Can you see how I learn? Human observers' inferences about Reinforcement Learning agents' learning processes

Anonymous Author(s)
Submission Id: 897

ABSTRACT

Reinforcement Learning (RL) agents often exhibit learning behaviors that are not intuitively interpretable by human observers. In settings where a human collaborates with a learning agent, this can result in suboptimal feedback. 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. As a first step, a novel paradigm for assessing observations of RL agent learning behavior was developed and applied in an exploratory qualitative interview study (N=9). Four common emerging themes were observed that users paid most attention to while observing a simple tabular agent solving a navigation task - Agent Goals, Knowledge, Decision Making and Learning Mechanisms. In a second confirmatory mixed-method experiment with repeated measurements (N=34), 816 individual answers were systematically analyzed in an expanded version of the introduced experimental paradigm with users reporting their observations in two different learning settings (a navigation and a manipulation task) each one with a separate underlying RL architecture (tabular and function approximation). The results showed that the paradigm was able to collect observer inferences in a reliable way across tasks and RL algorithms. In addition, the results confirmed and refined the 4 common themes that influence observer perceptions in understanding an RL agent's learning process and examined their interrelations, offering insights for transparency and interpretability in RL and Human Robot-Interaction.

KEYWORDS

Human-in-the-loop RL, Human Agent-Interaction, Human Robot-Interaction, Hybrid Intelligence

1 INTRODUCTION

Hybrid Intelligence (HI) focuses on joint Human-Machine Intelligence systems, "augmenting human and machine intelligence and capabilities to collaborate on tasks together and achieve goals that neither would have been capable on their own" [2].

Human-in-the-loop Reinforcement Learning (RL), where human users enrich the standard RL process by using teaching signals (i.e. [1, 11, 23]), is a prime example of a paradigm in which humans and AI agents collaborate on a task by adapting their behavior to each other's actions towards a joint outcome that benefits both.

While a human teacher can provide insight, expert knowledge or even just common sense to a learning agent [10], human feedback can also be suboptimal, for example, by giving delayed [3] or unbalanced feedback [32]. Since teaching signals are grounded in the teachers' interpretation of the agent's observed learning behavior, part of this misalignment in the teacher-learner loop may also arise from the human teacher's misunderstanding of the agent's learning process [25]. To improve the correct interpretation of the agent's observed learning behavior by human teachers, a deeper understanding of how teachers interpret agent learning processes from observed behavior is essential [9].

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. [9, 20, 25]) or in simulated settings [15]. More importantly, all human user studies on this topic employed paradigms in which teaching signals played an active role. While it is certainly important to maintain the interactive (teaching) framing for such settings, integrating teaching signals in the interaction can be problematic, since it not only leads to teacher expectations about how the agent "interprets" these signals in the first place [20, 29], but also influence agent behavior and thereby the human's subsequent observation. This may obstruct an unbiased examination of the human's perspective of the learning progress, especially at later learning stages.

To the best of our knowledge, no study has so far examined human inferences about the RL agent learning process from behavior observation alone. 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, bottom-up approach to investigate how individuals perceive and interpret the learning process of RL agents in a specifically developed experimental paradigm. In two sequential experiments, 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.

2 BACKGROUND AND RELATED WORK

The mechanisms of Reinforcement Learning (RL) are fundamentally difficult to assess for human observers from learning behavior alone, both in terms of complexity [17] but also in terms of correctly interpreting components such as reward functions in the first place [18]. However, in interactive settings, the perception of the learning behavior influences the interpretation of the learning process and the interaction itself in a recursive manner [25]. For example, Ho and colleagues demonstrate that teachers are strongly

Proc. of the Adaptive and Learning Agents Workshop (ALA 2025), Avalos, Aydeniz, Müller, Mohammedalamen (eds.), May 19 – 20, 2025, Detroit, Michigan, USA, ala-workshop.github.io. 2025.

