Can you see how I learn? Human Observers' Inferences about Reinforcement Learning Agents' Learning Processes

Extended Abstract

Bernhard Hilpert
Leiden University
Leiden, Netherlands
b.hilpert@liacs.leidenuniv.nl

Muhan Hou
Vrije Universiteit Amsterdam
Amsterdam, Netherlands
m.hou@vu.nl

Kim Baraka
Vrije Universiteit Amsterdam
Amsterdam, Netherlands
k.baraka@vu.nl

Joost Broekens
Leiden University
Leiden, Netherlands
joost.broekens@gmail.com

ABSTRACT

Human-in-the-loop Reinforcement Learning (RL) often suffers from suboptimal human teaching signals. Yet, how humans perceive and interpret RL agent's learning behavior is largely unknown. In a bottom-up approach with two experiments, this work provides a data-driven understanding of the factors in RL agents' behavior that influence the understanding of the agent's learning process for human observers. In two consecutive experiments with two different RL agents (a tabular and function approximation agent in a navigation and a manipulation task), human observations of agent learning behavior was assessed and systematically analyzed. Four common emerging themes were observed: Agent Goals, Knowledge, Decision Making and Learning Mechanisms, each with specific subclusters, offering insights for transparency in RL and HRI.

KEYWORDS

Human-in-the-loop RL; Explainability; Human Robot-Interaction; Hybrid Intelligence

ACM Reference Format:
Bernhard Hilpert, Muhan Hou, Kim Baraka, and Joost Broekens. 2025. Can you see how I learn? Human Observers' Inferences about Reinforcement Learning Agents' Learning Processes: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

In Human-in-the-loop Reinforcement Learning (RL), human users use teaching signals (i.e. [1, 7, 12]) to collaborate with AI agents towards achieving a joint outcome that benefits both [2].

While a teacher can provide insight and expert knowledge to a learning agent [6], human feedback can also be suboptimal, e.g. by giving delayed [3] or unbalanced feedback [15]. Since feedback is grounded in the teachers' interpretation of observed agent learning behavior, this misalignment in the teacher-learner loop could arise from teacher's misunderstanding of agent learning processes [13]. To improve collaboration, a deeper understanding of how teachers interpret agent learning processes from observation is essential [5].

However, up to this point, the process of drawing inferences about the learning process from observations has only been examined indirectly (i.e. as a function of teacher interaction behavior, see e.g. [5, 11, 13]) or in simulated settings [10]. More importantly, all human user studies on this topic employed paradigms in which teaching signals played an active role, which can lead to expectations of how the agent should react to the teaching and therefore a biased interpretation of observations [11, 14]. A neutral, direct assessment is necessary.

Open questions include the correct setting, time-chunking of observations, how to even query inferences about an RL agent's learning process from human observers, and what aspects of agent learning processes human observers draw inferences about.

The main Research Question for this study therefore is: *What do human teachers infer about RL agents' learning processes from observing the agents' learning behavior?* This work employs a human-centered approach to investigate how individuals perceive and interpret the learning process of RL agents. In an exploratory and a confirmatory experiment, a set of common emergent themes in teacher inferences regarding agent behavior is identified and validated across a range of tasks and RL algorithms. Additionally, a detailed analysis is provided that shows how these themes are applied to draw inferences about the ongoing learning task.

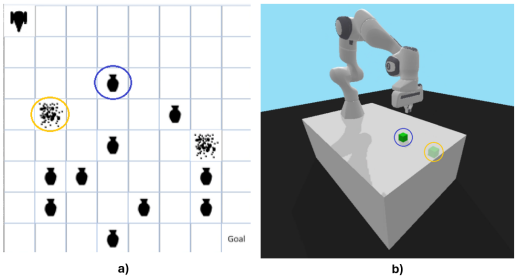

Figure 1: a) NT: Clean dirt (yellow) without breaking furniture (blue). b) MT: Push "medication" (blue) to target (yellow).

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

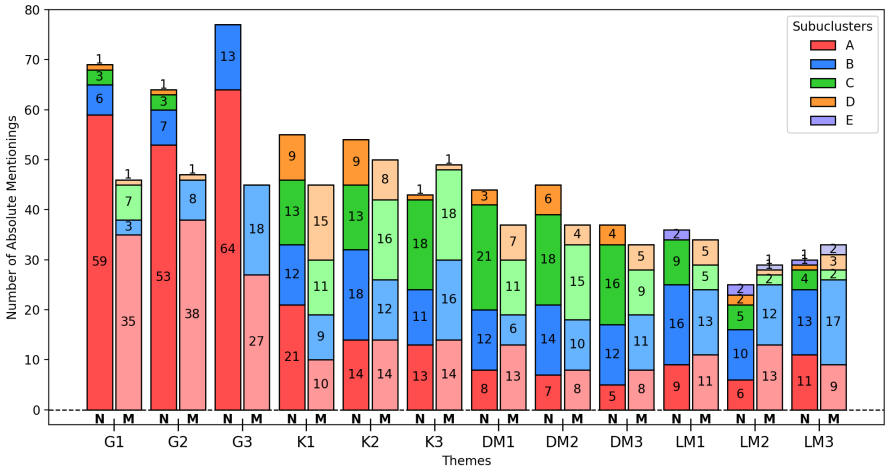

Figure 2: Themes + Subclusters across time split by task. Letters correspond to subclusters for each theme reported in section 3

2 METHODOLOGY

In two sequential experiments with a newly developed experimental paradigm, $N = 43$ participants were presented with video samples of the beginning, middle and end of the learning process of two types of RL agents learning two different tasks (see Figure 1): a Q learning agent with TD update in a Navigation Task (NT) and a DDPG agent in a manipulation task (MT). For each video, an iteratively developed modular qualitative questionnaire with 11 question blocks was used to assess the participants' observation of agent learning behavior. After this, a total of 832 participant statements were analyzed and clustered in an iterative 3-step analysis process based on thematic analysis [4] and grounded theory [9]. In line with Open Science Recommendations [8, 16], this experiment was preregistered and all materials are openly available on the OSF[1].

