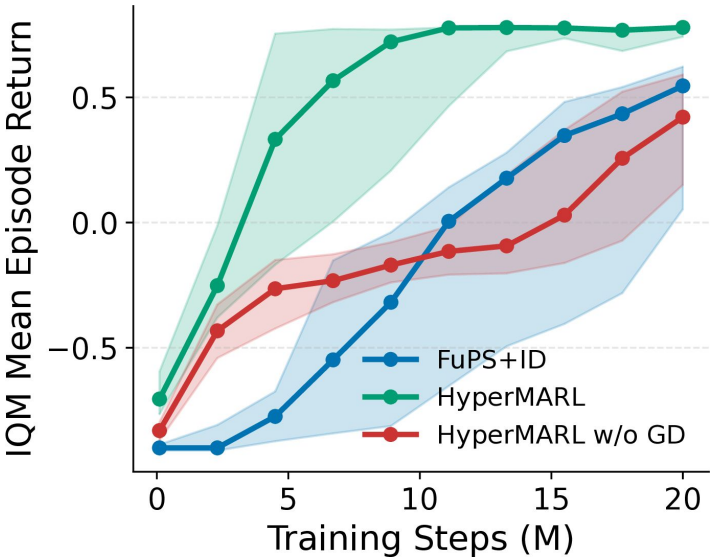


HyperMARL

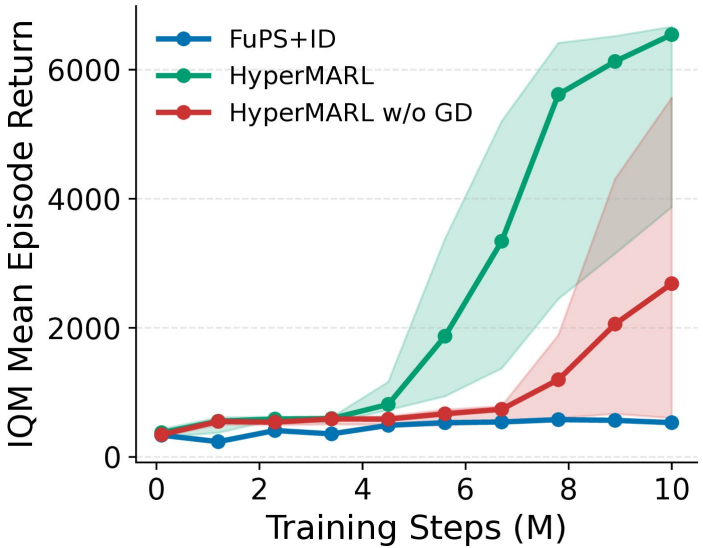


HyperMARL w/o decoupled grads

Ablations - Decoupled Grads



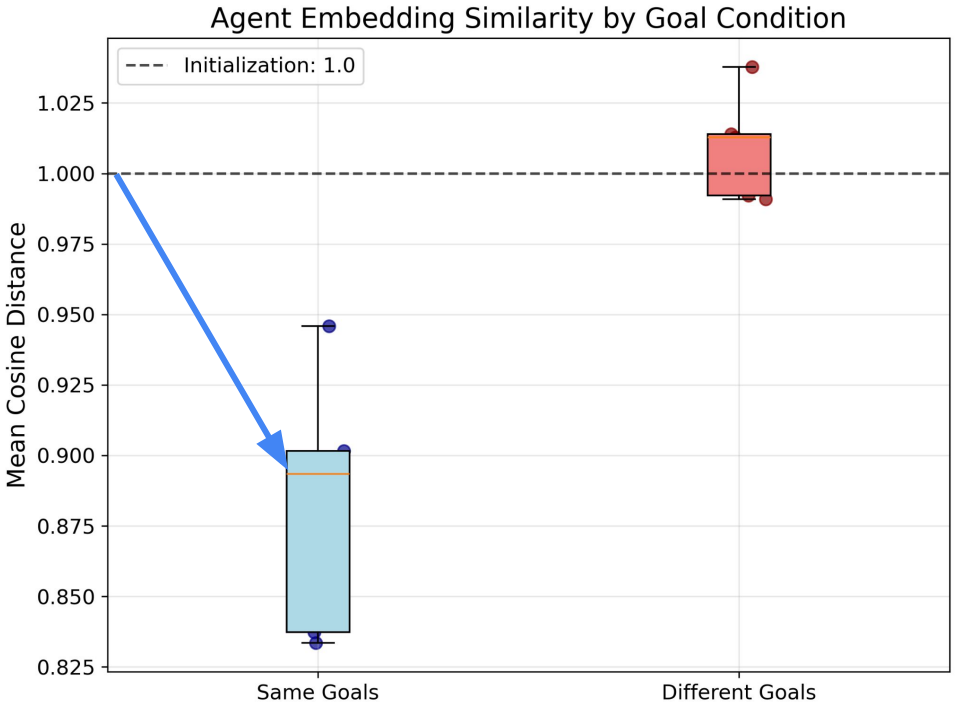
Dispersion



Humanoid MAMuJoCo

Ablations - Agent Embeddings - Match Goal

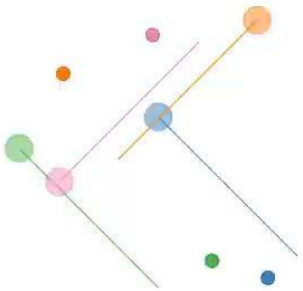
Same Goals

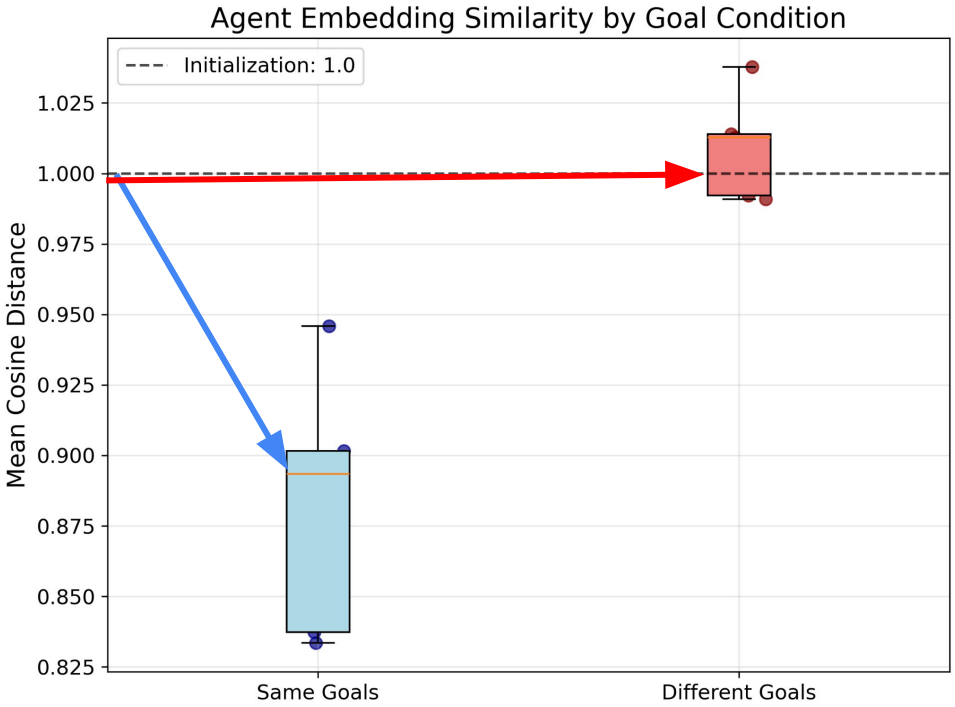


Ablations - Agent Embeddings - Match Goal

Same Goals

Diff Goals





Takeaway

- Simple Idea:
  - ***Gradient decoupling*** is a key ingredient in mitigating cross-agent interference – correlates with a lower grad variance.
- Simple Implementation:
