When One Agent Reshapes Another: Cross-Hessian Interpretation of Local Coupling in MARL

Risal Shahriar Shefin[1,†], Debashis Gupta[1,†], Murtadha Alsayegh[2] and Sarra Alqahtani[1,*]

[1]Wake Forest University, Computer Science Department, Winston-Salem, NC, USA
[2]University of Central Florida, School of Modeling, Simulation and Training, Orlando, FL, USA

Abstract

This late-breaking work presents an initial exploration of local inter-agent coupling in centralized training with decentralized execution (CTDE) multi-agent reinforcement learning (MARL). In particular, we study the question: *when does one agent's action locally reshape another agent's decision landscape?* We propose a gradient-based interpretation framework based on the cross-Hessian of a centralized critic and use its Frobenius norm, which we call the Frobenius Coupling Norm (FCN), as a compact summary of local directed coupling. As a preliminary step, we outline the local sensitivity interpretation of this quantity and examine its usefulness as an explainability signal in practice. We apply the framework to two standard CTDE algorithms, MADDPG and HATRPO, on the Simple Spread and Simple Adversary tasks. Using timestep-level FCN heatmaps, environment-frame inspection, and controlled single-step action perturbations, we show early evidence that the method can reveal interpretable coupling patterns and distinguish algorithmic behavior. In our experiments, HATRPO exhibits substantially smaller coupling magnitudes than MADDPG. Overall, these initial findings suggest that cross-Hessian analysis is a promising direction for explainability in MARL, while a broader empirical and theoretical treatment remains for future work.

1. Introduction

Centralized training with decentralized execution (CTDE) is a standard design principle in multi-agent reinforcement learning (MARL) [1, 2]. In CTDE, centralized critics are trained on joint information and can therefore capture dependencies among agents that decentralized policies do not explicitly reveal. This makes CTDE effective, but it also creates an explanation gap. After training, a practitioner may be able to inspect trajectories and returns while still struggling to answer time-local questions, such as: (i) *when* are two agents strongly dependent, (ii) *which* pairs dominate the interaction structure of an episode, and (iii) *how* does one agent's action locally reshape another agent's learning signal? These questions matter for explainability. In cooperative or mixed cooperative–competitive tasks, the same policy may move through phases of strong coordination, weak interaction, temporary role specialization, and near-independent execution. Episode-level reward does not reveal these shifts directly. Even when team performance appears stable, the local sensitivity structure that supports policy learning can change substantially over time.

In this paper, we propose a local interpretation framework based on the cross-Hessian of a centralized critic. Intuitively, the critic gradient for one agent describes how the critic changes with respect to that agent's action. The cross-Hessian then measures how a change in one agent's action locally alters the critic gradient associated with another agent. We interpret this quantity as a directed local coupling operator: it captures how one agent reshapes another agent's decision landscape at a particular state and timestep. Its Frobenius norm, which we call the Frobenius Coupling Norm (FCN), provides a compact scalar summary of the strength of this local coupling. To better frame the goals of this paper, we ask the following research questions. **RQ1:** Can the cross-Hessian of a centralized critic in CTDE MARL be

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil
*Corresponding author.
†These authors contributed equally.
shefrs24@wfu.edu (R. S. Shefin); guptd23@wfu.edu (D. Gupta); murtadha.alsayegh@ucf.edu (M. Alsayegh); alqahtas@wfu.edu (S. Alqahtani)

interpreted as a directed local coupling operator that captures how one agent's action locally reshapes another agent's decision landscape? **RQ2:** Can the Frobenius norm of this operator serve as a compact and interpretable summary of coupling strength over time and across ordered agent pairs? **RQ3:** Does stronger estimated coupling correspond to greater local gradient sensitivity under controlled one-step action perturbations?

We evaluate our framework as an explainability tool on Simple Spread and Simple Adversary from the Multi-Agent Particle Environments (MPE) [3], using MADDPG [1] and HATRPO [4] algorithms. We first visualize coupling heatmaps over time, then ground selected high- and low-coupling moments with environment frames, and finally test whether stronger estimated coupling corresponds to greater local gradient sensitivity under controlled one-step action deviations.

Contributions.