3 RESULTS AND DISCUSSION

The analysis revealed four common emergent themes in participants' inferences about the agents' learning processes, each with specific subclusters. A detailed analysis of mentions across all three time points can be found in Figure 2.

Agent Goals: participants infer the agent's learning process to be based on a set of goals. Subclusters include Outcome-related (A), Execution-related (B), Environment-learning (C) and Absence of (D) goals.

Agent Knowledge: participants infer the agent to have different types of knowledge throughout the task. Subclusters include Outcome (A), Procedural (B), Environment (C) and lack of (D) knowledge.

Agent Decision Making (DM): participants infer different ways, the agent takes decisions during the learning process. Subclusters include Undirected (A), Experience-based (B), Expected Outcome-based (C) and Absence of (D) DM.

Agent Learning Mechanisms (LM): participants infer different ways, the agent learns during the task. Subclusters include Exploring (A), Feedback (B), Reasoning (C) and Absence of (D) Learning.

All in all, the experiments revealed four detailed anthropomorphized common themes along which observers organize their inferences about the agent learning process, each with their own set of subclusters. Being highly conceptually interconnected, they seem to fit into a larger framework of observer inferences about agent behavior.

4 CONCLUSION

This work aimed to examine human observer inferences about agent learning processes in order to better understand in Human-in-the-loop RL systems. To this end, a new experimental paradigm was developed, to directly assess observation from a human-centered perspective and applied across different types or RL algorithms and tasks. The results showed that the paradigm was able to collect observer inferences in a reliable way. Furthermore, the experiments revealed a stable framework of four common themes (Goals, Knowledge, Decision Making and Learning Mechanisms) with interrelated subclusters along the lines of which observer inferences about the agent learning process are organized.

This could be the baseline for designing interaction and communication methods to align with this framework of agent perception along the lines of observer inferences to help to make agents even more explainable and align adaption in order to achieve more collaborative synergy. In total, this represents an important step towards Hybrid intelligence.

ACKNOWLEDGMENTS

This research is sponsored by the Hybrid Intelligence project, grant number 024.004.022. Special thanks to Jonne Goedhart, Felix Kleuker and Kevin Godin-Dubois for their support.

[1] https://osf.io/fumd8/?view_only=9cec60dccbd446f08bd818d0b3612705

REFERENCES

[1] David Abel, John Salvatier, Andreas Stuhlmüller, and Owain Evans. 2017. Agent-agnostic human-in-the-loop reinforcement learning. *arXiv preprint arXiv:1701.04079* (2017).

[2] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 8 (2020), 18–28.

[3] Christian Arzate Cruz and Takeo Igarashi. 2020. A survey on interactive reinforcement learning: Design principles and open challenges. In *Proceedings of the 2020 ACM designing interactive systems conference*. 1195–1209.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] 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).

[6] Carlos Celemin, Rodrigo Pérez-Dattari, Eugenio Chisari, Giovanni Franzese, Leandro de Souza Rosa, Ravi Prakash, Zlatan Ajanović, Marta Ferraz, Abhinav Valada, Jens Kober, et al. 2022. Interactive imitation learning in robotics: A survey. *Foundations and Trends® in Robotics* 10, 1-2 (2022), 1–197.

[7] Mohamed Chetouani. 2021. Interactive robot learning: an overview. *ECCAI Advanced Course on Artificial Intelligence* (2021), 140–172.

[8] Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015).

[9] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

[10] Clémence Grislain, Hugo Caselles-Dupré, Olivier Sigaud, and Mohamed Chetouani. 2023. Utility-based Adaptive Teaching Strategies using Bayesian Theory of Mind. *arXiv preprint arXiv:2309.17275* (2023).

[11] Mark K Ho, Fiery Cushman, Michael L Littman, and Joseph L Austerweil. 2019. People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General* 148, 3 (2019), 520.

[12] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. 9–16.

[13] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*. PMLR, 2285–2294.

[14] Arunima Sarin, Mark K Ho, Justin W Martin, and Fiery A Cushman. 2021. Punishment is organized around principles of communicative inference. *Cognition* 208 (2021), 104544.

[15] Andrea L Thomaz and Cynthia Breazeal. 2007. Asymmetric interpretations of positive and negative human feedback for a social learning agent. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 720–725.

[16] Janet Wessler, Tanja Schneeberger, Bernhard Hilpert, Alexandra Alles, and Patrick Gebhard. 2021. Empirical Research in Affective Computing: An Analysis of Research Practices and Recommendations. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. https://doi.org/10.1109/ACII52823.2021.9597418