  - Agent-conditioned Hypernetworks can achieve gradient decoupling – possible to learn adaptive behaviour, across diverse tasks, ***without altering the learning objective, preset diversity levels or sequential updates.***
- Challenges**:**
  - **#params -> chunked hypernets or low-rank approximations.**

Current MARL Baselines 🌍

Remembering the Markov Property in Cooperative MARL

https://arxiv.org/abs/2507.18333

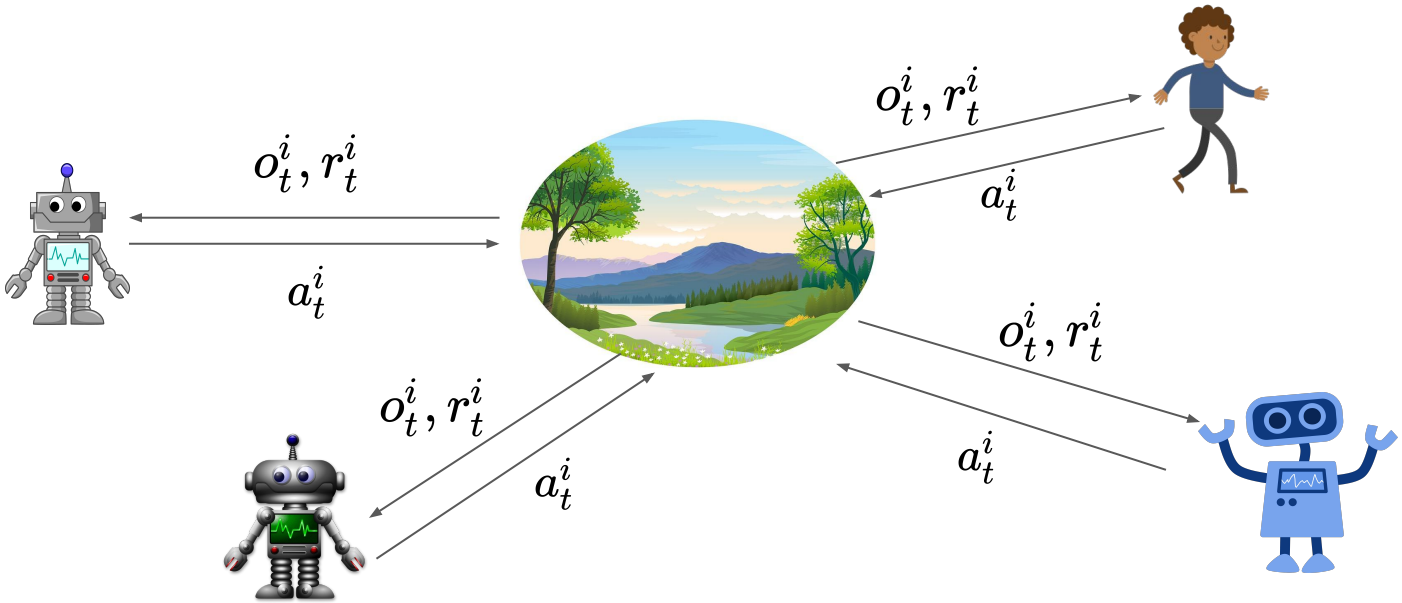
Problem

We know multi-agent systems are and will be everywhere.

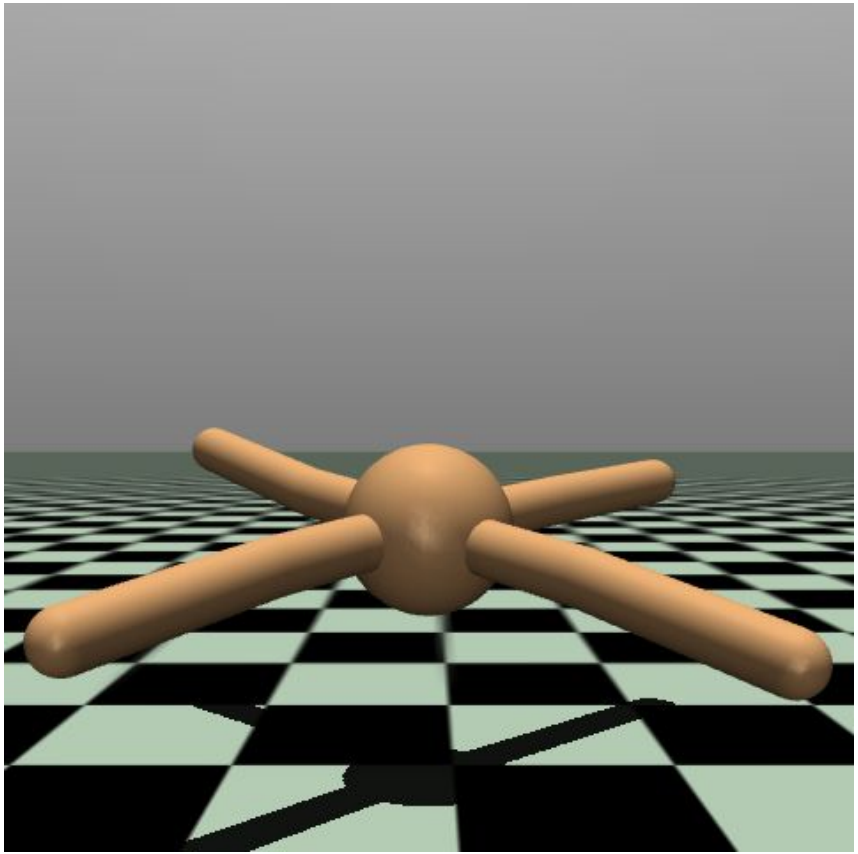
Therefore, we need to be able to adequately measure progress:

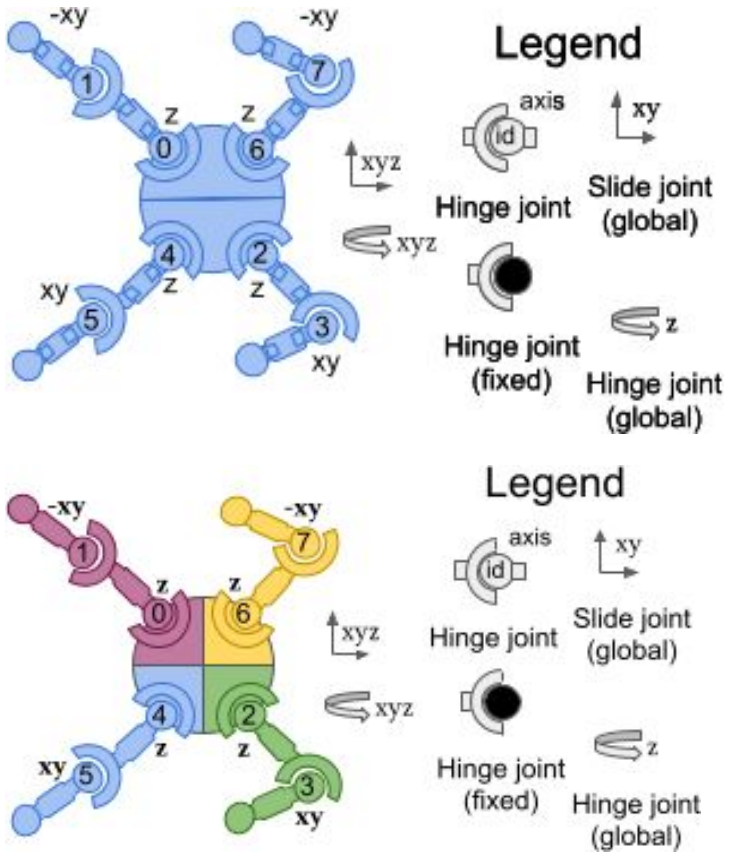
We need virtual MA environments that test ***properties*** of MA systems that we care about.

What is Multi-Agent Anyways?

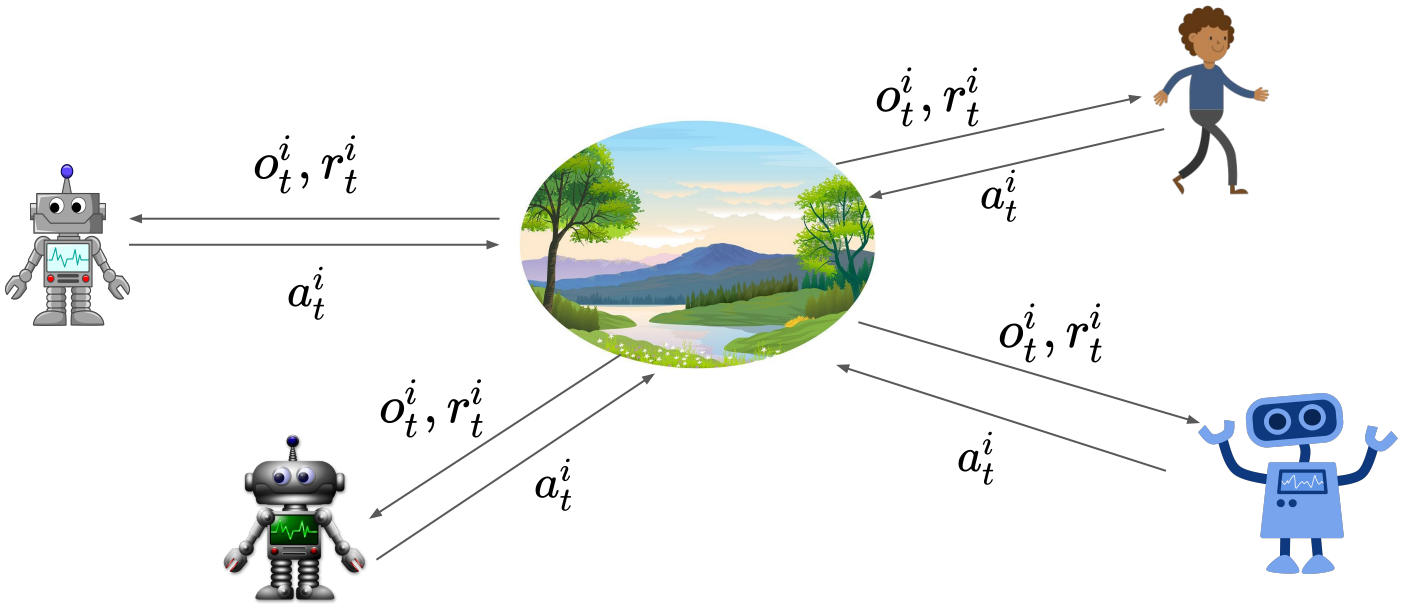


What is Multi-Agent Anyways?





MARL



1. ***Partial Observability*** - Agents can't see the full world.
2. ***Decentralised execution*** - agents can act based on their own obs.

Cooperative MARL - Dec-POMDPs

Cooperative MA Problems: Dec-POMDP:

$$(\mathcal{N}, \mathcal{S}, \mathbb{T}, \mathbb{O}, \mu, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \{\mathcal{O}^i\}_{i\in\mathcal{N}}, R, \gamma)$$

Action of all agents.

$R(s, \mathbf{a})$, common reward across agents

$\mathbb{T}(s_{t+1} \mid s_t, \mathbf{a}_t)$ If env is in state s_t and joint action a_t, what is the distribution of future state.

Goal:

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} \mathbb{E}_{s_0\sim\mu,\, \mathbf{a}_t\sim\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \mathbf{a}_t)\right].$$

Cooperative MARL - Dec-POMDPs

1. ***Partial Observability (POMDPs, Dec-POMDPs)* –** Agents receive incomplete information about the state of the environment. Typically modelled by using memory or recurrency.
2. ***Decentralised execution*** - agents can act based on their own obs.

Markov Property

A Dec-POMDP has the Markov property if the **current state** contains all relevant information for predicting the future [6,7].