inclined to use rewards in the teaching process intuitively as a communication instead of a reinforcement signal [19–21], indicating a clear misinterpretation of the agent's learning process. While agents learn through and chase cumulative rewards, the interaction pattern found in their work indicates that teachers fail to see this objective or actively choose to ignore it. Specifically, the researchers showed that the pattern persisted even when it explicitly hindered the agent's learning progress [20]. The authors explained this through assumed differences in interpretations of the state space on the teacher side, as well as the conclusion that teachers' ability judgments about the agent were based on the distance between the learner's policy (as inferred by the teacher) and the teacher's desired policy [20]. Although this may be a legitimate step in the teachers' analysis of how to understand and teach the problem, the creation of positive reward cycles in the teaching process based on the resulting feedback [20] also suggests a detachment between teachers' perception, interpretation and judgment of agent learning processes and the actually underlying RL mechanisms. Later research further shows that the expectation of feedback being organized around principles of inference in interactive settings generalizes to Human Human-Interaction [29].

More generally, the interpretation of behavior in interactive (social) settings "can be characterized as probabilistic inferences guided by an intuitive understanding of how people think, plan, and act", as outlined by the Inferential Social Learning framework [16]. This suggests that individuals acquire knowledge about their interaction partners by inferring insights from behavior observation. Understanding how this process applies in the context of Human Agent-Interaction settings, is therefore crucial in order to design effective Human-in-the-loop RL frameworks. However, teacher inferences about agent learners have been examined mostly indirectly through analyses of interaction behaviors, rather than examined as a direct factor in the process of understanding agent learning.

The goal of this work is therefore to examine directly how human teachers perceive and interpret RL agents' learning processes from inferences drawn about observed learning behavior.

3 EXPERIMENT 1: EXPLORATORY INTERVIEW STUDY

In this first exploratory study, an observational experimental paradigm is developed and its exploratory results presented.

3.1 Methodology

3.1.1 Participants: A total of n=9 participants (n=3 female, n=6 male). Five participants were bachelor students with an emphasis on RL in their course, 4 participants had no background in RL.

3.1.2 Experimental set-up and Procedure: In the absence of a validated paradigm to directly examine observer inferences about agent learning processes, an adapted version of the Grounded Theory framework was used to develop a reliable experimental procedure [14]. For this, an exploratory qualitative interview setting was selected for this experiment, due to the suitability of this method for studying undirected perceptual data [31]. In the presented paradigm, participants observed an RL agent's learning process with a framing of being a teacher but without the possibility of intervention. As a first step, participants received a short briefing about the general idea and functioning of an RL agent with an example video similar to the later stimuli and asked to imagine that they to should help the agent that to learn the task by teaching. Secondly, they were presented with videos of an RL agent learning to solve a task and were given the opportunity to share their observations about the learning process in a think aloud protocol [28] in order to elicit a verbal report that is as undirected, undisturbed and constant as possible [5]. The interviews were conducted in person, audio-recorded and took on average 45 minutes per participant; recordings were erased after an anonymized transcription. A written consent for this procedure was obtained from the participants beforehand. After the experiment, participants received a debriefing.

3.1.3 Stimuli: The agent to be examined was a tabular Q-learning agent with temporal difference update. The stimuli consisted of rendered recordings of this agent solving the two default versions of the OpenAI gym toolkits' *FrozenLake* environment (4x4 and 8x8 grid respectively) [7]. To account for both the dynamically progressing nature of RL but also the aim of a perception of the learning process in a coherent and unified way, two types of stimulus presentation were implemented between participant groups. The first group was presented with a video showing the full learning process from beginning to end (2 Videos), while the second group was presented with the same videos split up in 3 separate stimuli, with each video showing one third of the learning process (6 Videos). Both participant groups (RL experts and lay users) were assigned to these two conditions in a balanced way (2 participants per condition).

