Probing Dec-POMDP Reasoning in Cooperative MARL

Kale-ab Abebe Tessera
University of Edinburgh
Edinburgh, United Kingdom
k.tessera@ed.ac.uk

Leonard Hinckeldey
University of Edinburgh
Edinburgh, United Kingdom
l.hinckeldey@ed.ac.uk

Riccardo Zamboni
Politecnico di Milano
Milan, Italy
riccardo.zamboni@polimi.it

David Abel
University of Edinburgh
Edinburgh, United Kingdom
david.abel@ed.ac.uk

Amos Storkey
University of Edinburgh
Edinburgh, United Kingdom
a.storkey@ed.ac.uk

ABSTRACT

Cooperative multi-agent reinforcement learning (MARL) is typically framed as a decentralised partially observable Markov decision process (Dec-POMDP), a setting whose hardness stems from two key challenges: *partial observability* and *decentralised coordination*. Genuinely solving such tasks requires *Dec-POMDP reasoning*, where agents use history to infer hidden states and coordinate based on local information. Yet it remains unclear whether popular benchmarks actually demand this reasoning or permit success via simpler strategies. We introduce a diagnostic suite combining statistically grounded performance comparisons and information-theoretic probes to audit the behavioural complexity of baseline policies (IPPO and MAPPO) across 37 scenarios spanning MPE, SMAX, Overcooked, Hanabi, and MaBrax. Our diagnostics reveal that success on these benchmarks rarely requires genuine Dec-POMDP reasoning. Reactive policies match the performance of memory-based agents in over half the scenarios, and emergent coordination frequently relies on brittle, synchronous action coupling rather than robust temporal influence. These findings suggest that some widely used benchmarks may not adequately test core Dec-POMDP assumptions under current training paradigms, potentially leading to over-optimistic assessments of progress. We release our diagnostic tooling to support more rigorous environment design and evaluation in cooperative MARL.[1]

KEYWORDS

Multi-Agent Reinforcement Learning, Cooperative Multi-Agent Reinforcement Learning, Dec-POMDPs

ACM Reference Format:
Kale-ab Abebe Tessera, Leonard Hinckeldey, Riccardo Zamboni, David Abel, and Amos Storkey. 2026. Probing Dec-POMDP Reasoning in Cooperative MARL. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 25 pages. https://doi.org/10.65109/ECCJ1033

[1]The code is available at https://github.com/KaleabTessera/probing-dec-pomdps.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/ECCJ1033

1 INTRODUCTION

The widespread deployment of autonomous multi-agent systems is bounded by their ability to coordinate under uncertainty. In such settings, no single agent possesses a complete view of the world, yet outcomes depend on joint behaviour. This tension lies at the heart of cooperative multi-agent reinforcement learning [MARL, 2]. The standard formalism for these problems, decentralised partially observable Markov decision processes [Dec-POMDPs, 5, 21], capture this intrinsic hardness through two fundamental characteristics: *partial observability*, where agents cannot directly observe the full global state, and *decentralised coordination*, where agents must cooperate based on local and private information.

The intrinsic hardness of this setting stems directly from the interaction of these two factors. In principle, to act optimally, each agent must recover a *Markovian signal* by maintaining a *multi-agent belief* over the joint state and the policies (or histories) of other agents [21]. However, exact multi-agent belief computation is typically infeasible [5]. Consequently, practical model-free methods approximate this reasoning using finite-memory or recurrent policies (e.g., GRUs) [13], often instantiated within the *centralised training with decentralised execution* [CTDE, 17, 22] paradigm to leverage extra information during learning.

The empirical success of MARL approaches in benchmarks [among others, 23, 31] is often interpreted as evidence that practical approximations (e.g., recurrent policies) effectively capture the Dec-POMDP reasoning these problems demand. *We challenge this interpretation.* High returns can mask a failure to learn the underlying coordination challenge, as agents may exploit reactive shortcuts permitted by the task design rather than employing genuine history-based reasoning. This distinction is critical. If valid solutions exist that ignore the theoretical challenges of partial observability and coordination, then the environment can become a weak proxy for the Dec-POMDP formalism, yielding an illusion of progress on coordination under uncertainty. We therefore use trained policies as *diagnostic probes* to ask:

> *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?*

To answer this, we introduce a suite of *MARL diagnostics* that couple statistically grounded performance comparisons with information-theoretic probes to measure history dependence, private information flow, synchronous action coupling, and directed temporal influence. Together, these reveal whether learned policies genuinely employ Dec-POMDP reasoning, or bypass it entirely.

We apply these diagnostics to policies learned by standard baselines in *37 popular MARL scenarios*, across MPE [19], SMAX[2] [26], Overcooked (V1 and V2) [6, 11], Hanabi [4] and MaBrax [24, 26]. Across these settings, our analysis reveals three main takeaways: **(i)** history dependence rarely translates to history utility—while all learned policies encode some history dependence, only 43% actually need memory to achieve high returns, indicating that current observations often suffice for strong performance; **(ii)** hidden environment state and hidden teammate information act as separate drivers of difficulty, which our metrics successfully disentangle (e.g., empirically validating the design shift from Overcooked V1 to V2); and **(iii)** while coordination is common, its structure is highly variable—synchronous and temporal mechanisms dissociate across benchmarks. Notably, MPE emerges as the only suite where every scenario satisfies all four diagnostic criteria, consistently requiring both meaningful history use and decentralised coordination.

Ultimately, these findings suggest that, under current training paradigms, success on popular benchmarks often does not require the Dec-POMDP reasoning these tasks are intended to evaluate.

Contributions.

(1) **Diagnostic framework.** We introduce information-theoretic probes – measuring history dependence, private information flow, synchronous action coupling, and directed temporal influence – that audit whether learned policies actually exhibit Dec-POMDP reasoning, beyond what raw returns reveal.
(2) **Systematic benchmark audit.** We evaluate 37 scenarios across seven benchmark suites, revealing that history dependence is ubiquitous but rarely performance-critical, coordination structures vary qualitatively across domains, and few environments jointly test both partial observability and coordination.
(3) **Open-source tooling and implications.** We release diagnostic tools for researchers to audit their own environments, and discuss implications for designing tasks where partial observability and coordination are non-optional.