1. **Cross-Hessian coupling interpretation.** We introduce the cross-Hessian of a centralized critic as a directed local coupling operator for CTDE-based MARL.
2. **FCN as an interpretable scalar summary.** We define the Frobenius Coupling Norm and show that it upper-bounds local gradient sensitivity while exactly characterizing expected squared sensitivity under isotropic fixed-norm deviations.
3. **An explanation workflow for MARL.** We propose a step-wise evaluation protocol based on heatmaps, frame-level grounding, and controlled local deviations.

2. Background & Related Works

Centralized training with decentralized execution (CTDE) is a standard framework in multi-agent reinforcement learning (MARL) [1, 2]. In CTDE, each agent acts from its own local observation, so agent i uses a policy of the form $a^i \sim \pi_i(\cdot \mid o^i)$ at execution time. During training, however, learning is supported by a centralized critic that conditions on global state and joint actions, such as $Q(s, \mathbf{a})$ or $Q_i(s, \mathbf{a})$ with $\mathbf{a} = (a^1, \ldots, a^n)$. Because the critic depends on all agents' actions, it can encode inter-agent dependencies and represent how one agent's behavior affects another within the shared environment. Representative CTDE methods include MADDPG, which conditions each agent critic on joint actions, COMA, which uses a centralized counterfactual critic for credit assignment, QMIX, which factorizes a joint action-value under monotonicity constraints, and HARL/HATRPO, which extends trust-region style optimization to heterogeneous agents [1, 5, 6, 7]. This architectural choice makes the critic a natural explanatory object: if the critic aggregates state and joint action information, then its local derivatives can expose when one agent's action changes another agent's learning signal.

Early work in explainable RL visualized gradient-based policy saliency in Atari [8]; subsequent methods introduced policy-level abstractions [9], strategic XRL testbeds [10], causal and counterfactual explanations [11], and critical scrutiny of whether saliency is genuinely explanatory rather than merely exploratory [12]. Heuillet et al. synthesized this literature and argued that RL explanation requires methods tailored to temporally extended decision-making instead of static prediction settings [13].

Within MARL, however, interpretability remains comparatively underdeveloped and has mostly focused on agent-level attribution or policy summarization. Heuillet et al. took an early step toward collective XAI for MARL by using Shapley values to explain cooperative strategies and agent contribution [14]. Related work studies inter-agent effect through causal influence rewards [15], policy summarization and query-based explanations for centralized MARL [16], concept-based interpretable latent structure in multi-agent robotic control [17], explicit estimation of influence graphs over agents [18], and recent extensions of explanation methods to decentralized MARL policies [19]. These studies show clear demand for tools that reveal who influences whom, when, and why. Still, most existing approaches explain behavior through trajectories, concepts, queries, or contribution scores; they do not directly quantify how one agent locally reshapes another agent's decision landscape through the centralized critic. Our cross-Hessian view complements this literature by targeting that missing differential perspective: it treats inter-agent coupling as a local, directed, timestep-specific property of the critic and summarizes it with a compact scalar that can be visualized over an episode.

3. Approach: Cross-Hessian Coupling Interpretation

Let $Q(s, \mathbf{a})$ be a differentiable centralized critic, where $\mathbf{a} = (a^1, \ldots, a^n)$ and each $a^i \in \mathbb{R}^{d_i}$. At a fixed state–joint-action pair $(s, \mathbf{a})$, define the critic-gradient field for agent i as $g_i(s, \mathbf{a}) = \nabla_{a^i} Q(s, \mathbf{a})$. This quantity describes the local critic signal with respect to agent i's action. To understand how another agent influences that signal, we differentiate again with respect to agent j's action:

$$H_{i,j}(s, \mathbf{a}) = \nabla_{a^j} g_i(s, \mathbf{a}) = \nabla_{a^j} \nabla_{a^i} Q(s, \mathbf{a}). \tag{1}$$

The matrix $H_{i,j}(s, \mathbf{a}) \in \mathbb{R}^{d_i \times d_j}$ is interpreted as a directed coupling operator from agent j to agent i. This interpretation follows directly from a local expansion. If agent j's action is perturbed by a small vector δ, then a first-order Taylor expansion of the mapping $\phi_i(a^j) = \nabla_{a^i} Q(s, a^1, \ldots, a^j, \ldots, a^n)$ gives

$$\phi_i(a^j + \delta) - \phi_i(a^j) = H_{i,j}(s, \mathbf{a})\, \delta + r(\delta), \tag{2}$$

where the remainder satisfies $\frac{\|r(\delta)\|_2}{\|\delta\|_2} \to 0$ as $\|\delta\|_2 \to 0$. Thus, $H_{i,j}$ is the first-order linear map from agent j's action deviation to the induced change in agent i's gradient field, making it the core explanation object for local inter-agent coupling. For visualization and comparison across timesteps, we summarize this matrix with the Frobenius Coupling Norm (FCN):

$$\mathrm{FCN}_{i \leftarrow j}(s, \mathbf{a}) \triangleq \|H_{i,j}(s, \mathbf{a})\|_F. \tag{3}$$