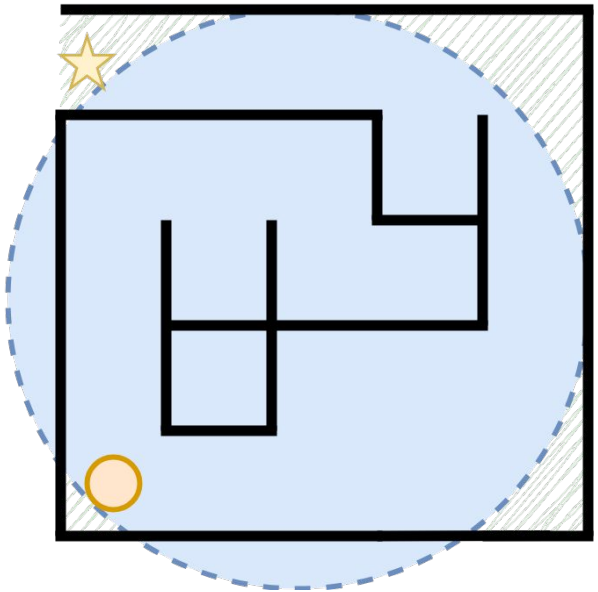
Formally, this means the transition dynamics do not depend on the full history:

$$\mathbb{T}(s_{t+1} \mid s_t, \mathbf{a}_t) = \mathbb{T}(s_{t+1} \mid s_t, \mathbf{a}_t, s_{t-1}, \mathbf{a}_{t-1}, \ldots, s_0, \mathbf{a}_0)$$

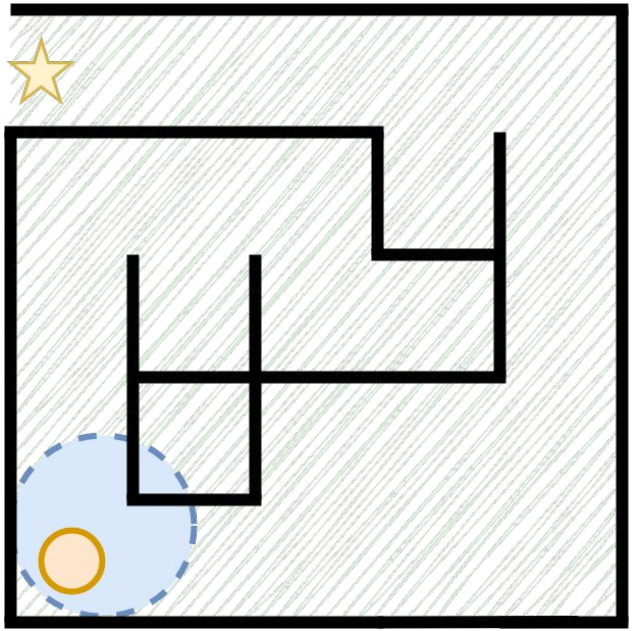
But in Dec-Pomdps, you don't have access to full state!

* Lots of RL learning methods rely on Markov Prop.