2 BACKGROUND

We introduce key concepts that will be needed throughout the paper.

Interaction Protocol. As a base model for interaction, we consider a discounted Dec-POMDP [5], defined by the tuple $\mathcal{M} = (\mathcal{N}, \mathcal{S}, \mathbb{T}, \mathbb{O}, \mu, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \{O^i\}_{i\in\mathcal{N}}, R, \gamma)$. Here, $\mathcal{N}$ is the set of $N \in \mathbb{N}$ agents and $\mathcal{S}$ is the set of global states. At each time step t, the system is in some state $s_t \in \mathcal{S}$. Each agent $i \in \mathcal{N}$ selects an action $a_t^i \in \mathcal{A}^i$, forming a joint action $\mathbf{a}_t = (a_t^1, \ldots, a_t^N)$ in the joint action space $\mathcal{A} = \times_{i=1}^N \mathcal{A}^i$. This action leads to a state transition according to the probability function $\mathbb{T}(s_{t+1}|s_t, \mathbf{a}_t)$ and a shared reward $R(s_t, \mathbf{a}_t)$. Agents do not observe the global state s_t, instead they receive a local observation $o_t^i \in O^i$. The joint observation $\mathbf{o}_t$ is drawn according to the observation function $\mathbb{O}(\mathbf{o}_t|s_t, \mathbf{a}_{t-1})$. The goal is to learn a joint policy $\boldsymbol{\pi}$ at which no agent has any incentive to deviate, while maximising the expected discounted return $\mathbb{E}_{\mathbf{a}_t\sim\boldsymbol{\pi},\mathcal{M}}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t, \mathbf{a}_t)\right]$. These solution concepts are usually described through various notions of *equilibria*: we report a brief description in Appendix A.

Mutual Information. To study the information embedded in agents' policies, we propose metrics based on mutual information (MI). For two discrete random variables X and Y with joint probability mass function[3] $p(x, y)$ and marginals $p(x)$, $p(y)$, we can measure MI as follows:

$$\mathbb{I}(X;Y) = H(X) - \mathbb{H}(X \mid Y) = \mathbb{H}(Y) - \mathbb{H}(Y \mid X), \tag{1}$$

$$= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{2}$$

where H is the Shannon entropy. Intuitively, $\mathbb{I}(X;Y)$ is the average amount of information that X conveys about Y, or vice versa. MI is symmetric and non-negative, and $\mathbb{I}(X;Y) = 0$ iff X and Y are independent.

We will also use metrics based on conditional mutual information (CMI), $\mathbb{I}(X;Y \mid Z)$. Intuitively, CMI measures the extra information that X tells us about Y, excluding what we know about Y given Z. $\mathbb{I}(X;Y \mid Z) = 0$ iff X and Y are conditionally independent given Z.

3 RELATED WORK

Benchmarking Partial Observability. Ellis et al. [9] found that many SMAC [27] maps admit open-loop solutions that ignore local observations. While they redesigned these maps to enforce "meaningful partial observability", they provided no metric to quantify it. In single-agent RL, Tao et al. [28] formalised *memory improvability* based on performance gaps between agents with access to more or less state information. Our framework provides quantitative tools for the multi-agent case, moving beyond raw performance metrics. We disentangle history dependence, private information flow, and coordination as separate dimensions of Dec-POMDP difficulty.

Conventions. Co-trained agents typically develop conventions that are efficient but arbitrary and brittle when paired with unfamiliar partners [10, 15]. Prior work shows that grounding these conventions in observations makes coordination more robust [14]. Our AA and DAI diagnostics explicitly quantify these dynamics, disentangling instantaneous, ungrounded conventions from coordination that is temporally responsive to a partner's trajectory.

4 PROBING DEC-POMDPS

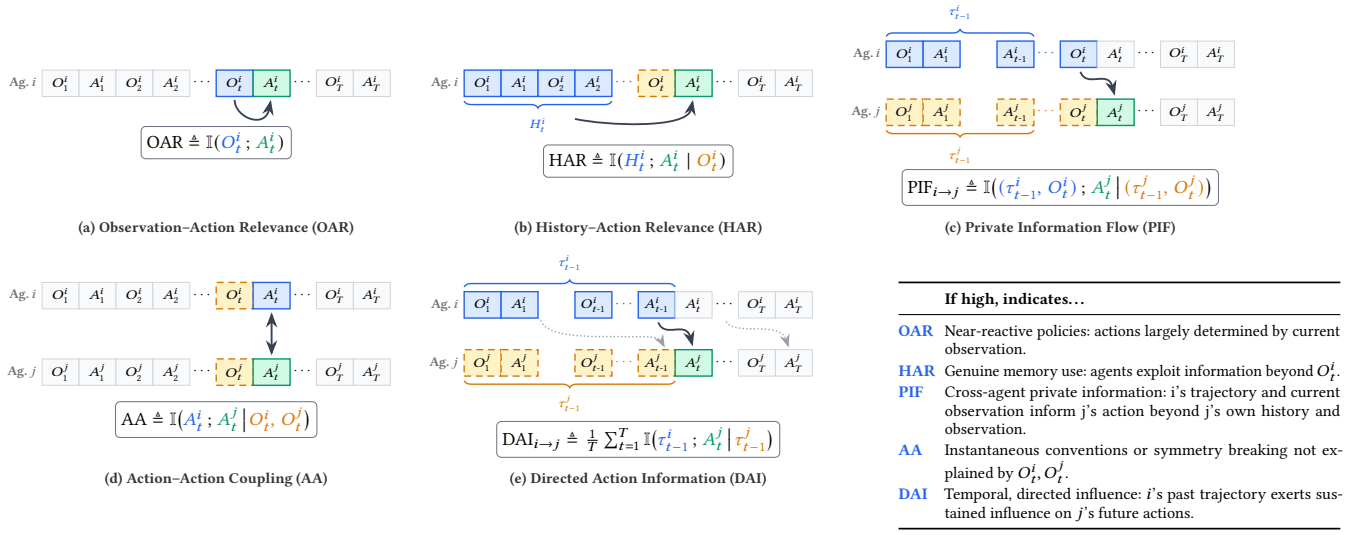
To probe the reasoning demands specific to MARL environments, we focus on two core properties of Dec-POMDPs – *partial observability* and *decentralised coordination*. While the interaction of these factors renders the general problem class NEXP-complete[4], theoretical worst-case hardness does not necessarily imply practical difficulty in specific benchmarks.