FCN is directed, so $\mathrm{FCN}_{i \leftarrow j}$ need not equal $\mathrm{FCN}_{j \leftarrow i}$. At timestep t in a rollout, we form a directed coupling matrix $C_t \in \mathbb{R}^{n \times n}$, where $(C_t)_{i,j} = \mathrm{FCN}_{i \leftarrow j}(s_t, \mathbf{a}_t)$. Stacking these matrices over timesteps yields an episode-level heatmap whose rows are directed pairs $j \to i$ and whose columns are timesteps. Equation (2) also gives the basic theoretical intuition for FCN. Taking norms yields

$$\|\phi_i(a^j + \delta) - \phi_i(a^j)\|_2 \le \|H_{i,j}(s, \mathbf{a})\|_F \|\delta\|_2 + o(\|\delta\|_2), \tag{4}$$

So a larger FCN implies greater possible local sensitivity under small deviations. In this way, FCN serves as a conservative scalar summary of coupling strength, while the stacked matrices provide a compact explanation artifact for showing when coupling spikes or collapses, which directed pairs dominate those phases, and whether the coupling profile is transient, sustained, or task-dependent. These heatmaps form the first stage of our empirical evaluation.

4. Experiments

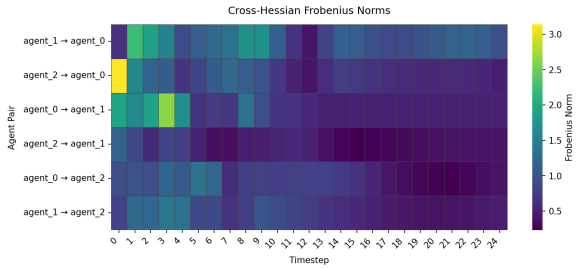
We evaluate the proposed interpretation framework on Simple Spread and Simple Adversary from MPE [3, 20]. All settings use three agents with continuous action spaces. We analyze trained policies and critics post hoc; no additional learning is performed during interpretation. All gradients and Hessian blocks are computed offline with automatic differentiation.

4.1. Algorithms and analysis setup

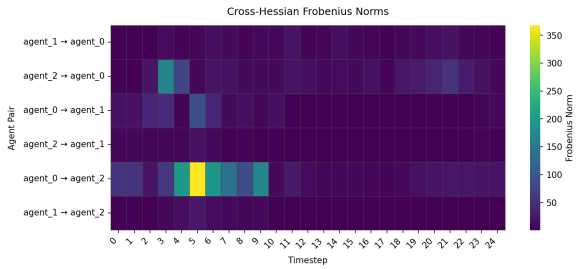
We study two representative CTDE methods: **(1) MADDPG** [1], a deterministic multi-agent actor–critic method with centralized critics. **(2) HATRPO** [4], a trust-region method for multi-agent settings.
For heatmaps and frame-level interpretation, we use representative rollout episodes. For quantitative sensitivity analysis, we use 2500 single-timestep deviation trials across 100 random seeds. In each trial, only one agent and one timestep are modified, with perturbation magnitude fixed at $\epsilon = 0.01$.

4.2. Coupling heatmaps over episodes

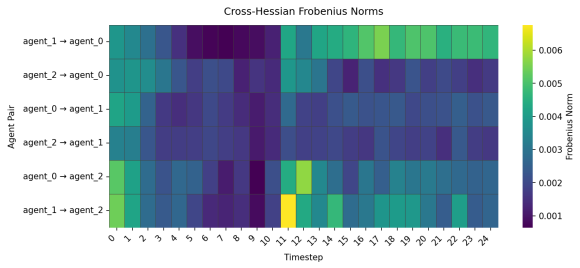
We present heatmaps before frames to follow the interpretation workflow: first, identify coupling structure globally, then ground it locally.