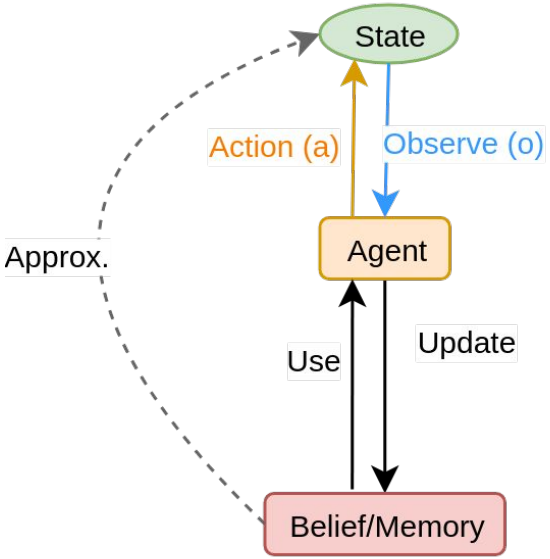
Partial Observability - Effective MDP

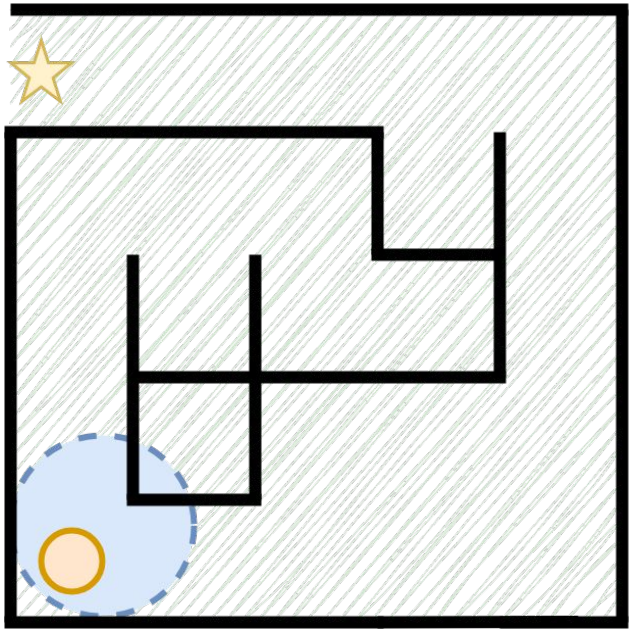


Partial Observability - POMDPs



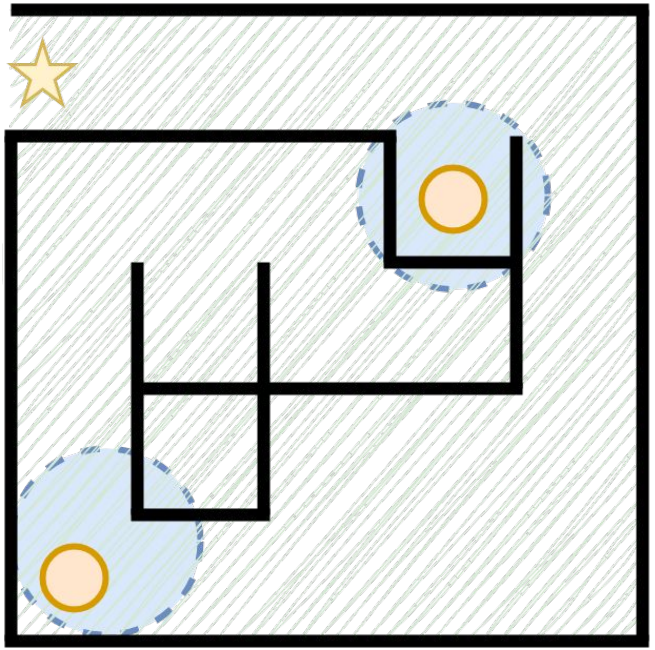
Partial Observability - POMDPs



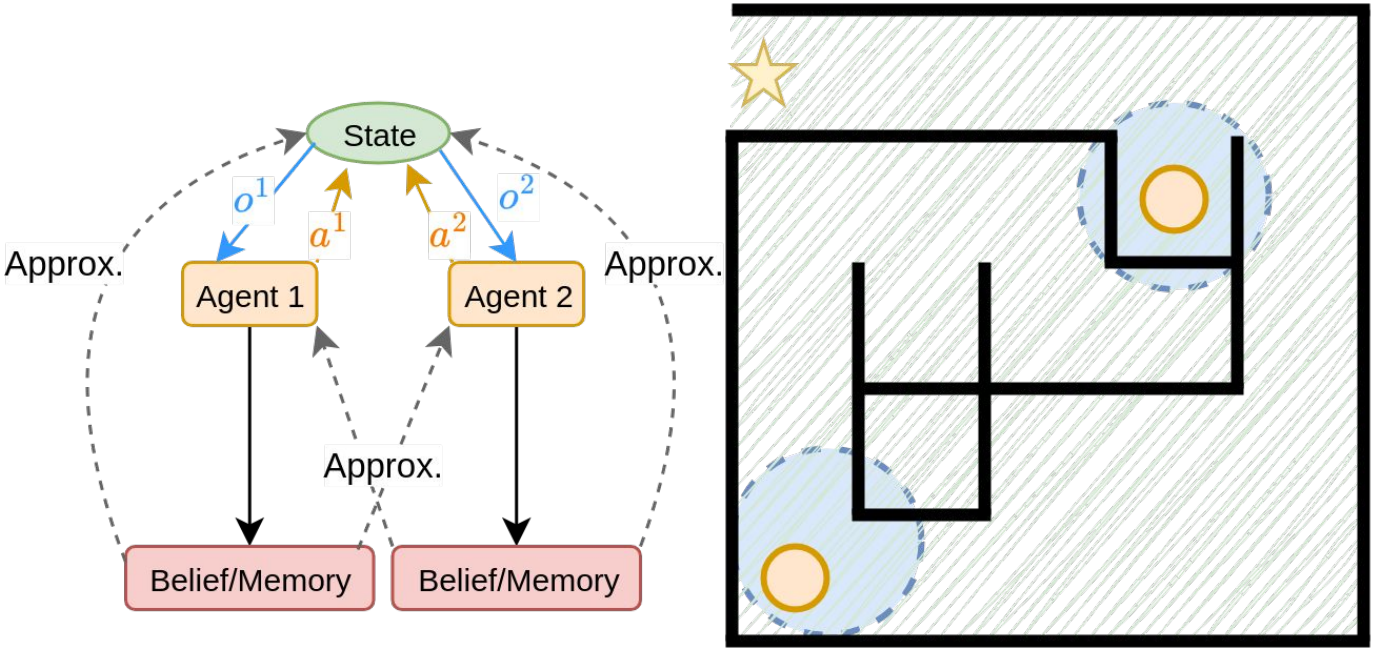


If you can't see, you must remember

Partial Observability - Dec-POMDPs



Partial Observability - Dec-POMDPs - Predict



If you can't see, you must predict

Partial Observability - Dec-POMDPs - Private Information

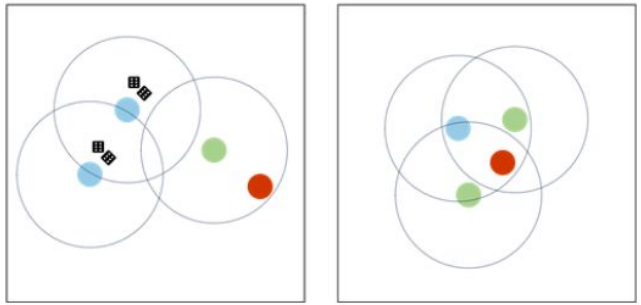


Figure 5: Example EPO enemy sighting. Allied units that do not observe the enemy are shown in *blue*, those that do are shown in *green* and the enemy unit in *dark red*. Initially, an ally spots an enemy. Later (right), when the enemy is within all allied sight ranges, only the first ally to observe the enemy and the ally for which the draw was successful can see it.

- Needs **private information** from other agents)
- “Meaningful Partial Observability” (Ellis et al.). Part of the proof that planning in Dec-POMDPs is NEXP-complete with n>=2 agents (Bernstein et al.).

Decentralisation

- **Easy**: Decentralised, but no coordination required.
- **Medium**: Agents need to coordinate with other agents.
- **Hard**: Medium + Agents need to predict the actions/behaviour of other agents to coordinate on tasks (other agent's non stationary).

MARL Environments - Prop of Cooperative MA Problems

Mutual Information

$\mathbb{I}(O; A)$ Dependence between action and obs.
Are agents using their observations?

$\mathbb{I}(H; A)$ Dependence between hist (rnn hidden state) and obs.
Are agents using their histories?

$\mathbb{I}(a_j; h_i | h_j)$ How much extra information does agent i's history/hidden state, provide about agent j actions, excluding what we know from agent j history?
Do agents have private information relevant to others?

$\mathbb{I}(A_j^{t+1}; Z_i^t | Z_j^t)$ How much extra information does agent i's final layer (Z) representation provide about agent j's next action, excluding what we know from agent j's final layer representation?
Are agents predicting each others' actions?

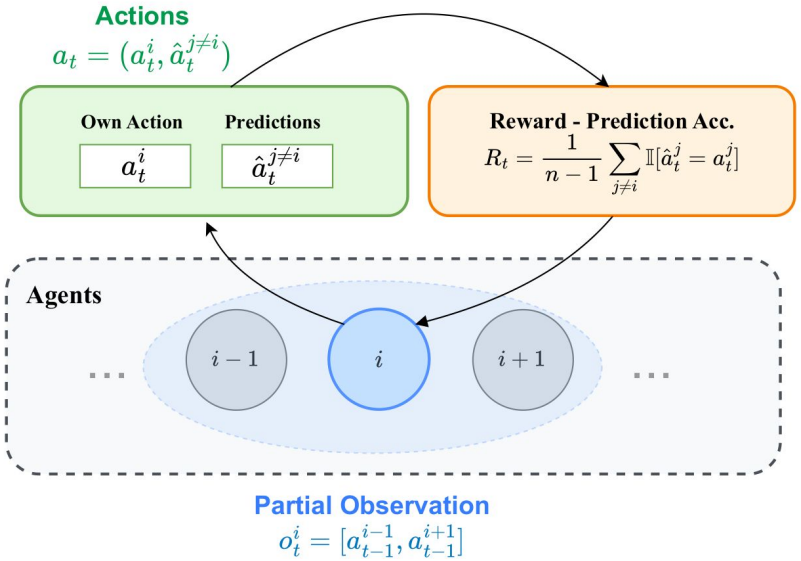
Metrics for MA Coordination

- **P**artial **O**bservability (**PO**MDPs, Dec-**PO**MDPs) – Agents receive incomplete information about the state of the environment. Typically modelled by using memory or recurrency.
  - ***Easy***: Env is not really partially observable.
    - Measure: Return(RNN) ≈ Return(MLP); I(H;A | O) = 0
  - ***Medium***: When hidden information is ***relevant*** to the task and ***cannot be inferred*** from the *single* observations alone. This is also the case in single-agent POMDPs.
    - Measure: Return(RNN) > Return(MLP); I(H;A | O) >0
  - ***Hard***: *Medium* + cannot be resolved by a ***single agent in isolation*** (needs private information from other agents).
    - Measure: Medium + I(a_j; h_i | h_j)>0
- **Dec**entralisation (**Dec**-POMDPs)/Coordination – Agents need to be able to act based on their local history of observations and/or actions.
  - ***Easy***: Decentralised, but no coordination required.
    - Measure: Return(IPPO) ≈ Return(MAPPO);
  - ***Medium***: Agents need to coordinate with other agents.
    - Measure: Return(MAPPO) > Return(IPPO)
  - ***Hard***: *Medium* + Agents need to predict the actions/behaviour of other agents to coordinate on tasks.
    - Measure: Medium + I(A_j^{t+1}; Z_i^t | Z_j^t)/ linear probe Z_i^t -> A_j^{t+1}