Our goal is therefore to characterise these properties *functionally*, measuring them only *as they matter for solving a task*. Consequently,

[2]Both SMAC-V1 [27] and SMAC-V2 [9] maps were tested.

[3]For continuous variables, we use the probability density function.

[4]The worst case complexity of DEC-MDPs is the same as Dec-POMDPs [5], as such hardness comes from decentralisation as well, and not (only) from the presence of hidden states.



(a) Observation–Action Relevance (OAR) (b) History–Action Relevance (HAR) (c) Private Information Flow (PIF)

(d) Action–Action Coupling (AA) (e) Directed Action Information (DAI)

	If high, indicates…
OAR	Near-reactive policies: actions largely determined by current observation.
HAR	Genuine memory use: agents exploit information beyond O_t^i.
PIF	Cross-agent private information: i's trajectory and current observation inform j's action beyond j's own history and observation.
AA	Instantaneous conventions or symmetry breaking not explained by O_t^i, O_t^j.
DAI	Temporal, directed influence: i's past trajectory exerts sustained influence on j's future actions.

Figure 1: Summary of information-theoretic diagnostics. Colours denote source (blue), target (green), and conditioning (amber, dashed) variables. All quantities are expectations under the converged joint policy p^{π}; O_t^i is agent i's observation, A_t^i its action, H_t^i the local history (RNN hidden state or finite window), τ_{t-1}^i agent i's action–observation history up to $t-1$, and T the episode horizon.

we define every diagnostic as an expectation under the trajectory distribution of a joint policy $\boldsymbol{\pi}$ after convergence. We do not define *purely structural* properties of Dec-POMDPs independent of behaviour, but rather, we quantify *the specific reasoning capabilities necessitated by the task*. Figure 1 presents a summary and interpretation of our proposed diagnostic measures, and we discuss the technical details in the following sections.

4.1 Partial Observability

Is Partial Observability Relevant?

While many environments are *structurally* partially observable (states are hidden), this does not guarantee that the missing information is *functionally relevant* to solving a task. For example, the hidden state may not affect the rewards or dynamics, or it may be redundant given the current observations.

We are therefore interested in identifying when partial observability strictly affect success. If a task requires memory, it confirms that immediate observations are insufficient and that history contains decision-relevant information. Therefore, we measure this using *history dependence*.

Definition 4.1 (Relevant Partial Observability). An environment exhibits relevant partial observability if:

(i) memory-based agents (π_{RNN}) outperform reactive agents (π_{FF}) under matched training conditions; and
(ii) learned policies actively exploit history, rather than relying solely on immediate observations.

This definition requires that memory is both *beneficial* (producing higher returns) and *active* (influencing decisions). We quantify this with a performance diagnostic, and with two complementary information-theoretic probes.

Diagnostic 1 (Memory–Reactive Gap). We test whether memory results in a performance gain by comparing recurrent and feed-forward policies. For matched training runs (sharing seed, environment, and algorithm), let $J(\pi)$ denote the mean evaluation return. We define the paired performance gap as:

$$\Delta_{\text{Mem}} \triangleq J(\pi_{\text{RNN}}) - J(\pi_{\text{FF}}).$$

We test $\mathbf{H_0}$: $\text{median}(\Delta_{\text{Mem}}) \leq 0$ vs. $\mathbf{H_1}$: $\text{median}(\Delta_{\text{Mem}}) > 0$ using a one-sided Wilcoxon signed-rank test [30] over the paired differences. A significant result ($p < 0.05$) indicates a reliable performance advantage from memory under matched training.

Diagnostic 2 (History–Action Relevance (HAR)). We quantify memory use beyond the current observation via conditional mutual information:

$$\text{HAR} \triangleq \mathbb{I}(H_t^i; A_t^i \mid O_t^i), \qquad \text{HAR}^{\text{norm}} \triangleq \frac{\text{HAR}}{\mathbb{H}(A_t^i \mid O_t^i)} \in [0,1]. \quad (3)$$

Here, H_t^i denotes the agent's history representation, for reactive policies, $H_t^i = O_{t-k:t-1}^i$ (a length-k window excluding O_t^i), and for recurrent policies, H_t^i is the RNN hidden state.

Diagnostic 3 (Observation–Action Relevance (OAR)). We quantify reactivity by measuring how informative the current observation is about the agent's action:

$$\text{OAR} \triangleq \mathbb{I}(O_t^i; A_t^i), \qquad \text{OAR}^{\text{norm}} \triangleq \frac{\text{OAR}}{\mathbb{H}(A_t^i)} \in [0,1]. \quad (4)$$

High OAR^{norm} indicates that A_t^i is largely predictable from the current observation O_t^i (i.e., near-reactive behaviour). Conversely, low OAR^{norm} combined with high HAR^{norm} provides evidence that history contributes information for selecting A_t^i beyond what is contained in O_t^i.

Is Partial Observability Reliant on Private Information?

The previous diagnostics measure whether agents benefit from *history* or *memory*, which acts as a behavioural proxy for functionally relevant partial observability. Crucially, history dependence alone does not imply that the hidden information is relevant for *coordination*. An agent may use its history only to infer latent *environment* state, as in single-agent POMDPs [3, 16], even if this provides no additional information about coordinating with teammates.

We therefore introduce a cross-agent diagnostic that quantifies whether the private information of one agent helps predict the actions of another. This metric is related to the intuition behind *meaningful partial observability* [9], where hidden information observed by one agent is critical for the actions of another. Such cross-agent information asymmetries are central to the hardness of Dec-POMDPs [5].