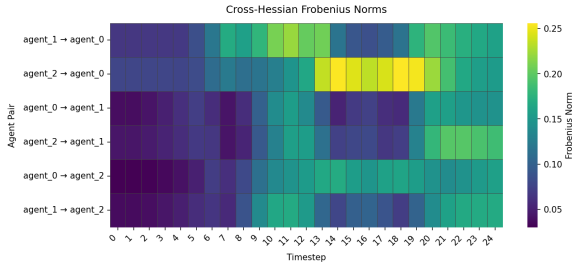
(a) MADDPG (Simple Spread)



(b) MADDPG (Simple Adversary)



(c) HATRPO (Simple Spread)



(d) HATRPO (Simple Adversary)

Figure 1: FCN heatmaps over episode timesteps. Each row is a directed pair $j \to i$ and each column is a timestep. Brighter values indicate stronger directed coupling.

4.2.1. MADDPG

Figure 1a shows the FCN heatmap for a representative MADDPG rollout in Simple Spread. Coupling is strongest early and decreases as the system stabilizes, consistent with an initial coordination phase followed by more independent navigation. Figure 1b shows MADDPG on Simple Adversary. The dominant early interaction is $0 \to 2$, peaking around timestep 5, then rapidly stabilizing. Smaller transient spikes appear for $2 \to 0$ and $0 \to 1$, while other pairs remain near baseline.

4.2.2. HATRPO

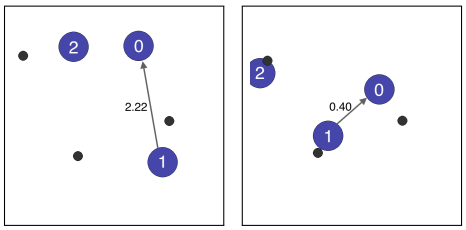
Figure 1c shows the HATRPO heatmap on Simple Spread. Overall coupling is much weaker than in MADDPG, but the heatmap still reveals a clear regime shift near the middle of the episode. Even at small absolute magnitudes, it highlights the timing of the coordination-to-stabilization transition. Figure 1d shows HATRPO on the Simple Adversary task. Unlike MADDPG, coupling grows during the middle-to-late episode, especially for $2 \to 0$, suggesting more sustained strategic dependence near the critical stage of the task.

4.3. Semantic grounding with environment frames

After locating informative moments in the heatmaps, we inspect environment frames to interpret their behavior.

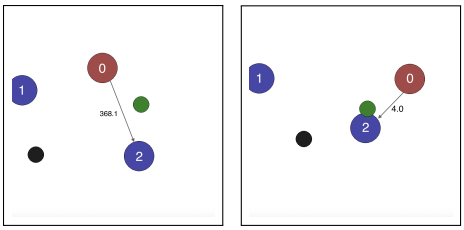
4.3.1. MADDPG

In Simple Spread, for the pair $1 \to 0$, FCN is maximal at $t = 1$ and minimal at $t = 12$ in the representative rollout. Figure 2a grounds these moments. At $t = 1$, the episode is in a high-coupling regime: agents are still negotiating role assignment and relative motion, so one agent's movement strongly reshapes another's decision landscape. By $t = 12$, agents have largely committed to distinct landmarks and navigation is more locally determined. Together, the heatmap and frames suggest that coupling is strongest during assignment and weakest during stabilized execution. In Simple Adversary, the pair $0 \to 2$ reaches maximum FCN at $t = 5$ and minimum FCN at $t = 17$. Figure 2b shows these moments. At peak coupling, both agents are still contesting or reacting to motion near the landmark, so the adversary's



(i) Max FCN ($t = 1$) (ii) Min FCN ($t = 12$)

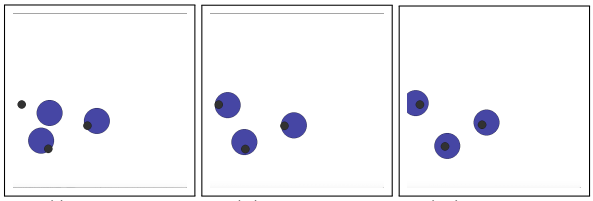
(a) Simple Spread



(i) Max FCN ($t = 5$) (ii) Min FCN ($t = 17$)

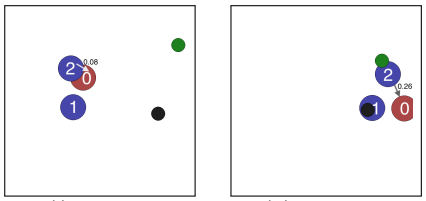
(b) Simple Adversary

Figure 2: MADDPG: Example environment frames at the max and min FCN moments of a directed agent pair.



