# Probing Dec-POMDP Reasoning in Cooperative MARL 🔍

Kale-ab Tessera, Leonard Hinckeldey, Riccardo Zamboni, David Abel, Amos Storkey



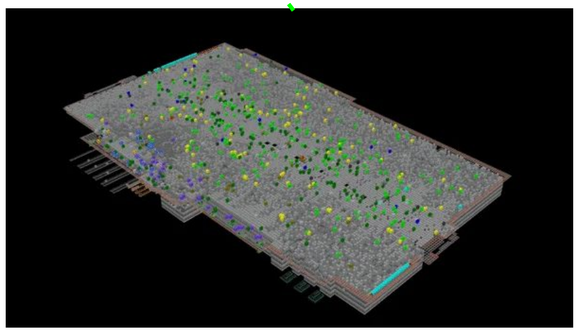


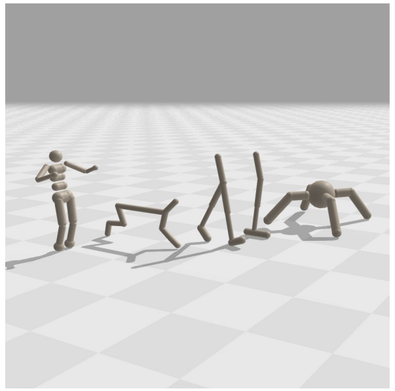
THE UNIVERSITY of EDINBURGH informatics

POLITECNICO MILANO 1863

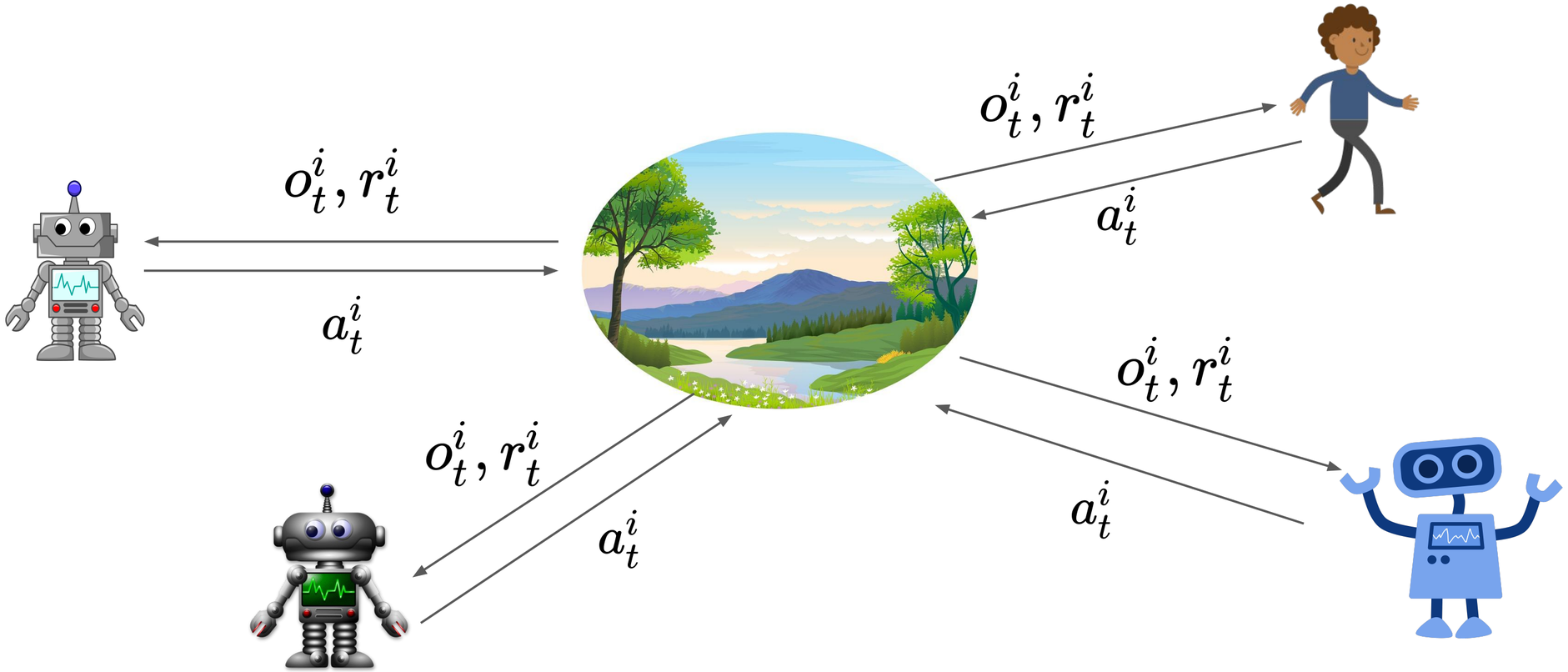
# **Motivation**: Build Useful Cooperative Multi-Agent Environments

Useful for.