Diagnostic 4 (Private Information Flow (PIF)). We measure how much additional information agent i's history provides about agent j's action, beyond what is already contained in j's own history. We define this using conditional mutual information:

$$\begin{aligned} \mathrm{PIF}_{i\to j} &\triangleq \mathbb{I}(\tau^i_{t-1}, O^i_t;\, A^j_t \mid \tau^j_{t-1}, O^j_t), \\ \mathrm{PIF}^{\mathrm{norm}}_{i\to j} &\triangleq \frac{\mathrm{PIF}_{i\to j}}{\mathbb{H}(A^j_t \mid \tau^j_{t-1}, O^j_t)} \in [0,1]. \end{aligned} \tag{5}$$

Here, τ_{t-1} denotes an agent's action-observation history[5]. We explicitly condition on the current observations O^i_t, O^j_t alongside the past τ_{t-1} to capture information asymmetries at decision time.

$\mathrm{PIF}_{i\to j}$ quantifies how much information about A^j_t is contained in agent i's trajectory that is *not* already captured by agent j. $\mathrm{PIF}^{\mathrm{norm}}_{i\to j}$ rescales this as the fraction of agent j's residual action uncertainty (given its own history τ^j_{t-1} and observation O^j_t) that is explained by agent i.

4.2 Decentralised Coordination

The previous diagnostics quantify whether *hidden information* is relevant to decision-making, specifically, whether agents require memory of local state (HAR; Diagnostic 2) or access to a teammate's private information (PIF; Diagnostic 4). However, they do not characterise the *form* of coordination that emerges in the joint behaviour induced by the converged policies (if any). We therefore introduce coordination probes that separate instantaneous action coupling from temporally extended, more directional dependence.

Is Coordination Synchronous?

Diagnostic 5 (Action–Action Coupling (AA)). We quantify *instantaneous* action dependence via the coupling of actions at time t:

$$\mathrm{AA} \triangleq \mathbb{I}\left(A^i_t; A^j_t \mid O^i_t, O^j_t\right), \quad \mathrm{AA}^{\mathrm{norm}} \triangleq \frac{\mathrm{AA}}{\mathbb{H}\left(A^j_t \mid O^i_t, O^j_t\right)}. \tag{6}$$

AA measures same-timestep dependence between agents' actions beyond what their current observations explain, and $\mathrm{AA} > 0$ is consistent with symmetry breaking or instantaneous conventions (e.g., agents taking distinct roles such as heading to different landmarks).

[5]In practice, we approximate τ_{t-1} using the RNN hidden state (for recurrent policies) or a finite window of size k, $(O^i_{t-k:t-1}, A^i_{t-k:t-1})$ (for reactive policies).

Is Coordination Temporally Responsive?

AA alone cannot distinguish *task-driven role differentiation* from *arbitrary, ungrounded conventions*, as it detects instantaneous coupling beyond shared observations, but cannot distinguish static conventions (e.g., fixed roles) from agents adapting to evolving partner behaviours.

To probe this *temporally extended, directional dependence*, we test whether agent i's *past* provides additional predictive information about agent j's *current* action, conditioned on agent j's own history. While a lagged AA could measure this, it would rely on fixed windows that are brittle to unknown or variable delays. We instead use *Directed Information* [20], which aggregates directional cross-timestep dependence over the episode, capturing dependencies regardless of the temporal lag.

Diagnostic 6 (Directed Action Information (DAI)). We measure the average directional, cross-timestep dependence from agent i to agent j as follows:

$$\begin{aligned} \mathrm{DAI}_{i\to j} &\triangleq \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left(\tau^i_{t-1}; A^j_t \middle| \tau^j_{t-1}\right), \\ \mathrm{DAI}^{\mathrm{norm}}_{i\to j} &\triangleq \frac{\mathrm{DAI}_{i\to j}}{\frac{1}{T}\sum_{t=1}^{T} \mathbb{H}\left(A^j_t \middle| \tau^j_{t-1}\right)} \in [0,1]. \end{aligned} \tag{7}$$

Here, τ^i_{t-1} is agent i's action-observation history up to $t-1$, including A^i_{t-1}, the last act before agent j selects A^j_t. Conditioning on τ^j_{t-1} controls for what is already predictable from agent j's own past, so $\mathrm{DAI}_{i\to j} > 0$ indicates that agent i's past carries additional predictive information about agent j's current action. Unlike PIF, which includes current observations to capture information at decision time, DAI conditions only on the causal past (the trajectory completed before j's action), isolating strictly temporal, directional influence.

5 CASE STUDY: HOW OBSERVATION STRUCTURE SHAPES BEHAVIOUR

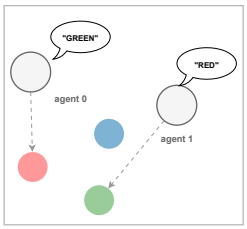
Multi-Particle Environments (MPE) [19] provide a controlled testbed with differing observation and communication structures. We examine three cooperative tasks—*Simple Reference*, *Speaker–Listener* and *Simple Spread*—using our diagnostics (Section 4) and MAPPO.

Performance. We see from Tbl. 2d, recurrent policies (RNN) outperform feed-forward (FF) baselines in all three tasks ($p < 0.05$, one-tailed Wilcoxon), confirming that memory provides a reliable advantage across MPE.

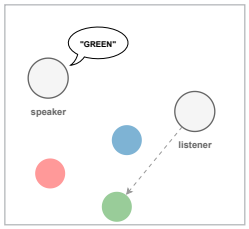
What the diagnostics reveal. Viewing MPE through our diagnostics shows that learned behaviour varies sharply across tasks, not because the algorithm changes, but because the observation/communication structure does.

Simple Reference (Fig. 2a). In *Simple Reference*, two agents move and observe the *other's* goal alongside a rich communication channel ($\dim_c$=10). Goal information is thus redundantly available at every timestep, reducing the need for history: $\mathrm{HAR}^{\mathrm{norm}}$ is the lowest across tasks and declines over training (≈ 0.06, Fig. 3b), and PIF/DAI remain low (Fig. 3c, 3d).

