Evaluating XAI Support From A Hierarchical Reinforcement Learning Policy in Human-Agent Collaboration

Extended Abstract

Mateus Levi Simões Fernandes
PUC-Rio
Rio de Janeiro, Brazil
mfernandes@inf.puc-rio.br

Alberto Sardinha
PUC-Rio
Rio de Janeiro, Brazil
sardinha@inf.puc-rio.br

ABSTRACT

Explainable AI (XAI) research in human-agent collaboration has relied on hand-crafted policies in custom environments, limiting generalizability to state-of-the-art teaming benchmarks. We provide the first systematic evaluation of XAI support generated from a learned, intrinsically explainable policy in an established benchmark environment. We generate real-time explanations from hierarchical subtask selections via a trigger-based delivery system, comparing text versus audio modalities in a between-subjects experiment with gaming-experienced participants. While we found no statistically significant performance benefits, preliminary patterns suggest expertise may moderate explanation utility. This work establishes a methodological foundation for evaluating intrinsically explainable reinforcement learning in collaborative contexts, along with the first comparison of explanation modality effects in real-time human-agent teaming.

KEYWORDS

Human-Agent Teaming; Hierarchical Reinforcement Learning; XAI

ACM Reference Format:
Mateus Levi Simões Fernandes and Alberto Sardinha. 2026. Evaluating XAI Support From A Hierarchical Reinforcement Learning Policy in Human-Agent Collaboration: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/HDMG2174

1 INTRODUCTION

Research in explainable AI (XAI) and human-agent teaming (HAT) has remained largely disconnected: while XAI techniques have been developed for various AI systems [9], systematic evaluation of XAI in established HAT benchmarks remains limited [12]. HAT research in benchmark environments like Overcooked-AI [3] has prioritized performance optimization over transparency [6, 10], while the few works applying XAI to collaborative contexts have relied on custom environments [7] or non-deployable techniques [13].

We bridge this gap by leveraging the Hierarchical Ad Hoc Agents (HA^2) architecture [1], which achieves state-of-the-art performance through intrinsically explainable hierarchical policies. HA^2's Manager selects human-interpretable subtasks while Workers execute primitive actions, providing natural explanation content without post-hoc training [11], yet this explanatory potential remains unexplored.

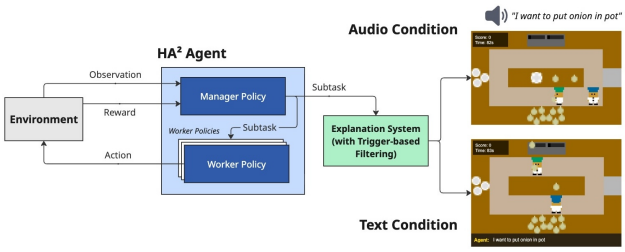

Figure 1: System architecture and experimental conditions. The Manager's subtask selections flow to both the Worker (for action execution) and the Explanation System with trigger-based filtering. Identical explanation content ("I want to put onion in pot") is delivered via audio speech synthesis (top right) or text display (bottom right, black bar).

We generate real-time explanations from HA^2's subtask selections during collaborative Overcooked-AI tasks. Since pilot testing revealed that continuous explanation display overwhelmed users, we developed a trigger-based delivery system generating explanations only at coordination-relevant moments. We compare text versus audio modalities, hypothesizing audio reduces cognitive interference in visually demanding tasks in a between-subjects experiment (n=38).

To our knowledge, we provide the first systematic evaluation of XAI support generated from a learned policy in an established benchmark, comparing different modalities while keeping content identical as well as providing a baseline trigger-based delivery system for managing explanation frequency for real-time collaboration. While our exploratory study found no statistically significant performance benefits from explanations, preliminary patterns suggest the peak performance of gaming experts may not benefit from simple status-based explanations.

2 METHODOLOGY

Agent Implementation: We use HA^2 [1] with the original paper's Worker policies and a Manager trained via PPO self-play on Overcooked-AI's Counter-Circuit layout. The Manager selects from twelve domain-specific subtasks (e.g., "pick up onion from dispenser", "place onion in pot") each timestep, with Workers executing corresponding primitive actions.

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/HDMG2174

Trigger-Based Explanation System: Since the Manager updates subtask selections every timestep, pilot studies revealed that directly exposing these choices created excessive and confusing communication. To remedy this, we developed a priority-based trigger system generating explanations only at coordination-relevant moments: (1) *Blocking* – when the human occupies the agent's path, (2) *Critical Path* – when the agent selects task-essential subtasks, (3) *Distance Threshold* – when the agent travels predetermined distance since last explanation, and (4) *Subtask Change* – when the Manager switches to new subtask. Each trigger maps to natural language templates (e.g., "Moving around you to [subtask]"). A six-second cooldown prevents excessive communication; text explanations are displayed on the game interface for 4 seconds, while audio explanations are generated via browser-native speech synthesis.

Experimental Design: Between-subjects design with three conditions: Control (no explanations), Text (visual display, 4-second duration), Audio (identical content via browser-native speech synthesis). Following recommendations established in [8], participants completed four 80-second Counter-Circuit sessions with NASA-TLX assessments [4] after each session and Godspeed Questionnaire [2] and Human-Agent Fluency Assessment [5] post-experiment. We implemented the experiment via a web-based platform supporting remote participation.