Metrics for MA Coordination

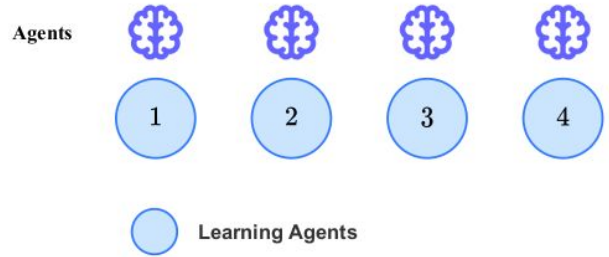
Question	ENV
i) Is Partial Observability relevant to the task?	
ii) Is Partial Observability reliant on private information across agents?	
iii) Is Coordination non-trivial?	
iv) Does Coordination require the prediction of other agents' actions?	

Case Study: Brittle Conventions vs. Robust Coordination

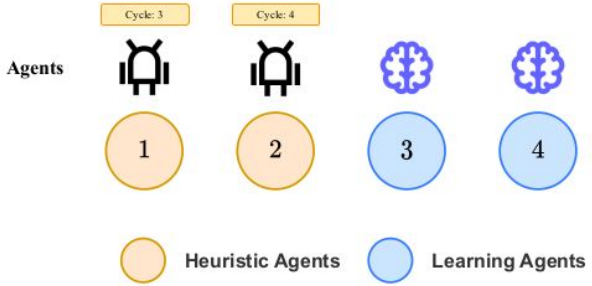


Prediction Game

Case Study: Brittle Conventions vs. Robust Coordination

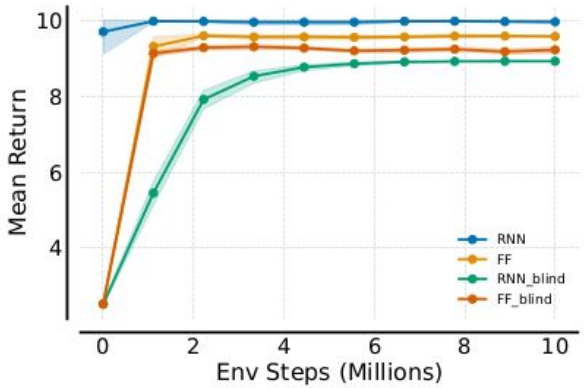


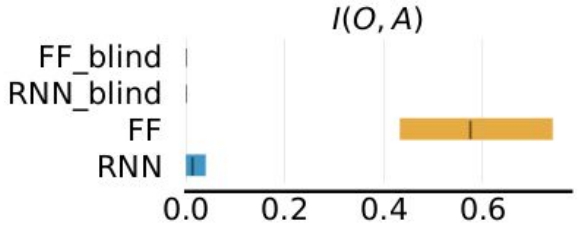
(b) Scenario A: Co-adapting



(c) Scenario B: Mixed

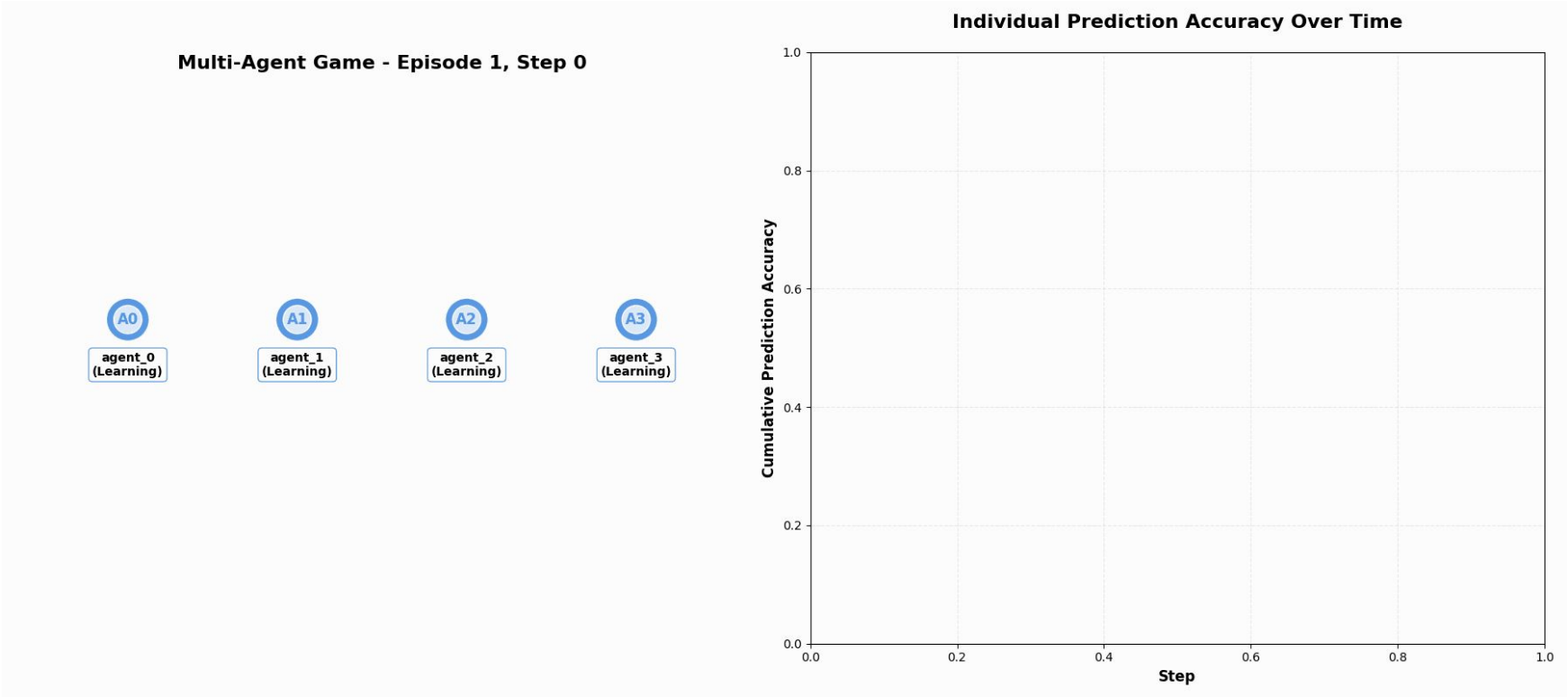
Case Study: Scenario A



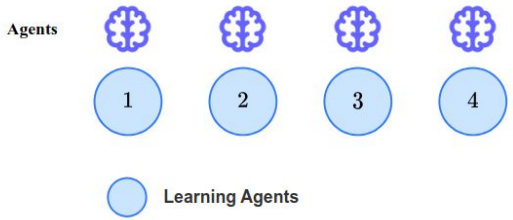


(a) Co-adapt: High Perf. (b) **Low** $I(O, A)$

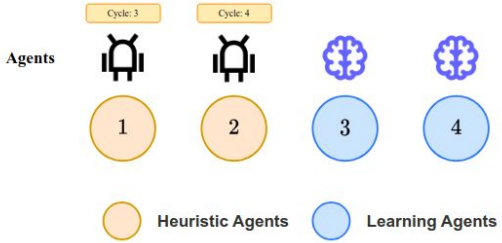
MARL Environments - Why are “blind conventions” bad?



MARL Environments - Why are “blind conventions” bad?



(b) A homogeneous setup with four learning agents.

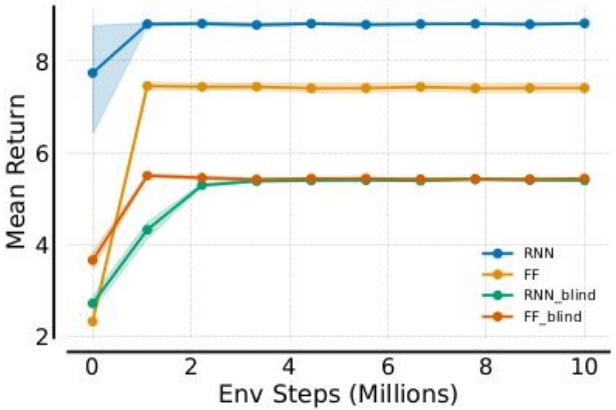


(c) A heterogeneous setup with two learning and two heuristic agents.

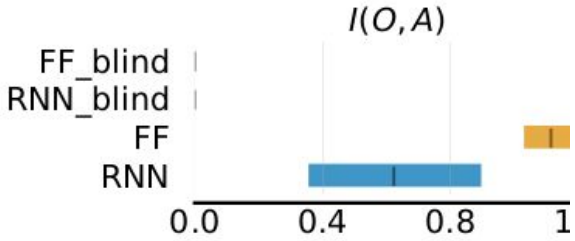
Scenario	RNN	FF
Baseline	9.97 (9.92, 10.03)	9.59 (9.55, 9.63)
Add Heuristic	4.80 (4.41, 5.20)	4.04 (3.26, 4.81)

- Brittle to **stochastic changes** in env.