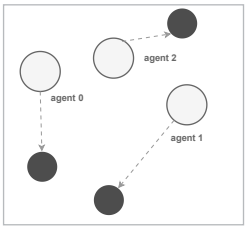
Speaker–Listener (Fig. 2b). In this scenario, a stationary speaker observes a hidden goal and must guide a listener that receives



(a) Simple Reference



(b) Speaker Listener

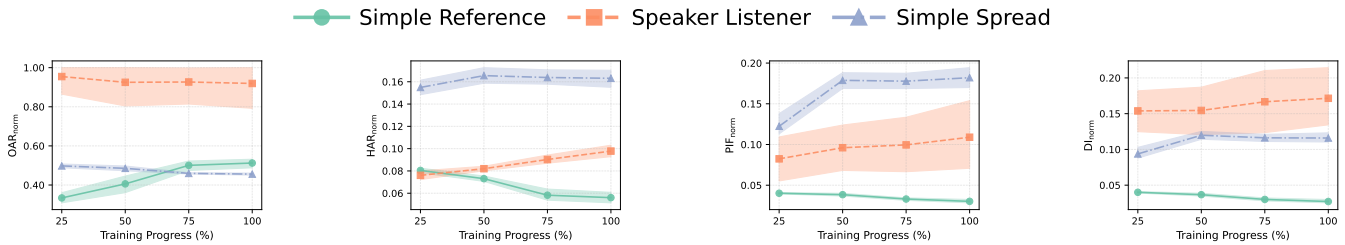


(c) Simple Spread

Scenario	Δ (RNN−FF)
Simple Reference	**6.50**
Speaker Listener	**14.84**
Simple Spread	**2.50**

(d) Δ_{Mem}; bold: $p<0.05$.

Figure 2: MPE tasks and per-environment performance deltas (RNN−FF). Memory improves performance across tasks.



(a) Observation–Action Relevance (OAR)

(b) History–Action Relevance (HAR)

(c) Private Information Flow (PIF)

(d) Directed Action Information (DAI)

Figure 3: Evolution of diagnostic metrics *during* training in MPE with recurrent MAPPO (mean and 95% CI): (a) Observation–Action Relevance, (b) History–Action Relevance, (c) Private Information Flow, and (d) Directed Action Information. For the same algorithm and training paradigm, environment modifications can have a large impact on the kinds of behaviour learned.

no goal information except through a narrow message channel ($\dim_c$=3). This dependency produces the highest DAI^{norm} across tasks (> 0.15, Fig. 3d), reflecting sustained directional influence from speaker to listener. HAR^{norm} rises over training but remains moderate (≈ 0.10), suggesting that the listener's history use, while present, is secondary to the cross-agent information channel.

Simple Spread (Fig. 2c). Here, agents must cover distinct landmarks without explicit communication. Consequently, HAR^{norm} and PIF^{norm} are the highest across tasks (Fig. 3b, Fig. 3c), indicating that agents condition on each other's private trajectories to avoid overlapping landmarks. DAI^{norm} is also substantial (≈ 0.12), confirming coordination is both temporally extended and reliant on private information.

Takeaway. The form of coordination that emerges is shaped primarily by **information bottlenecks in the environment**. When task-relevant information is fully available at each timestep (*Simple Reference*), agents default to **reactive behaviour** despite having recurrent architectures. Conversely, when information is restricted, e.g., funnelled through a narrow channel (*Speaker–Listener*) or left implicit in a partner's trajectory (*Simple Spread*), agents develop **qualitatively different coordination structures**: higher directional influence in the former, and higher private information flow in the latter.

6 RESULTS

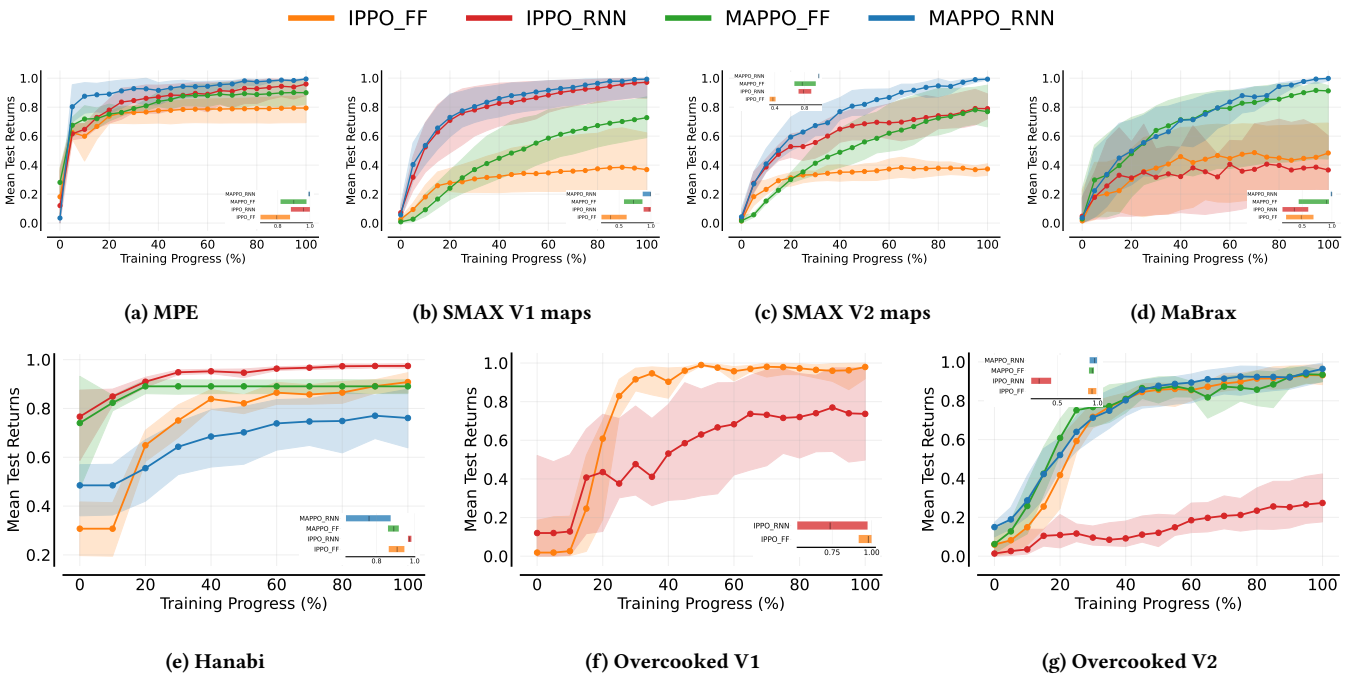
We apply our diagnostics (Sec. 4) to widely used cooperative MARL benchmarks, using learned policies as probes of *partial observability* and *decentralised coordination* as they arise in behaviour. Concretely, we ask a fundamental question: do these tasks genuinely elicit *Dec-POMDP reasoning*, where agents exploit history to infer decision-relevant hidden states and coordinate based on private information, or do they permit solutions that largely bypass these demands?