# **Setting:** Cooperative MARL - **Dec-POMDP**



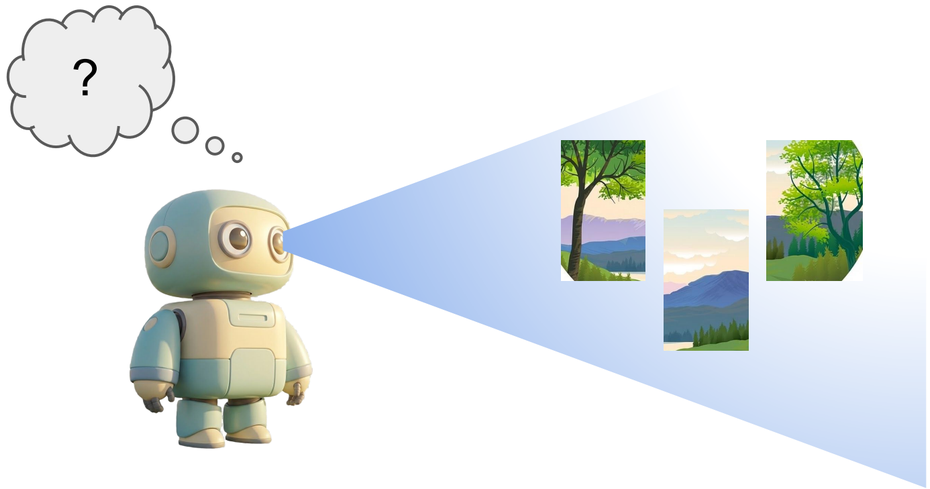

Goal:

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

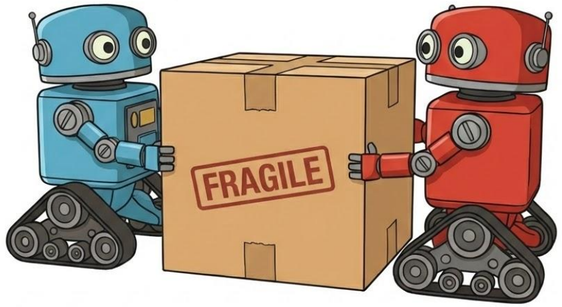
# **Probing** Dec-POMDP Reasoning in Cooperative MARL