3.1.4 Measurements: Questioning focused on participants' observation about the agent learning process in the videos and participants were instructed to explicitly report on the learning process (as opposed to describing the learning behavior). Furthermore, a question regarding their teaching intentions was added in order to continuously remind the participants of the interactive teaching framing (results are reported elsewhere). Since no validated questionnaire was available, the first participant (an RL expert) was used for a piloting session to query the question formulation for the development of the experimental paradigm in line with [14]. Specifically, they were questioned on formulations that queried most sensibly the human inferences about the learning process. Based on these suggestions, two questions were conceived and used subsequently: "*What do you think is going on in the brain of the agent?*" and "*Why do you think the agent is doing what it is doing?*"

3.1.5 Analysis: Data preparation: As a first step, audio recordings were transcribed. Secondly, participant answers were pooled per video, facilitating the full anonymization of the data. In a third step, to allow for a disentangled analysis of the data collected as free speech through the think-aloud method, individual statements were then isolated.

Qualitative Analysis: An iterative, three-step approach of thematic analysis [6] was used to systematically cluster the qualitative reports. As a first step, data dimensionality in participant answers was reduced by extracting a core meaning from each statement. Secondly, initial themes were generated by clustering statements of thematic similarity and defining an initial descriptive label. As a third step, these themes were reviewed again against the full data

set and definitions refined when necessary. After this step number 2 and 3 were repeated a total of 3 times.

3.2 Results

The thematic analysis revealed four common emergent themes in participants' inferences about the agents' learning processes.

Agent Goals: includes statements in which participants infer the agent's learning process for this task is based on a set of goals. This theme was referred to 23 times. Examples include statements such as "*The agent wants to explore and understand the environment that it is in*", "*It is trying to avoid the ice*", "*It is trying to find a way that it can navigate the environment as quick as possible*" and "*seems not to realize what the idea is*".

Agent Knowledge: includes statements in which participants infer that the agent's learning process incorporates different types of knowledge throughout the task. This theme was referred to 36 times. Examples include statements such as "*It seems to have gathered some sort of sense of the second row being a bad row*", "*it remembers, "alright I was in that tile and then I moved to that tile and there was a lake right there"*", "*he knows exactly which tiles to not step onto, because he memorized it*".

Agent Decision Making: includes statements in which participants make inferences about the different ways that the agent takes decisions during the learning process. This theme was referred to 50 times. This includes statements about actions, options (action sequences) but also overall policies such as "*It is just trying out different moves and it only learns something when it actually tried the action*", "*He follows a system: 1 down, one right, then 2 down one right, then three down and falling in*", "*He realized he cannot take the first second rows and thinks whether he can take the third one*".

Agent Learning Mechanisms: includes statements in which participants make inferences about the different ways that the agent learns during interaction with the environment. This theme was referred to 18 times. Examples include statements such as "*it seems that it is just mapping certain actions to certain states*", "*He has moved two tiles to the right in the first try and realized it was not a mistake*".

Other: includes additional themes that appeared with too low frequencies in the participants' answers to be considered their own theme, such as inferences about agent perception mechanisms (i.e. vision), feelings and communicative signals, environment or behavior descriptions and statements about the learning progress.

Overall, data collection with the presented paradigm resulted in 264 isolated statements, of which 127 could be clustered into the four common themes, amounting to roughly 48% explanatory value. Furthermore, the number of statements in the split-up video condition was 3 times higher than the whole video condition, indicating that it is better suited for a collection of differentiated results.

3.3 Discussion

The experiment aimed at understanding observer inferences about agent learning processes with a newly developed paradigm. The data-driven analysis revealed four major common themes: Agent Goals, Agent Knowledge, Agent Decision Making and Agent Learning Mechanisms that participants inferred about the learning process from observing RL agent learning behavior over time.

Agent Goals emerging as a distinctive theme indicates task understanding seems to play a key role in observer inferences of the agent's learning process. Therefore, the theme of Agent Goals could also be interpreted as reflecting the observers' understanding of the task and environment that the agent is trying to optimize and learn. This is especially important since every user approaches each task with their own set of assumptions. An observer misunderstanding of the task at hand might subsequently skew other inferences about the agent's learning process. A first indication are statements like "[reaching the exit] as safely as possible" (which was not part of the reward structure).