Experimental Setup. We evaluate 37 scenarios across MPE [19], SMAX (V1 maps and V2-style maps) [26, 27], Overcooked (V1 and V2) [6, 11], Hanabi [4] and MaBrax [24, 26].

Evaluation Protocol. We train with 10 seeds, matching original training budgets, and evaluate every 5% of training (mean evaluation return over 32 episodes) [12]. For aggregate comparisons, we report min–max normalised interquartile mean (IQM) with 95% stratified bootstrap CIs [1]. Hyperparameters are tuned per scenario, full details in App. A.1.

Algorithms. We use Independent PPO [IPPO, 8] and Multi-Agent PPO [MAPPO, 31] as they are widely used MARL baselines. We treat them as two training paradigms: IPPO uses independent critics, whereas MAPPO uses a centralised critic. Additionally, we compare feed-forward (FF) and recurrent (RNN) policies to study the role of memory and temporal information flow in these settings. Finally, to avoid confounders from optimisation and representation

[5]MAPPO is omitted for Overcooked V1 as full observability renders a centralised critic redundant.



(a) MPE (b) SMAX V1 maps (c) SMAX V2 maps (d) MaBrax

(e) Hanabi (f) Overcooked V1 (g) Overcooked V2

Figure 4: Sample efficiency of IPPO and MAPPO across diverse MARL benchmarks. We show the min-max normalised interquartile mean (IQM) with 95% stratified bootstrap confidence intervals (CIs). Detailed plots in App. B.

choices associated with shared weights in heterogeneous tasks [7, 29], we do not use parameter sharing in any baseline.

6.1 Diagnostic Probes

To answer the questions from Section 4, we use a two-stage protocol. First, we compute diagnostics on *converged* policies. Then, we determine whether each value reflects genuine structure or finite-sample noise by comparing against a *permutation null baseline*.

Permutation null baselines. Information-theoretic estimators (e.g., kNN/KSG [18, 25]) can exhibit bias when working with finite samples, resulting in non-zero values even under independence. We therefore construct an empirical null by independently permuting each agent's action sequence within each episode, which destroys temporal and cross-agent dependencies while preserving each agent's marginal action distributions. We recompute each diagnostic on the permuted data and deem the result meaningful only if its value on the original trajectories *exceeds* the mean of the corresponding permutation baseline.

Aggregation. We apply a two-stage aggregation to probe for the emergence of Dec-POMDP reasoning capabilities. First, within each run, we compute the *maximum* diagnostic value across agents, asking whether *any* agent exhibits the property. Second, we maximise across training configurations (IPPO/MAPPO × FF/RNN) to determine if *any* algorithm elicits the behaviour. This returns a conservative, per-scenario verdict: a property is flagged as absent only if no agent under any tested paradigm displays it.

Decision Rules. We now translate the conceptual questions from Section 4 into concrete decision rules, evaluating partial observability and coordination directly through agent behaviour.

Decision Rule 1 (Do agents benefit from memory?)**.** *Following Definition 4.1, agents* benefit from memory *iff both:*

(1) ***Significant performance gap:*** *The memory–reactive gap* Δ_{Mem} *is significant (one-tailed Wilcoxon signed-rank,* $p < 0.05$*), see Diagnostic 1.*

(2) ***Meaningful history use:*** *Under the memory-based policy,* $\mathrm{HAR}^{\mathrm{norm}}$ *exceeds its permutation null baseline, see Diagnostic 2.*

Criterion (1) establishes a reliable performance advantage from memory, while criterion (2) confirms that this advantage reflects active use of history rather than other confounding factors, such as optimisation dynamics.

Decision Rule 2 (Do agents use hidden teammate information?)**.** *Agents* use hidden teammate information *iff* $\mathrm{PIF}^{\mathrm{norm}}$ *exceeds its permutation null baseline, indicating that agent i's trajectory and observation inform agent j's action beyond agent j's own history (Diag. 4).*

Decision Rule 3 (Does synchronous coordination emerge?)**.** *Instantaneous, synchronous coordination* emerges *iff* $\mathrm{AA}^{\mathrm{norm}}$ *exceeds its permutation null baseline, indicating coupling beyond shared observations (Diag. 5).*

Decision Rule 4 (Does temporal coordination emerge?)**.** *Temporal, directional coordination* emerges *iff* $\mathrm{DAI}^{\mathrm{norm}}$ *exceeds its permutation null baseline, indicating genuine causal influence from past actions (Diag. 6).*

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)
Do agents use hidden teammate information?	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)

6.2 The Relevance of Partial Observability

How often does memory really matter?

Applying Decision Rule 1, we find that memory-based policies yield a statistically significant performance advantage in **43.2%** (16/37) of tested scenarios ($\Delta_{\text{Mem}} > 0$; see Tbls. 1, 14, and Fig. 4). However, we observe a clear dissociation between history *dependence* and *utility*. HAR^{norm} exceeds its permutation null in all 37 scenarios (App. Tbl. 15), confirming that trained policies universally encode *some* history dependence, yet this dependence translates into a measurable performance gain in less than half of the cases. Hanabi illustrates this disconnect. Despite being a canonical partially observable task, the memory–reactive gap is not significant under our baselines ($\Delta_{\text{Mem}} = 0.279$, Tbls. 1, 14), as IPPO/MAPPO fail to meaningfully exploit recurrent architectures to improve performance on this task (Fig. 11).