**Dec-POMDP planning is NEXP-complete** (doubly-exponential worst-case)(Bernstein et al.).

**1. Partial observability**.




**2. Decentralised coordination.**



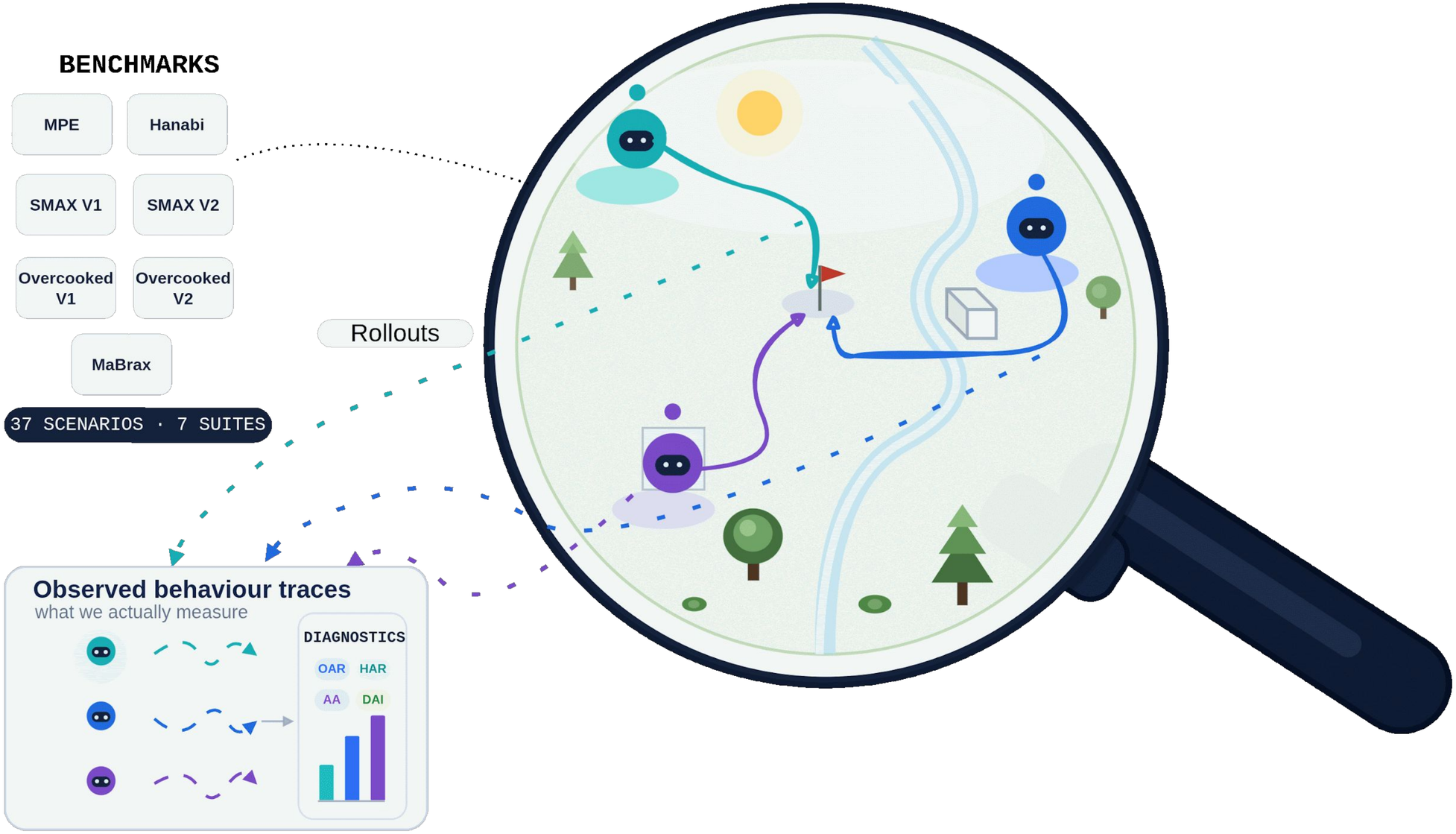

- Hardness drivers are also relevant in the ***real world***, where problems often involve ***partial observability*** and ***decentralisation***.

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. Mathematics of operations research 27, 4 (2002), 819–840.

# **Probing** Dec-POMDP Reasoning in Cooperative MARL

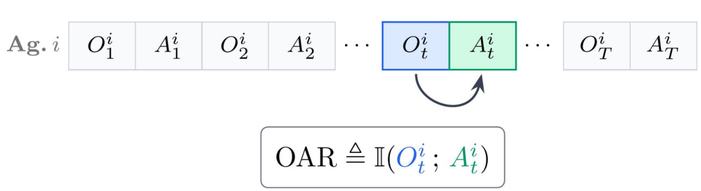
*Do modern cooperative MARL environments truly test the **Dec-POMDP properties that make these problems hard**,*
*or do they permit **success via strategies that bypass** them?*

# **Probing** Dec-POMDP Reasoning in Cooperative MARL

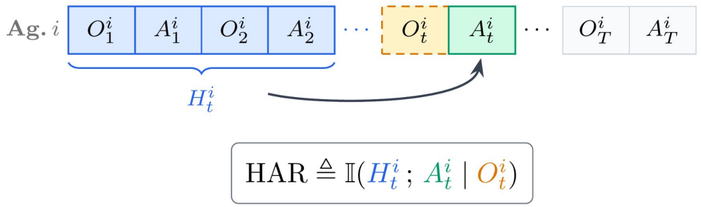


# Information-Theoretic Metrics

## Partial observability

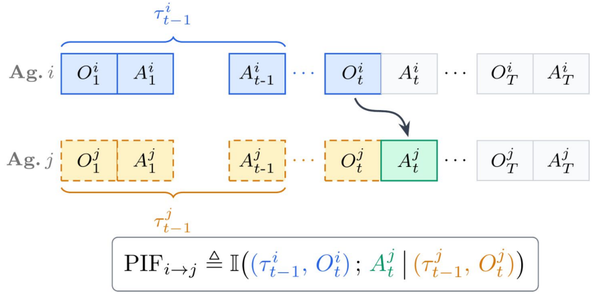



(a) Observation–Action Relevance (OAR)

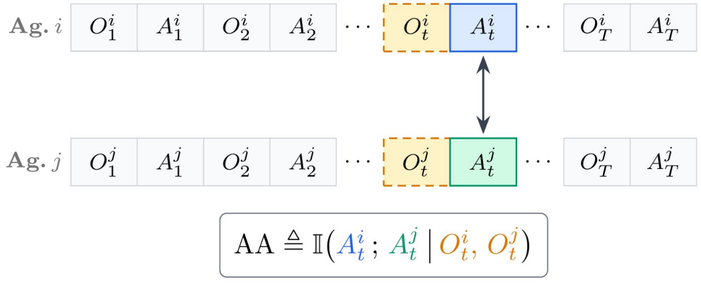



(b) History–Action Relevance (HAR)