(i) $t = 9$. (ii) $t = 10$. (iii) $t = 11$.

(a) Simple Spread: Representative frames around the sharp heatmap phase shift.



(i) $t = 4$ (ii) $t = 18$

(b) Simple Adversary: environment frames at the min and max FCN for the directed pair $2 \to 0$.

Figure 3: HATRPO: Example environment frames across tasks.

movement strongly reshapes agent 2's local critic-gradient field. Later, once the configuration has settled, the same interaction is much weaker.

4.3.2. HATRPO

For HATRPO on Simple Spread, the heatmap shows a sharp regime change between timesteps 10 and 11. Rather than global extrema, Figure 3a shows representative frames around this transition. Agents are still approaching landmarks at $t = 9$, reach them around $t = 10$, and maintain stable coverage from $t = 11$ onward. The frame sequence explains the phase shift: once coverage is established, cross-agent dependence drops sharply. In Simple Adversary, the largest FCN values occur for $2 \to 0$. Coupling is weakest at $t = 4$ and strongest at $t = 18$ in the representative rollout. Figure 3b shows that weak coupling appears early, before the strategic interaction becomes critical, while strong coupling emerges late, when agent 2 more directly shapes the adversary's response.

4.4. Controlled deviations as validation of the interpretation framework

We next use controlled one-step action deviations to test whether larger estimated coupling corresponds to larger local gradient sensitivity.

Deviation protocol. At each trial, for a chosen ordered pair $(j \to i)$ and timestep t, we compute the singular value decomposition $H_{i,j}(t) = U\Sigma V^\top$, and take the top right singular vector $v_{\max}$, and perturb agent j's action by $\delta = \epsilon v_{\max}$, where $\epsilon = 0.01$. We then measure the induced shift in agent i's gradient field: $\|\Delta g_i(t)\|_2 = \left\|\nabla_{a^i} Q(s_t, \dots, a_t^j + \delta, \dots) - \nabla_{a^i} Q(s_t, \dots, a_t^j, \dots)\right\|_2$. From the local expansion in Equation (2), the first-order change in agent i's gradient field under a small action deviation δ is governed by $H_{i,j}(t)\delta$. Choosing δ along the top right singular vector therefore probes the direction of largest local linear amplification by the cross-Hessian. For this reason, the deviation experiment should be interpreted as a test of the worst-case local sensitivity of the coupling operator, rather than a direct test of FCN alone. For completeness, we also record episodic reward change. $\Delta R_i = R_i^{\text{normal}} - R_i^{(j,t)}$, where $R_i^{(j,t)}$ is the episodic reward when only one action at timestep t is perturbed for agent j.

Hypothesis. If the cross-Hessian interpretation is informative, then states with larger estimated coupling should produce larger induced gradient shifts. Under the current protocol, that prediction is

Table 1
MADDPG ($n = 3$): per-pair summary for FCN, induced gradient shift, and episodic reward change under $\epsilon = 0.01$ for Simple Spread and Simple Adversary.

		Simple Spread			Simple Adversary		
Agent j	Agent i	$\|H_{i,j}\|_F$ (mean ± std)	$\|\Delta g_i\|_2$ (mean ± std)	ΔR_i (mean ± std)	$\|H_{i,j}\|_F$ (mean ± std)	$\|\Delta g_i\|_2$ (mean ± std)	ΔR_i (mean ± std)
0	0	2.038958 ± 1.599183	0.018846 ± 0.015716	-0.000010 ± 0.003836	122.084 ± 264.406	1.138 ± 2.173	-0.000497 ± 0.004715
0	1	1.206132 ± 0.957657	0.011124 ± 0.009311	0.000104 ± 0.004029	13.121 ± 25.225	0.130 ± 0.249	0.000518 ± 0.003629
0	2	1.477626 ± 1.195228	0.013854 ± 0.011874	0.000076 ± 0.003990	55.140 ± 79.228	0.511 ± 0.697	0.000054 ± 0.001779
1	0	1.311637 ± 1.062565	0.012281 ± 0.010538	0.000144 ± 0.004624	14.617 ± 28.799	0.143 ± 0.284	-0.000086 ± 0.002779
1	1	2.042311 ± 1.440923	0.018528 ± 0.013972	-0.000308 ± 0.005144	5.509 ± 9.667	0.054 ± 0.097	0.000791 ± 0.007483
1	2	1.692610 ± 1.407330	0.015915 ± 0.013916	0.000242 ± 0.005046	5.050 ± 7.096	0.049 ± 0.073	0.000964 ± 0.007993
2	0	1.214197 ± 0.987272	0.011388 ± 0.009893	0.000028 ± 0.005581	48.848 ± 72.407	0.484 ± 0.719	0.000060 ± 0.001525
2	1	1.046113 ± 0.846585	0.009601 ± 0.008299	0.000242 ± 0.005270	4.081 ± 5.481	0.039 ± 0.055	-0.000196 ± 0.005538
2	2	2.183318 ± 2.184370	0.020528 ± 0.021857	-0.000188 ± 0.005494	12.605 ± 17.698	0.122 ± 0.181	-0.000673 ± 0.004021

