Remembering the Markov Property in Cooperative MARL

K. Tessera, L. Hinckeldey, R. Zamboni, D. Abel, A. Storkey
k.tessera@ed.ac.uk

The University of Edinburgh

Do current MARL environments test genuine cooperative reasoning, the kind that requires **behaviours grounded in observations** and **memory**?

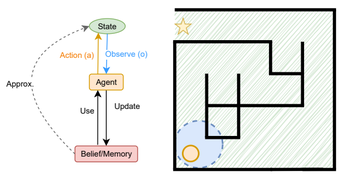
1) Problem Formulation

The Problem: Recovering a Markov Signal

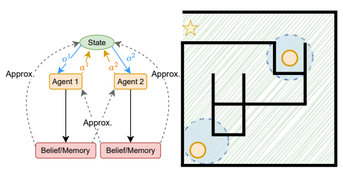
Setting: Dec-POMDP
$\mathcal{M} = (\mathcal{N}, \mathcal{S}, T, O, \mu, \{\mathcal{A}^i\}, \{\mathcal{O}^i\}, R, \gamma)$; agents receive partial o_t^i and act with $\pi^i(a_t^i \mid h_t^i)$.

Goal: From histories h_t^i, recover a *Markovian signal*: a belief/approx. over environment state and teammates' behaviour that is sufficient for control.

Question: Do current MARL environments require reconstructing this signal? Do MARL agents do this?



(a) **POMDP:** *If you can't see, you must remember*



(b) **Dec-POMDP:** *If you can't see, you must predict*

Mutual Information

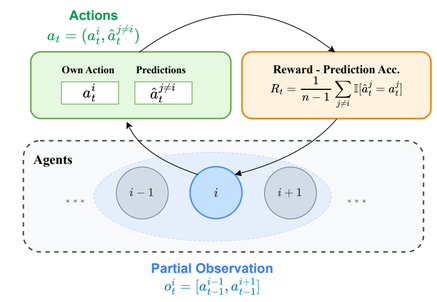
Observation-grounding:

- $\mathbb{I}(O; A)$: dependence between actions and current observations
- $\mathbb{I}(H; A)$: dependence between actions and agent history (e.g., RNN hidden state)

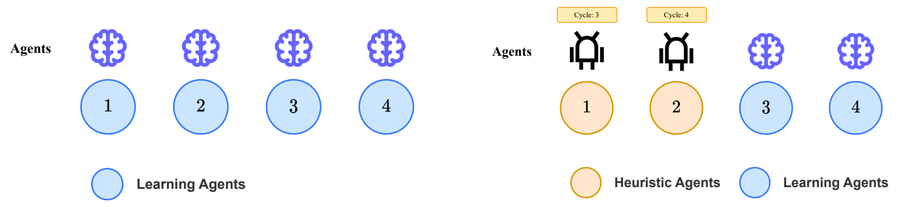
Memory:

- *Empirical test:* Performance with and without memory
- *Compare $\mathbb{I}(H; A)$ and $\mathbb{I}(O; A)$* If $\mathbb{I}(H; A) \mathbb{I}(O; A)$, agents rely more on memory than immediate observations.

2) Case Study: Brittle Conventions vs. Robust Coordination

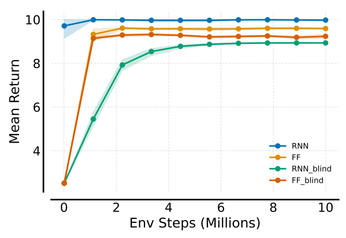


(a) **Prediction Game**

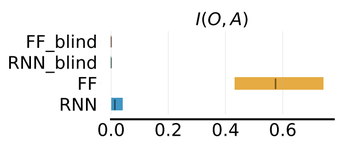


(b) **Scenario A: Co-adapting**

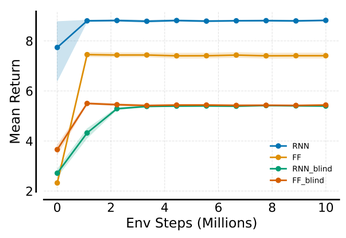
(c) **Scenario B: Mixed**



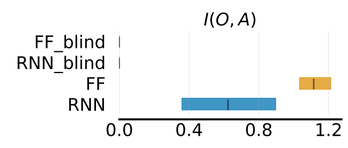
(a) Co-adapt: High Perf.