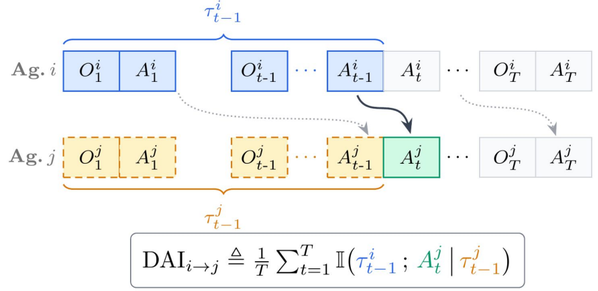
## Decentralised Coordination




(c) Private Information Flow (PIF)




(d) Action–Action Coupling (AA)




(e) Directed Action Information (DAI)

# Information Theoretic Metrics

| If High | Suggests… |
| --- | --- |
| High OAR + low HAR | C*onsistent with* ***reactive policies***. |
| High AA + low DAI | Consistent with ***instantaneous conventions*** or ***symmetry breaking***. |
| High PIF + DAI | *Consistent with* ***Dec-POMDP reasoning.*** |

# Results - Findings

**Table 1: Diagnostics of learned behaviour across cooperative MARL benchmarks. We report the share of scenarios (count/total) where trained policies satisfy our decision criteria (Sec. 6.1). Crucially, these reflect dependencies *induced by the policy* rather than strict environment requirements. Per-scenario metrics are detailed in App. D.**

| | MPE | SMAX V1 | SMAX V2 | MaBrax | Hanabi | Overcooked V1 | Overcooked V2 |
|---|---|---|---|---|---|---|---|
| Do agents benefit from memory? | 100% (3/3) | 100% (9/9) | 100% (3/3) | 20% (1/5) | 0% (0/1) | 0% (0/5) | 0% (0/11) |
| Does teammate info. help predict actions? | 100% (3/3) | 67% (6/9) | 67% (2/3) | 100% (5/5) | 0% (0/1) | 20% (1/5) | 82% (9/11) |
| Does synchronous coordination emerge? | 100% (3/3) | 44% (4/9) | 0% (0/3) | 60% (3/5) | 0% (0/1) | 100% (5/5) | 82% (9/11) |
| Does temporal coordination emerge? | 100% (3/3) | 67% (6/9) | 67% (2/3) | 100% (5/5) | 100% (1/1) | 40% (2/5) | 100% (11/11) |

# Results - Findings

1. ***History dependence ≠ utility.***
    - All policies show some history dependence, but only 43% need memory for high returns.

2. ***Hidden state vs. teammate info***.
    - Probes disentangle hidden state and hidden teammate information as separate difficulty drivers (e.g., Overcooked V1→V2).

# Results - Findings

3. ***Coordination structure varies.***
    - Coordination structure varies. Synchronous and temporal mechanisms dissociate across benchmarks.

4. ***Few benchmarks jointly test partial observability and coordination.***
    - MPE is the only suite in which every scenario satisfies all diagnostic criteria.

Many detected dependencies are above null baselines but still modest, leaving plenty of room to design better envs. 🔥

# Takeaway & Thanks

- 💡 Contribution:
  - **Diagnostic framework.** We introduce ***information-theoretic diagnostics*** *that reveal* ***how*** *policies solve tasks*, beyond raw returns.
  - **Systematic benchmark audit.** We audit ***37 popular MARL scenarios*** showing some admit shortcuts that bypass Dec-POMDP reasoning.
  - Open-source tooling and implications.

- ⚠️ Challenges & Limitations
  - **Policy-dependent probes** - so they characterise learned behaviour under a specific policy, and not worst-case or best-case properties of the environment.
  - Noise in estimators for different mutual information metrics - used null baselines to address this.

# Takeaway & Thanks!

🔍 Takeaway - These tools enable designing tasks where ***partial observability*** and ***coordination*** are non-optional.

```
pip install dec-pomdp-diagnostics

import dec_pomdp_diagnostics as dpd

data = dpd.UserData(
    observations  = {"agent_0": obs0, "agent_1": obs1},   # (N, obs_dim)
    actions       = {"agent_0": act0, "agent_1": act1},   # (N,) int
    timesteps     = {"agent_0": ts0,  "agent_1": ts1},    # (N,) int
    episode_ids   = {"agent_0": eps0, "agent_1": eps1},   # (N,) int
    hidden_states = {"agent_0": h0,   "agent_1": h1},     # optional, RNN only
    env_name="my_env", alg_name="IPPO_RNN", seed=0,
)

result = dpd.compute_diagnostics(data, history_k=3, null_reps=5)
print(result.describe())
# ✓  History dependence > null
# ✓  Does teammate information help predict actions?
# ×  Does synchronous coordination emerge?
# ✓  Does temporal coordination emerge?
```

Email: k.tessera@ed.ac.uk