The emergence of Agent Knowledge indicates that participants assume, the agent relies on a representation of task understanding and experience. This is important since the inference of knowledge is a latent, less directly visible concept compared to Agent Goals or Agent Decision Making indicating a mental model with (humanized) intelligence capabilities about the agent on the observer side that informs inferences about the agent's learning process. Furthermore, since both Decision Making and Learning Mechanisms have a clear connection to Knowledge, knowledge seems to be a crucial component to understanding observer inference processes about agent learning.

The emergence of Decision Making shows the existence of a procedural view on the agent learning process (as compared to rather descriptive, structural components such as goals and knowledge). It reveals that observers see the agent as a proactive entity that does not simply react to the environment but autonomously influences outcomes. In addition to that, Decision Making is highly interlinked with both Goals and Knowledge, indicating that observer inferences of agent learning are based on an integrated representation.

Lastly, the theme of Learning Mechanisms confirms the observers' integrated and procedural view of agent learning. Further, they appear to understand agents also as adaptive entities.

The anthropomorphized structure of this framework aligns with previous results from cognitive science. An anthropomorphic perception increases an individual's understanding by attributing preexisting knowledge and structures, which could mean, that observers anthropomorphize agent learning in general.

All in all, this initial exploratory study revealed four major common themes along which observers organize their inferences about the agent learning process. Being highly conceptually interlinked, they seem to fit into a larger framework of observer inferences about agent behavior.

Limitations and learned lessons Given the small sample size of the experiment and the exploratory nature of the paradigm, no reliable conclusions can be made about the results so far. Experiential interview data like this can be seen as an indication about the participants' process of interpreting agent behavior, but due to their self-report nature it runs the risk of only indirectly reflecting the participants' authentic underlying inference processes [4]. As a next step, the framework needs to be evaluated in a more systematic, empirical setting with a larger number of participants.

Secondly, while the collected data proved to be rich and detailed (in the split-up condition), up until now the four common themes remain undifferentiated. More data needs to be collected to gain more precise insights both about each of the common themes themselves as well as their interrelations. This includes a refining of

the experimental paradigm with a more differentiated quantitative questioning format. Next, the framework proved to be highly anthropomorphized. This could in part be due to the human tendency to apply pre-existing knowledge to virtual agents in order to facilitate understanding [30]. However, it could also be due to both the stimulus material (i.e. the FrozenLake agent having a human like appearance) and the question being asked (i.e. "What do you think is going on in the *brain* of the agent?). A next step should examine whether the anthropomorphizing of themes replicates in a modification of the paradigm with less anthropomorphized stimuli and questioning. Lastly, this study examined observer inferences about an agent's learning process only for a very simple tabular agent. Since most interactive teaching settings - especially in robotic applications - employ more sophisticated forms of RL agents, future work should also validate this framework for other types of agents.

4 EXPERIMENT 2: CONFIRMATORY QUESTIONNAIRE STUDY

The first exploratory experiment examined observer inferences about an RL agent's learning process and revealed four common themes along which these inferences are made when observing RL agent learning behavior. This second, confirmatory experiment has three primary objectives:

(1) Validating the initial findings on a larger sample size
(2) Adding granularity to the four common themes found in the initial study by examining them in greater detail
(3) Generalizing the findings by expanding the experimental paradigm to a broader range of tasks and RL algorithms

In order to address these research objectives, a larger-scale questionnaire study was implemented.

4.1 Methodology

Following the recommendations for open research practices [12, 33], this experiment was preregistered and all materials (Questionnaire, coding plan, instructions, scripts, stimuli, annotated data corpus and the preregistration protocols) are openly available on the OSF[1].

4.1.1 Participants: Participants were recruited through Prolific. A total of 35 participants was recruited of which one had to be excluded for providing inattentive answers. This resulted in a final sample of N = 34 (26 Male, 8 Female, 0 Non-binary, self-assigned or undisclosed). The participants were between 19 and 65 years old (M = 38.79 years, SD = 10.68 years). 23 participants owned a pet, 20 participants reported to have children. Participants received on average £10.42 GBP. Screening criteria included being 18 years or older, a minimum of a 95% approval rate on Prolific and 100 completed tasks, and self-reported fluency in English, confirmed by residence in English-speaking or high English-fluency countries.