Table 2
HATRPO ($n = 3$): per-pair summary for FCN, induced gradient shift, and episodic reward change under $\epsilon = 0.01$ for Simple Spread and Simple Adversary.

		Simple Spread			Simple Adversary		
Agent j	Agent i	$\|H_{i,j}\|_F$ (mean ± std)	$\|\Delta g_i\|_2$ (mean ± std)	ΔR_i (mean ± std)	$\|H_{i,j}\|_F$ (mean ± std)	$\|\Delta g_i\|_2$ (mean ± std)	ΔR_i (mean ± std)
0	0	0.007022 ± 0.005913	0.000040 ± 0.000030	0.001292 ± 0.038191	0.186 ± 0.117	0.00129 ± 0.00080	0.00041 ± 0.00801
0	1	0.001668 ± 0.001193	0.000016 ± 0.000012	0.000287 ± 0.041169	0.055 ± 0.044	0.00049 ± 0.00043	-0.00011 ± 0.00820
0	2	0.001780 ± 0.001277	0.000017 ± 0.000013	0.000421 ± 0.041056	0.050 ± 0.046	0.00046 ± 0.00045	0.00032 ± 0.00903
1	0	0.001996 ± 0.001295	0.000019 ± 0.000013	-0.000394 ± 0.037343	0.093 ± 0.061	0.00071 ± 0.00048	0.00031 ± 0.01196
1	1	0.006438 ± 0.006067	0.000034 ± 0.000029	-0.000679 ± 0.036236	0.238 ± 0.142	0.00155 ± 0.00093	-0.00070 ± 0.01511
1	2	0.001550 ± 0.001259	0.000015 ± 0.000013	0.001427 ± 0.039290	0.061 ± 0.052	0.00056 ± 0.00050	-0.00062 ± 0.01020
2	0	0.001923 ± 0.001315	0.000018 ± 0.000013	0.000725 ± 0.028522	0.098 ± 0.064	0.00076 ± 0.00051	-0.00010 ± 0.00945
2	1	0.001497 ± 0.001040	0.000014 ± 0.000011	0.000480 ± 0.020256	0.068 ± 0.057	0.00062 ± 0.00054	-0.00058 ± 0.00937
2	2	0.007193 ± 0.007244	0.000038 ± 0.000035	0.000280 ± 0.018496	0.198 ± 0.124	0.00116 ± 0.00072	-0.00030 ± 0.01104

most direct for the operator and its spectral norm, and is expected to hold empirically for FCN as a practical scalar proxy.