(b) **Low** $I(O, A)$



(c) Mixed: High Perf.

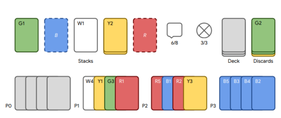


(d) **Higher** $I(O, A)$

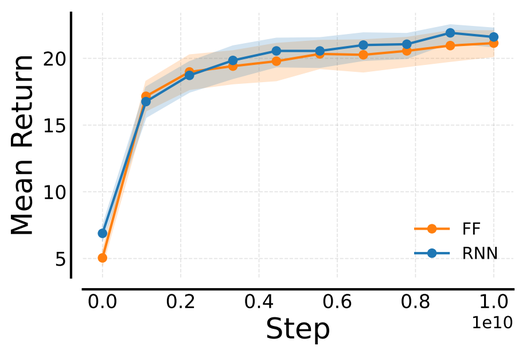
Brittle Conventions vs. Robust Coordination

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.

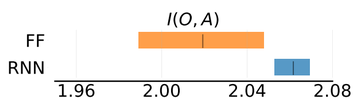
3) Experiments: Modern Environments



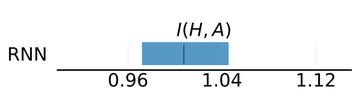
Hanabi



Hanabi (Two-Players).

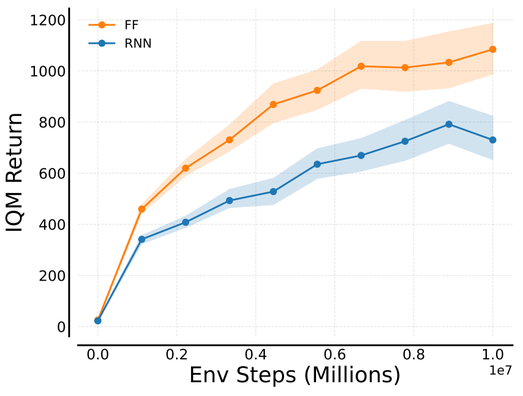


$\mathbb{I}(O; A)_{max} \approx 3$



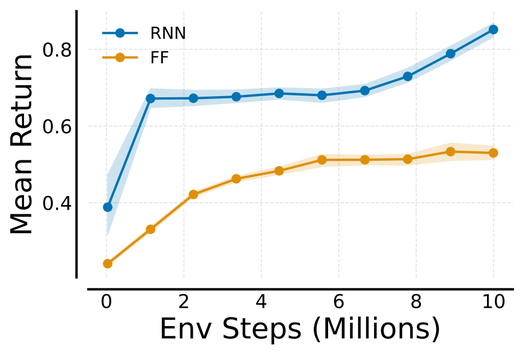
$\mathbb{I}(H; A)_{max} \approx 3$

MaBrax

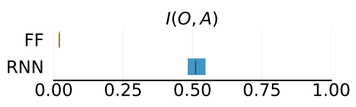


MaBrax Results (Across 5 settings)

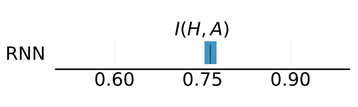
SMAX



SMAX-v2 style (5 Units).



$\mathbb{I}(O; A)_{max} \approx 2.30$



$\mathbb{I}(H; A)_{max} \approx 2.30$

Summary

Environment	Obs-Grounded	Memory-Based
Hanabi	✓	×
MaBrax	×	×
SMAX	✓ / ×	✓

Takeaways

- Modern model-free recurrent MARL methods **can** learn robust cooperative behaviour when environments **necessitate** this.
- Yet current MARL environments may inadvertently allow success through **alternative means** (e.g. blind conventions, memoryless coordination) rather than genuine cooperation.

We therefore advocate for new cooperative environments built upon two core principles: (1) **behaviours grounded in observations** and (2) **memory-based reasoning about other agents**, ensuring success requires *genuine multi-agent cooperative reasoning*.