4.1.2 Expanded experimental paradigm: In order to examine whether the results of experiment 1 generalize beyond the initial setting of an anthropomorphic tabular RL agent in a deterministic navigation task, the experimental paradigm was expanded. Firstly, instead of two tabular agents in differently sized grid worlds, the 4x4-maze was substituted with a Deep Deterministic Policy Gradient agent

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

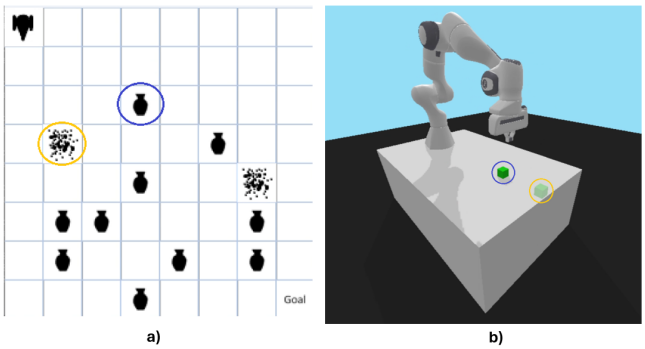

Figure 1: a): NT: The cleaning robot needs to clean the dirt patches (yellow) and avoid breaking the furniture (blue). b) MT: the assistive robot needs to push "medication" (blue) to the target position (yellow).

(DDPG; an off-policy, actor-critic algorithm designed for continuous action spaces) performing a manipulation task. This serves to test whether observer inferences about RL agent learning processes also apply to function approximation algorithms. Secondly, in order to induce more realism to the teaching settings and reduce anthropomorphism, the agents were framed as two types of service robots performing everyday tasks, to test whether the results generalize to non-human, yet realistically teachable tasks.

The navigation task (NT) was now framed as a cleaning robot that had to learn to clean and exit a room without breaking furniture. To this end, the appearance of the 8x8 FrozenLake environment was adapted, substituting the elf by an iconized cleaning robot with conic shape (to indicate orientation), ice tiles by white "floor" tiles, holes in the maze by iconized vases and the treasure chest by a white tile with "exit" inscription (see Figure 1). Additionally, 2 dirt patches were added to the task, implemented as dynamic flags: the agent received an additional reward (+1, equal to the "exit" tile) upon visiting after which they disappeared. This additional milestone in the learning process served to increase face validity of the setting, making the learning progress more visible to the observer. Other than that, the task and algorithm remained the same.

The manipulation task (MT) was framed as an assistive robot in elderly care learning to push a box of "medication" over a table to a target location (where a patient that has trouble reaching could more easily pick it up). The agent performed a modified version of the Push task within the PandaGym environment [13], with fixed starting positions for the object and the target closer to the edge than in the original task and received a dynamic reward for both the distance between the end effector and the object and the distance between the object and the target. Additionally, upon collision between object and target, the agent received a large reward (+1000).

4.1.3 Experimental set-up and procedure: In this experiment participants again observed the RL agents' learning process with a teaching framing. After being recruited through Prolific, participants were forwarded to the landing page of the questionnaire which welcomed them and explained the procedure and options for stopping the experiment. Then, it led them to the informed consent form and a commitment request. In the second step, they

received an introductory description of the general idea and functioning of RL in lay user terms[2], before the questionnaire started. The assessment presented participants with two task scenarios (see Section 4.1.2) with 6 videos in total (see Section 4.1.4) in which an RL agent learned to solve the task. For each video, participants were in turn presented with a block of questions (see Section 4.1.5. After completing the questionnaire, participants were thanked and redirected to Prolific, where they received their payment after 24 hours. The questionnaire took on average 57 minutes to complete.