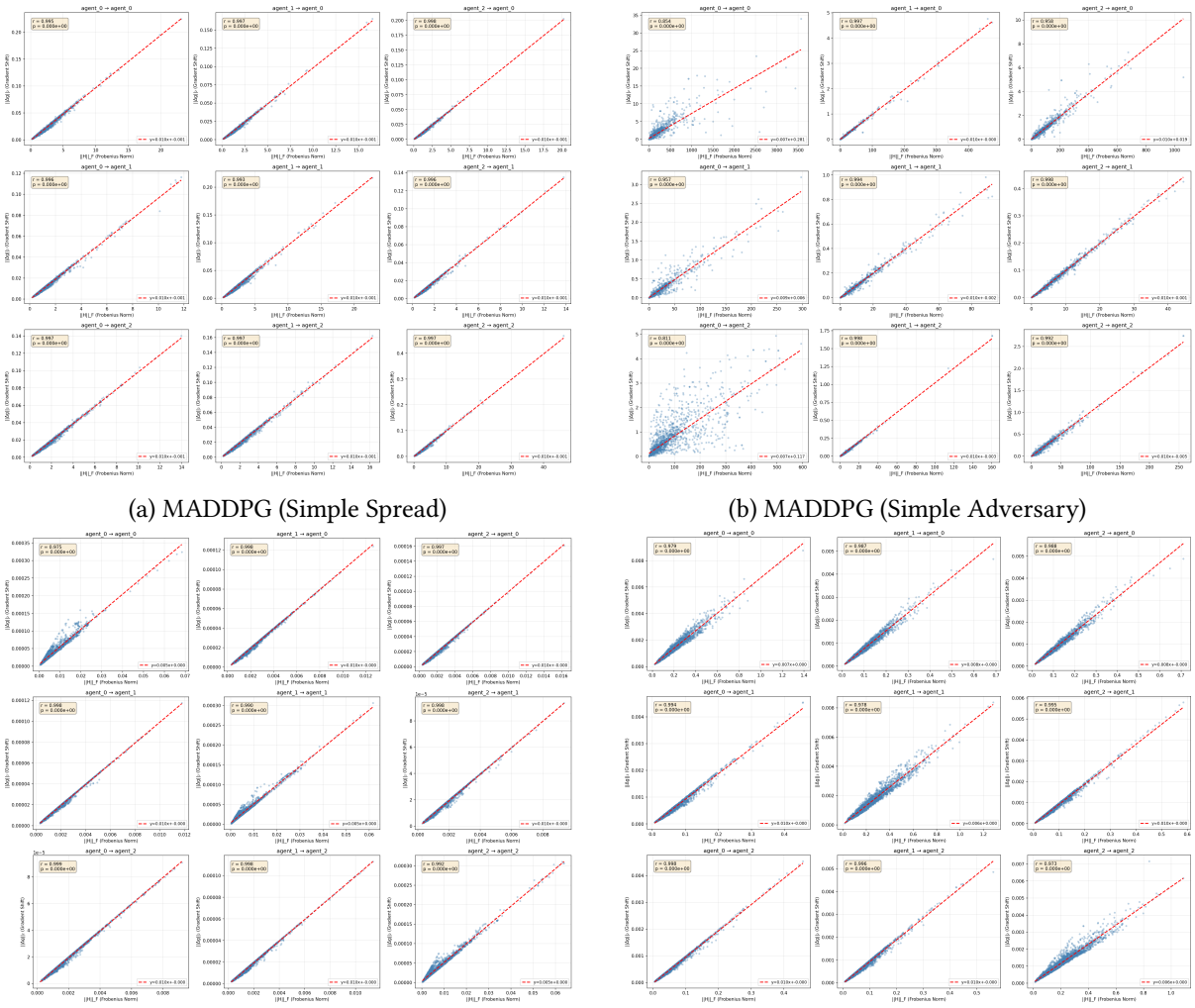
4.5. Quantitative results

4.5.1. MADDPG

Figures 4a and 4b plot induced gradient shift $\|\Delta g_i\|_2$ against FCN under the controlled single-step deviations above. Across directed pairs, the relationship is strongly monotonic and visually close to linear. Because the perturbation direction follows the top singular vector, these plots are best read as evidence that larger cross-Hessian magnitude yields larger worst-case local sensitivity, with FCN serving as a practical summary statistic. Table 1 gives per-pair summaries. Pairs with larger mean FCN also show larger mean gradient shifts. Diagonal entries are included only as a scale reference: they reflect within-agent curvature rather than inter-agent coupling and are therefore expected to be larger. For explainability, the important point is that cross-agent pairs in MADDPG remain large enough to substantially reshape another agent's gradient field. A second observation is that episodic reward changes remain near zero on average and do not mirror the coupling pattern. This shows why gradient-level explanation is useful: substantial local sensitivity can exist even when reward degradation from a single-step deviation is negligible or noisy.

4.5.2. HATRPO

Figures 4c and 4d show the same analysis for HATRPO, with Table 2 reporting aggregate statistics. Within HATRPO, FCN values and induced gradient shifts remain small in absolute magnitude across the evaluated settings, while still exhibiting a clear positive association under the controlled perturbation protocol. This supports the intended interpretation of the cross-Hessian-based signal: larger estimated coupling corresponds to larger local sensitivity, even when the overall scale is modest. Reward change again remains noisy and only weakly aligned with the coupling pattern. By contrast, the gradient-based analysis provides a more stable way to distinguish relatively stronger and weaker directed dependencies within the analyzed HATRPO runs.