- Not robust to **partners**.
- Poor **generalisation**.

Case Study: Scenario B



(c) Mixed: High Perf.



(d) **Higher** $I(O, A)$

MARL Environments - Why are “blind conventions” bad?

BRITTLE CONVENTIONS VS. ROBUST COORDINATION

Same method, but the **mechanism** for success changes with environment modifications (partner composition).

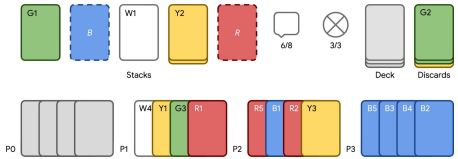
Implication: Current MARL environments may **enable** fragile co-adaptation rather than robust cooperation.

Partial observability become not relevant to the task – allowed forming on conventions.

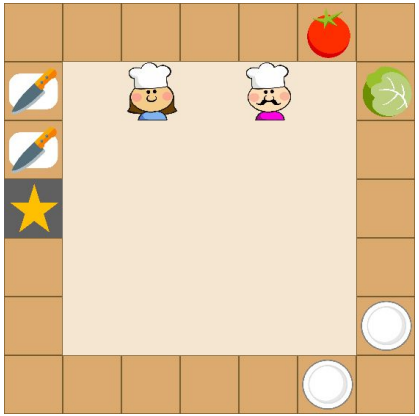
What kind of equilibria does your environment make easy to find (attractive)?

MARL Environments

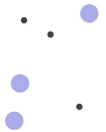
Hanabi



Overcooked

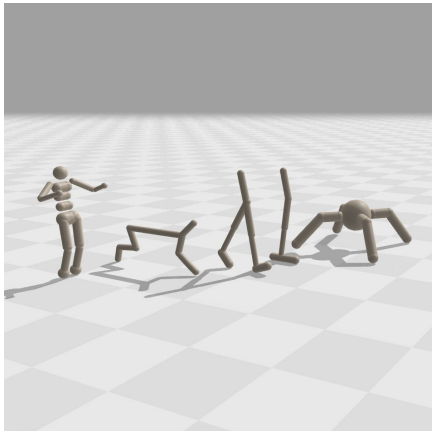


MPE



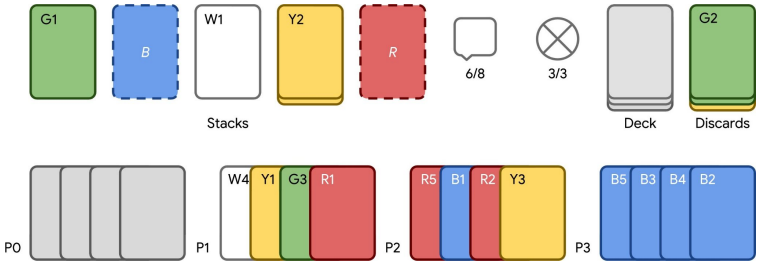
SMAX v1 & v2

MaBrax

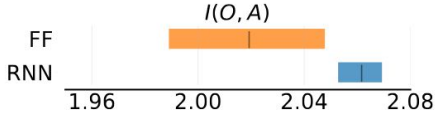


MARL Environments

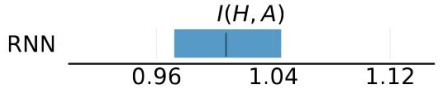
Hanabi



1. Policies grounded in obs/hist ✅

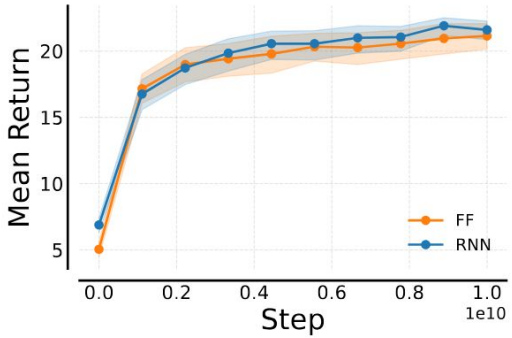


(d) Hanabi $(\mathbb{I}(O,A)_{max} \approx 3)$



(e) Hanabi $(\mathbb{I}(H,A)_{max} \approx 3)$

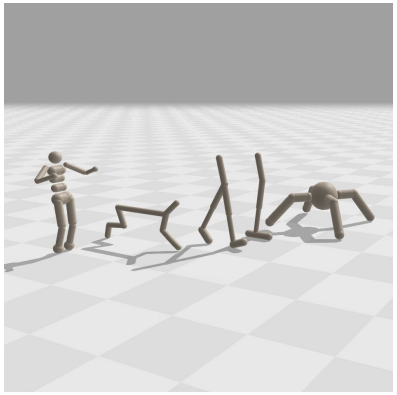
2. Recurrent reasoning about other agents. ❌