4.1.4 Stimuli: In order to reflect an actual agent learning process, stimulus videos were created by training each agent until convergence and subsequently rendering roll-out trajectories under the saved checkpoint policies. The training process was considered converged after 5 consecutive successful episodes. In the NT, the agent was trained for 10000 episodes and with checkpoint policies saved every 500 training episodes. In the MT, the agent was trained for 100000 steps, with checkpoint policies saved every 1000 steps. This resulted in two source videos, one lasting 94 seconds (NT) and the other 192 seconds (MT). For each source video, we further divided it into 3 stimulus videos of roughly equal lengths, representing the early, middle, and end phases of the learning process.

4.1.5 Measurements: In order to examine observer inferences about the agents' learning process in greater detail, a modular questionnaire was designed, consisting of a combination of qualitative and quantitative assessments. The questionnaire consisted of 2 parts (1 for each task scenario). In the beginning of each part, participants were briefed on the specific task scenario at hand, including crucial information about the specific objectives of the robotic agent as well as the reward structure of the environment. This was followed by a block of 11 questions presented in pairs of a qualitative and a quantitative query. The video stimulus at hand was available for rewatching above each question pair. The first assessment served as a manipulation check. For each stimulus video participants were asked to rate in percent to what extent they believed the agent had learned the task at hand. This served to assess whether participants perceived a learning progress (LP) in the agent's behavior (i.e. whether they perceived a learning). This was followed by the qualitative question "*What do you think is happening in this video?*", that served as a baseline assessment for qualitative data and to make sure, participants had actually watched the video. After this, each of the four common themes (Goals, Knowledge, Decision Making, Learning Mechanisms) was assessed by a qualitative and a quantitative question. Firstly, a rating question was presented regarding the relevance of the theme for the agent learning at this point in the learning process. Participants answered on a 7-point scale from 1 (*irrelevant*) to 7 (*very relevant*). The qualitative assessment queried participants' inferences about the individual theme in the following way: "*What goals/knowledge do you think the robot has in this video?*" or "*How do you think the robot makes decisions/learns in this video?*". Finally, participants were asked to describe their teaching intentions, emphasizing the interactive framing of the setting. This block of questions was applied for each of the video stimuli in a modular way. The questionnaire concluded with demographic questions assessing age, gender and yes/no questions whether the participants had kids or owned a pet.

4.1.6 Analysis: The analysis process of experiment 1 was applied in similar manner to the newly collected data in order to cluster the statements regarding each of the four common themes.

4.2 Results

This experiment aimed at capturing a more detailed insight into the observers' inferences about the agent's learning process over a more generalized experimental paradigm. Specifically, subjective ratings on LP, the relevance of each previously found common theme at each of the three time points as well as open-ended statements on observer inferences regarding each theme were collected.

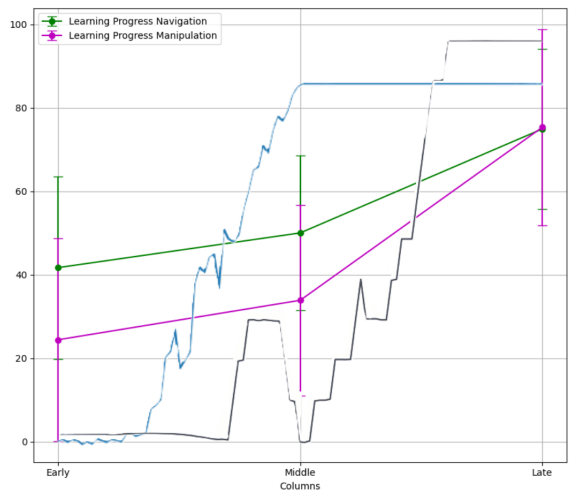

Figure 2: LP. Green = Perceived LP NT, violet = Perceived LP MT, Blue = Actual LP in NT, Black = Actual LP in MT

As a first step, the perceived LP was assessed to ensure that a learning was actually perceived by the observers and therefore validate subsequent inferences drawn about the observed learning process. Results shown in Figure 2 confirm a steady increase in perceived LP for both tasks. Between tasks, observers perceive a faster learning progress in the NT in the beginning, while this development is inverted in the end, mirroring the agents' actual LP (as assessed by cumulative rewards).