Participants: Following IRB approval, we recruited 41 participants via convenience sampling from the academic community, yielding n=38 after excluding participants with <80% instructional check accuracy (ages 19-43, M=25.2, SD=4.5; ≈66% male; Control n=14, Text n=14, Audio n=10). Post-hoc analysis revealed high gaming familiarity (M=8.1/10, SD=2.1; 92% scoring ≥6/10). We targeted n=30 based on prior work [7] expecting large effects, but sensitivity analysis revealed we were underpowered for the medium effects we observed (H1: d=0.51; H2: d=0.14).

3 PRELIMINARY FINDINGS

Performance: Gaming-experienced participants trended toward higher scores *without* explanations (Control: M=125.00, SD=18.19 vs. Explanations: M=114.17, SD=22.59; t(36)=-1.53, p=.136, d=-0.51), though this did not reach significance. Modality comparison showed equivalent performance (Text vs. Audio: t(22)=0.33, p=.745, d=0.14); NASA-TLX cognitive workload results revealed no significant differences across any dimension after Bonferroni correction, and subjective collaboration measures (Godspeed, Human-Agent Fluency) showed no significant effects (all p>0.05).

Learning Rates: Despite equivalent peak performance, learning trajectories revealed a monotonic pattern (Figure 2). Audio participants improved fastest (M=11.20 points/session), followed by Text (M=8.00), then Control (M=5.39) – a 2.1× difference between Audio and Control. While not reaching significance (F(2,34)=1.51, p=.236), this medium effect (d=0.50, p=.158) suggests explanations may accelerate adaptation rather than improve execution. This pattern of equivalent final performance despite different learning rates indicates that transparency may facilitate mental model development without necessarily translating to peak performance advantages, particularly among expert users who can rapidly identify efficient strategies.

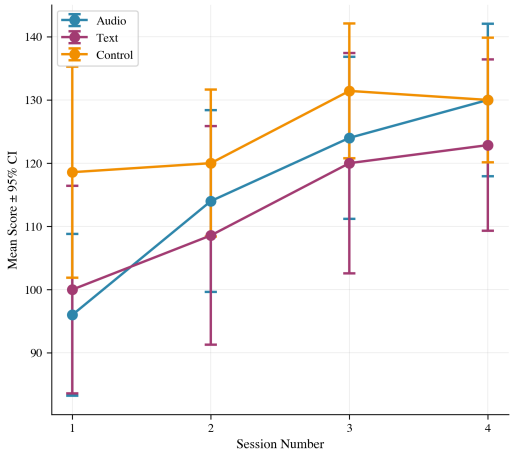

Figure 2: Learning curves by condition showing equivalent peak performance but differential adaptation rates. Audio explanations produced the steepest learning trajectory despite no final performance advantage.

Modality Salience: In qualitative feedback, multiple text participants reported not noticing explanations until later sessions, with no equivalence in audio participant responses. This salience difference may explain differential learning rates despite equivalent cognitive load and peak performance. Audio participants also trended toward rating the agent as less anthropomorphic (p=.081) and trustworthy (p=.100), suggesting that more transparent decision-making visible may expose the agent's artificial nature – particularly when the Manager policy changed subtasks mid-execution, creating mismatches between stated intentions and observed actions.

4 CONCLUSIONS

This exploratory work demonstrates that intrinsically explainable RL architectures can generate authentic real-time explanations in benchmark environments. Our trigger-based delivery system addresses the practical challenge of managing explanation frequency in fast-paced tasks, while our modality comparison reveals that audio ensures greater salience than text – though neither improved peak performance for our gaming-experienced sample. The preliminary pattern of faster adaptation with audio explanations despite equivalent final scores suggests XAI may facilitate mental model development rather than direct execution benefits, particularly among expert users. Adequately-powered further studies with more diverse populations and multiple layouts are needed to confirm these exploratory findings.

ACKNOWLEDGMENTS

This work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The experimental study was approved by PUC-Rio's Research Ethics Committee (Protocol SGOC 545928, dated June 11, 2025). The authors thank the National Council for Scientific and Technological Development (CNPq), Brazil, for the Research Productivity Fellowship (PQ) with reference 312699/2025-5.

REFERENCES

[1] Stéphane Aroca-Ouellette, Miguel Aroca-Ouellette, Katharina von der Wense, and Alessandro Roncone. 2025. Implicitly aligning humans and autonomous agents through shared task abstractions. In *Proceedings of the thirty-fourth international joint conference on artificial intelligence, IJCAI-25*, James Kwok (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4101–4109. https://doi.org/10.24963/ijcai.2025/457

[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[3] 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).

[4] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[5] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[6] Yi Loo, Chen Gong, and Malika Meghjani. 2023. A hierarchical approach to population training for human-AI collaboration. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 3011–3019.

[7] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems* 34 (2021), 610–623.

[8] Rohan Paleja, Michael Munje, Kimberlee Chestnut Chang, Reed Jensen, and Matthew Gombolay. 2024. Designs for Enabling Collaboration in Human-Machine Teaming via Interactive and Explainable Systems. *Neural Information Processing Systems (NeurIPS)* (2024).

[9] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable goal-driven agents and robots-a comprehensive review. *Comput. Surveys* 55, 10 (2023), 1–41.

[10] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. *Advances in neural information processing systems* 34 (2021), 14502–14515.

[11] Chenxu Wang, Zilong Chen, and Huaping Liu. 2024. On the Utility of External Agent Intention Predictor for Human-AI Coordination. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2546–2548.

[12] Rosina O Weber, Adam J Johs, Prateek Goel, and João Marques Silva. 2024. XAI is in trouble. *AI Magazine* 45, 3 (2024), 300–316. https://doi.org/10.1002/aaai.12184 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12184.

[13] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–31. https://doi.org/10.1145/3610072