This suggests that much of the observed history dependence could be redundant, i.e., policies learn to track past information that offers no functional advantage over current observations O_t^i. Consequently, to genuinely test Dec-POMDP reasoning, environments should ensure decision-relevant information is *exclusively* available through history, rendering reactive policies insufficient.

Is partial observability reliant on private information?

From applying Decision Rule 2, we find that PIF^{norm} exceeds its permutation null in **70.3%** (26/37) of tested scenarios (Tbl. 1). Notably, many of these are *not* the same scenarios flagged by the HAR criterion, confirming that hidden environment state and hidden teammate information are distinct drivers of difficulty that our metrics can successfully disentangle (App. Tbl. 15).

This separation is especially visible in Overcooked. Overcooked V1 is fully observable and triggers PIF in only **20%** of layouts, while Overcooked V2, which introduces hidden teammate information by design [11], rises to **82%**. This serves as an external validation of our diagnostic, as PIF recovers the design intentions of the environment authors. SMAX V2 maps, following SMACv2, were similarly motivated by "meaningful partial observability" [9], however, PIF is detected in 67% of both V1 and V2 maps. This suggests that, at least under current baselines, several V1 maps already exhibit meaningful cross-agent information flow, and the redesign may not have widened this gap as intended.

6.3 Decentralised Coordination

Synchronous vs. Temporal coordination.

Decision Rules 3 and 4 probe two distinct coordination mechanisms. Synchronous coordination (AA) captures instantaneous action coupling conditioned on current observations, and **64.9%** (24/37) of scenarios exceed the null permutation. While high AA indicates action-action dependence, this coupling can be brittle, e.g. when it reflects rigid, ungrounded conventions that do not generalise [15]. Nonetheless, it remains a signature of coordination.

Directed Action Information (DAI), by contrast, measures temporal influence between agents. Under this measure, **81.1%** (30/37) of scenarios exceed the null permutation. Notably, **10/37** scenarios lack synchronous coupling yet exhibit significant temporal influence (App. Tbl. 15), indicating that meaningful sequential coordination can arise without simultaneous conventions.

These two mechanisms dissociate systematically across benchmarks, revealing the underlying coordination structure each environment induces. SMAX V2 maps show the starkest separation—none trigger AA, yet 67% elicit DAI, suggesting that SMAX V2-style combat micro-management relies on sequential positioning rather than synchronous actions. Overcooked V1 presents a contrasting profile (100% AA, 40% DAI), reflecting rigid positional conventions in many scenarios. However, Overcooked V2's introduction of hidden information strengthens temporal dependence (100% DAI) while retaining synchronous coupling (82% AA). Finally, MPE stands out as the only suite where every scenario demands both coordination forms (100% AA and 100% DAI).

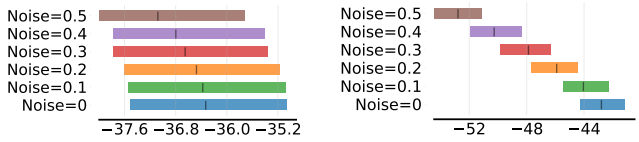
Summary. Our audit yields four main takeaways:

(1) **History dependence ≠ history utility.** All policies exhibit detectable history dependence (HAR^{norm} > null in 37/37 scenarios), yet only **43.2%** show a significant performance gain from memory (Fig. 4, Tbl. 14).

(2) **Hidden state and private information are separable.** PIF flags **70.3%** of scenarios, often different ones from HAR, confirming these are separate drivers of difficulty. The Overcooked V1→V2 contrast (**20% → 82%**) validates PIF as an environment-agnostic audit tool.

(3) **Coordination is structurally diverse.** AA (**64.9%**) and DAI (**81.1%**) dissociate across benchmarks: SMAX V2 exhibits temporal but not synchronous coordination, Overcooked V1 the reverse, while MPE, SMAX V1, MaBrax, and Overcooked V2 trigger both.

(4) **Few benchmarks jointly test partial observability and coordination.** MPE is the only suite in which every scenario satisfies all diagnostic criteria. Most scenarios do not require meaningful history use for strong performance despite being framed as Dec-POMDP challenges.



(a) MPE Simple Reference **(b) MPE Simple Spread**

Figure 5: Robustness to observational noise in MPE. We report the mean return and 95% stratified bootstrap confidence intervals for IPPO FeedForward (FF) across varying noise scales. Scenarios with higher $\mathrm{OAR}^{\mathrm{norm}}$ (*Simple Spread*) correlate with greater sensitivity to sensory perturbations.

Our diagnostics expose the divergence between what a benchmark *intends* to test and what it *actually requires*. By characterising *how* agents coordinate rather than just how well, these tools enable researchers to verify Dec-POMDP demands and deliberately select environments that stress-test specific capabilities. Furthermore, as demonstrated in Section 5, our metrics capture the behavioural impact of structural environment changes, potentially providing actionable guidance for designing more rigorous cooperative environments.

7 IMPLICATIONS

In many real-world cooperative systems, agents are expected to adapt to changes in their environment and to the behaviour of other agents. While our probes and metrics do not directly measure generalisation, they allow us to detect when policies exhibit weak statistical dependence between observations/histories and actions, i.e., low $\mathrm{OAR}^{\mathrm{norm}}$ and $\mathrm{HAR}^{\mathrm{norm}}$. Such instances suggest that agents may be relying on learned conventions or implicit coordination strategies rather than actively conditioning on current observations.

This distinction has nuanced implications. On the one hand, environments in which agents can solve the task via conventions without relying on observations may yield policies that are robust to sensory noise or partial occlusion. On the other hand, such policies may be brittle under structural changes to the environment, to the behaviour of other agents, or even to minimal variations in the task definition [32], since coordination may depend on fixed joint strategies rather than observation-driven adaptation.

To examine how our diagnostics relate to behavioural robustness, we conduct controlled evaluations under noisy observations in two MPE tasks with differing $\mathrm{OAR}^{\mathrm{norm}}$ values: *Simple Spread* and *Simple Reference*. The former exhibits substantially higher estimated mutual information between observations and actions than the latter (IPPO FF, App. D).