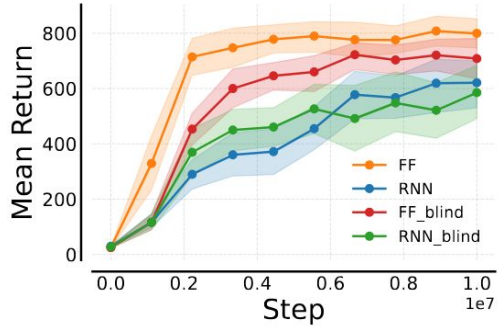
$$\mathbb{I}(O;A) > \mathbb{I}(H;A)$$

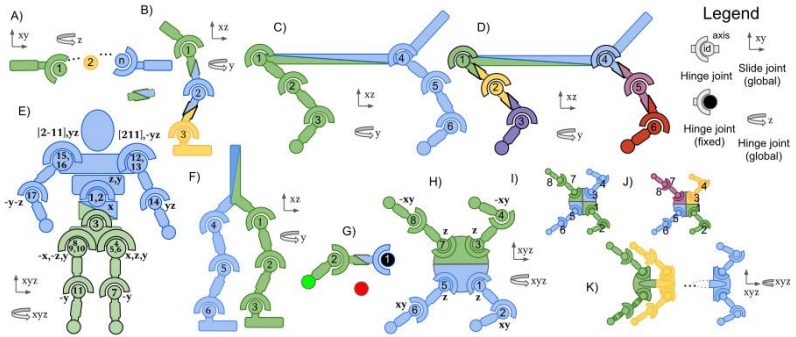
MaBrax

MARL Environments

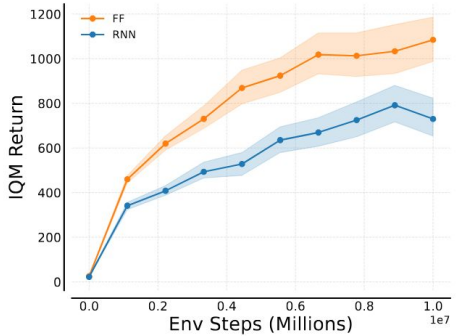


1. Policies grounded in obs/hist ❌



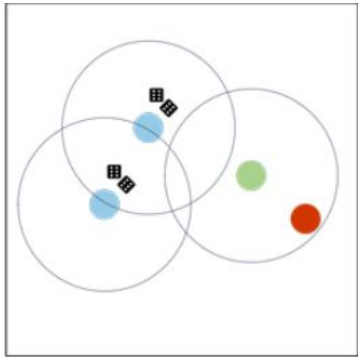


2. Recurrent reasoning about other agents. ❌

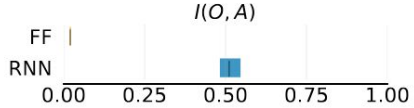


MARL Environments

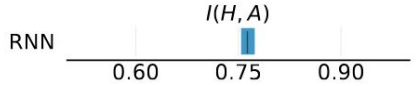
SMAX v2 (approx. of Starcraft 2)



1. Policies grounded in obs/hist ❌ ✅

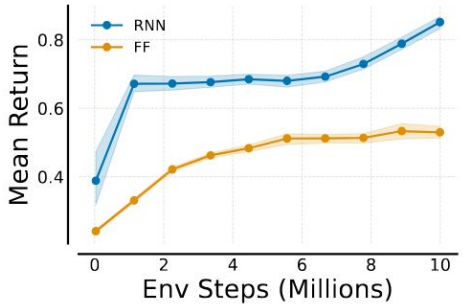


(f) SMAX ($\mathbb{I}(O, A)_{max} \approx 2.30$)



(g) SMAX ($\mathbb{I}(H, A)_{max} \approx 2.30$)

2. Recurrent reasoning about other agents. ✅



$$\mathbb{I}(O; A) < \mathbb{I}(H; A)$$

Summary

	MPE	SMAX v1	SMAX v2	MaBrax	Hanabi	Overcooked
Is PO relevant?	❌	✅	✅	❌	❌	
Is PO reliant on private information?	❌	❌	✅	❌	❌	
Is Coordination non-trivial?	❌	❌	✅			
Does Coordination require anticipating others?						

Summary

1. We will increasing be seeing more Multi-Agent systems in practise, therefore it is important to measure progress in these settings using environments that test properties we want.
2. Design of environments have implications in the what behaviours is easy to learn i.e. attractive optima.

Limitations & Next Steps

1. **Single algorithm** (IPPO) with MLP and RNN (GRU) architectures. Claims specific and not board. Include more Multi-Agent algorithms such as MAPPO & expand environments.
2. **Measures**:
   a. I(O,A) and I(H,A) - since obs and H are correlated, switch to conditional MI
      i. I(O,A | H) - policies grounded in obs
      ii. I(H,A | O) - are policies relying on memory?
   b. Distengle recurrent reasoning
      i. Used for reconstructing state or predicting other agents actions - use ideas from linear probing e.g. (Mon-Williams et al., 2025).
   c. Only measuring MI at end? We should measure over time. – Not fair!
3. What are properties of the dec-pomdp vs properties of the learning algorithm vs both?
4. Measures for coordination?

Questions &/or Comments

References

1. Tran, K.T., Dao, D., Nguyen, M.D., Pham, Q.V., O'Sullivan, B. and Nguyen, H.D., 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. arXiv preprint arXiv:2501.06322.
2. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T. and Han, T.A., 2025. Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.
3. Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S.R., Rocktäschel, T. and Perez, E., Debating with More Persuasive LLMs Leads to More Truthful Answers. In Forty-first International Conference on Machine Learning.
4. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B. and Mordatch, I., 2023. Improving factuality and reasoning in language models through multiagent debate. In Forty-first International Conference on Machine Learning.
5. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z. and Zhou, L., MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In The Twelfth International Conference on Learning Representations.
6. Sutton, R.S. and Barto, A.G., 1998. Reinforcement learning: An introduction (Vol. 1, No. 1, pp. 9-11). Cambridge: MIT press.
7. Oliehoek, F.A. and Amato, C., 2016. A concise introduction to decentralized POMDPs (Vol. 1). Cham, Switzerland: Springer International Publishing.