(a) MADDPG (Simple Spread)

(b) MADDPG (Simple Adversary)

(c) HATRPO (Simple Spread)

(d) HATRPO (Simple Adversary)

Figure 4: Induced gradient shift $\|\Delta g_i\|_2$ versus FCN under single-step deviations aligned with the top singular-vector direction of $H_{i,j}$. Each panel corresponds to a directed pair $j \to i$.

5. Conclusion

This late-breaking work presents an initial cross-Hessian-based perspective for explaining local inter-agent dependence in CTDE MARL. Rather than offering a complete solution, we aim to highlight a promising direction for understanding when one agent locally reshapes another agent's decision landscape. Across Simple Spread and Simple Adversary, our early results suggest that this framework can reveal interpretable coupling patterns over time, connect them to concrete task situations, and distinguish algorithmic behavior beyond what reward alone shows. In particular, HATRPO exhibits substantially weaker coupling than MADDPG in our experiments. Overall, these findings position cross-Hessian analysis as a useful preliminary explainability tool and motivate broader future study.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under grant no. 2442581.

Declaration on Generative AI

The authors used ChatGPT to improve the writing and fix the grammar.

References

[1] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, Advances in neural information processing systems 30 (2017).

[2] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, Y. Wu, The surprising effectiveness of ppo in cooperative multi-agent games, Advances in neural information processing systems 35 (2022) 24611–24624.

[3] J. Terry, B. Black, N. Grammel, M. Jayakumar, A. Hari, R. Sullivan, L. S. Santos, C. Dieffendahl, C. Horsch, R. Perez-Vicente, et al., Pettingzoo: Gym for multi-agent reinforcement learning, Advances in Neural Information Processing Systems 34 (2021) 15032–15043.

[4] J. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, Y. Yang, Trust region policy optimisation in multi-agent reinforcement learning, in: ICLR 2022-10th International Conference on Learning Representations, The International Conference on Learning Representations (ICLR), 2022, p. 1046.

[5] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, S. Whiteson, Counterfactual multi-agent policy gradients, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[6] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, S. Whiteson, Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 4295–4304.

[7] Y. Zhong, J. G. Kuba, X. Feng, S. Hu, J. Ji, Y. Yang, Heterogeneous-agent reinforcement learning, Journal of Machine Learning Research 25 (2024) 1–67.

[8] S. Greydanus, A. Koul, J. Dodge, A. Fern, Visualizing and understanding atari agents, in: International conference on machine learning, PMLR, 2018, pp. 1792–1801.

[9] N. Topin, M. Veloso, Generation of policy-level explanations for reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 2514–2521.

[10] R. Pocius, L. Neal, A. Fern, Strategic tasks for explainable reinforcement learning, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 10007–10008.

[11] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 2493–2500.

[12] A. Atrey, K. Clary, D. Jensen, Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning, arXiv preprint arXiv:1912.05743 (2019).

[13] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowledge-Based Systems 214 (2021) 106685.

[14] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Collective explainable ai: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values, IEEE Computational Intelligence Magazine 17 (2022) 59–71.

[15] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, N. De Freitas, Social influence as intrinsic motivation for multi-agent deep reinforcement learning, in: International conference on machine learning, PMLR, 2019, pp. 3040–3049.

[16] K. Boggess, S. Kraus, L. Feng, Toward policy explanations for multi-agent reinforcement learning, in: International Joint Conference on Artificial Intelligence (IJCAI), 2022.

[17] R. Zabounidis, J. Campbell, S. Stepputtis, D. Hughes, K. P. Sycara, Concept learning for interpretable multi-agent reinforcement learning, in: Conference on robot learning, PMLR, 2023, pp. 1828–1837.

[18] F. R. Pieroth, K. Fitch, L. Belzner, Detecting influence structures in multi-agent reinforcement learning, in: Forty-first International Conference on Machine Learning, 2024.

[19] K. Boggess, S. Kraus, L. Feng, Explaining decentralized multi-agent reinforcement learning policies, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, 2026, pp. 17357–17365.

[20] I. Mordatch, P. Abbeel, Emergence of grounded compositional language in multi-agent populations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.