To test robustness to noise, we perturb observations x with additive Gaussian noise scaled by the feature-wise standard deviation σ_x, computed over N initial rollouts. For noise scale $k \in [0, 0.5]$. For more details on how we add noise see the (App. C).

Fig. 5 shows that performance in *Simple Spread* degrades more substantially under increasing noise than in *Simple Reference*. This is consistent with the higher $\mathrm{OAR}^{\mathrm{norm}}$ observed in *Simple Spread*, when using IPPO FF.

A key take-away is that information-theoretic diagnostics can provide structured signals about how policies utilise observations and interact with other agents under the training distribution. When interpreted jointly, they can indicate whether behaviour appears observation-driven or convention-driven. However, these metrics quantify statistical dependence rather than causal relationships. As a result, high mutual information does not guarantee sensitivity to noise, and low values do not necessarily imply the absence of structured coordination. Careful behavioural evaluation alongside the use of diagnostics can however provide indications of robustness and generalisation of learned policies.

8 CHALLENGES AND LIMITATIONS

Policy-dependent probes. All diagnostics are expectations under the converged joint policy $p^{\boldsymbol{\pi}}$ and therefore characterise *learned behaviour* under IPPO/MAPPO with FF/RNN architectures, not worst-case or best-case properties of the environment. This is deliberate, as we probe behaviours induced by widely used algorithms; however, stronger or weaker algorithms may yield different diagnostic profiles for the same scenario.

Estimation noise. Our MI/CMI/DI estimators (kNN and KSG [18]) are biased in finite samples, especially with long histories or large action spaces. We mitigate this via permutation null baselines that account for estimator-specific bias, and report bootstrap confidence intervals throughout. Nonetheless, these probes are diagnostic tools, not hard pass/fail filters, and borderline cases should be interpreted with caution.

9 CONCLUSION

In this work, we introduce a principled diagnostic framework to probe whether cooperative MARL agents genuinely exhibit Dec-POMDP reasoning. By coupling information-theoretic metrics with simple decision rules, our diagnostics evaluate *how* policies solve tasks, not just *how well*, moving evaluation beyond raw returns.

Applied to 37 scenarios across seven environments, our analysis reveals that: **(i)** history dependence is ubiquitous but rarely yields a performance advantage; **(ii)** hidden state and private teammate information are separable drivers of difficulty; and **(iii)** synchronous and temporal coordination frequently dissociate across domains. Notably, MPE is the only environment in which every scenario satisfies all diagnostic criteria. Our case study further demonstrates that the form of emergent coordination is shaped primarily by information bottlenecks in the environment design.

These findings motivate a shift toward benchmarks that strictly compel agents to exploit historical context and coordinate under private information—making **partial observability** and **decentralised coordination** non-optional for success.

ACKNOWLEDGMENTS

An author on this project received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101120726. This work was also supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding Guarantee 10085198.

REFERENCES

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G Bellemare. 2021. Deep Reinforcement Learning at the Edge of the Statistical Precipice. *Advances in Neural Information Processing Systems* (2021).

[2] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. 2024. *Multi-Agent Reinforcement Learning: Foundations and modern approaches*. MIT Press.

[3] Karl Johan Åström. 1965. Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.* 10, 1 (1965), 174–205.

[4] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280 (2020), 103216.

[5] 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.

[6] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems* 32 (2019).

[7] Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, and Stefano V Albrecht. 2021. Scaling multi-agent reinforcement learning with selective parameter sharing. In *International Conference on Machine Learning*. PMLR, 1989–1998.

[8] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533* (2020).

[9] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. 2023. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 37567–37593.

[10] Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. 2019. Bayesian action decoder for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1942–1951.

[11] Tobias Gessler, Tin Dizdarevic, Ani Calinescu, Benjamin Ellis, Andrei Lupu, and Jakob Nicolaus Foerster. 2025. OvercookedV2: Rethinking Overcooked for Zero-Shot Coordination. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=hlvLM3GX8R

[12] Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. 2022. Towards a standardised performance evaluation protocol for cooperative marl. *Advances in Neural Information Processing Systems* 35 (2022), 5510–5521.

[13] Matthew J Hausknecht and Peter Stone. 2015. Deep Recurrent Q-Learning for Partially Observable MDPs.. In *AAAI fall symposia*, Vol. 45. 141.

[14] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. 2021. Off-belief learning. In *International Conference on Machine Learning*. PMLR, 4369–4379.

[15] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*. PMLR, 4399–4410.

[16] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[17] Landon Kraemer and Bikramjit Banerjee. 2016. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing* 190 (2016), 82–94.

[18] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 69, 6 (2004), 066138.

[19] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).

[20] James Massey et al. 1990. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, Vol. 2. 1.

[21] Frans A Oliehoek, Christopher Amato, et al. 2016. *A concise introduction to decentralized POMDPs*. Vol. 1. Springer.

[22] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (2008), 289–353.

[23] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. 2020. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869* (2020).

[24] Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. 2021. Facmac: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems* 34 (2021), 12208–12221.

[25] Brian C Ross. 2014. Mutual information between discrete and continuous data sets. *PloS one* 9, 2 (2014), e87357.

[26] Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, et al. 2023. Jaxmarl: Multi-agent rl environments in jax. *arXiv preprint arXiv:2311.10090* (2023).

[27] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043* (2019).

[28] Ruo Yu Tao, Kaicheng Guo, Cameron Allen, and George Konidaris. 2025. Benchmarking partial observability in reinforcement learning with a suite of memory-improvable domains. *arXiv preprint arXiv:2508.00046* (2025).

[29] Kale-ab Abebe Tessera, Arrasy Rahman, Amos Storkey, and Stefano V Albrecht. 2025. HyperMARL: Adaptive Hypernetworks for Multi-Agent RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=56CgYnf9Dr

[30] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin* 1, 6 (1945), 80–83.

[31] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[32] Riccardo Zamboni, Mirco Mutti, and Marcello Restelli. 2025. Towards Principled Unsupervised Multi-Agent Reinforcement Learning. https://openreview.net/forum?id=XF1OzY8mEI