A second step of validation consisted of a relevance rating to assess whether the common themes of inferences are also perceived as influential by the observers themselves (see Figure 3). One-sample t-tests revealed that for each of the 3 time points, observers assigned a relevance significantly above a "neutral" rating of 4. Additionally, the subjective relevance even seems to increase as time progresses.

Finally, the qualitative analysis examined thematic clusters of observer inferences within Goals (G), Knowledge (K), Decision Making (DM) & Learning Mechanisms (LM). The results of this analysis will be reported separately for each of the common themes.

[2]Instructions, questionnaire and stimuli were piloted with N=4 test participants before recruiting in order to ensure understandability

A quantification of absolute mentions of each category split by task and time point of assessment can be found in Figure 4. Out of a total of 816 answers across all tasks and time points, 1105 individual statements were extracted and analyzed.

Goals : Observer inferences regarding agent goals could largely be clustered into four categories.

(A) *Outcome-related Goals:* Inferred goals concerning task achievement, that define what constitutes success and failure in the task respectively. Outcome-related goals were mentioned a total of 276 times across all tasks and time points. For the NT, this included all statements about achieving subgoals such as "*find the dirt and clean it*", "*avoid breaking any furniture*" and "*find the exit*" but also "*finding a pathway*". For the MT, such subgoals included "*locate the medicine*", "*hold onto it*", "*not to have it fly off of the table*" as well as "*move the box to the target zone*".

(B) *Execution-related Goals:* Inferred goals concerning improvement and optimization of task execution, that relate to how the agent goes about solving the task (55 mentions). For both tasks, this included specific statements regarding efficiency (e.g. "*to clean the room efficiently*") or speed (e.g. "*the robot wants to push the box quick*") of task solving. Additionally, in the MT, this encompassed statements about solving the task with a high accuracy ("*to achieve accuracy of movement*") and safety ("*safley place it across the table for the patient*") as well as control in executing the pushing ("*To improve the force, angle and direction in which it pushes the box*").

(C) *Learning about Environment:* Inferred the explicit goal of learning about and understanding the environment (13 mentions). For both tasks, this encompassed statements on learning the overall layout (e.g. "*Map the area out in more detail*") and learning about individual components (e.g. "*Learn what constitutes "furniture"*"). Additionally, for the MT, this included statements about testing the limits (e.g. "*how far its arm can extend*"") and understanding possible movements (e.g. "*how the box moves on the table*").

(D) *Absence of Goals:* Statements in which observers explicitly inferred an absence of goals (e.g. "*I'm not sure this robot has any idea of a goal in mind*"), a total of 4 mentions.

A total of 8 statements could not be assigned to one of these clusters, while still referring to some sort of a goal (e.g. "*To help the elderly achieve certain duties that can not be done alone by them*"). Similarly, 9 statements were classified as *Unclear*.

Knowledge : For observer inferences regarding agent knowledge a similar pattern of four clusters emerged.

(A) *Outcome Knowledge:* The agent's understanding of its overall objectives and tasks including specific goals within the environment (86 mentions). For both tasks this included statements of its overall purpose (e.g. "*It knows that it needs to pick up the dirt and to get out especially*", "*it's realised it had to touch the medicine in order to get it to do anything*") as well as the consequences of specific actions (e.g. "*it's dangerous to go near furniture in case it breaks.*").

(B) *Procedural Knowledge:* Inferences describing the agent's understanding of task execution (78 mentions), for both settings describing inferences on how to move (e.g. "*Being able to stretch*"), how to implement behavior (e.g. "*It knows that it needs more deliberate and softer pushes and to stay down on the table more*"), how to interact with the environment (e.g. "*The robot probably has the knowledge to identify which is a vase and which is a piece of dirt*") and how to solve the task (e.g. "*There's some knowledge here in that it learns to do the dirt first before it heads for the exit*").

(C) *Knowledge about the Environment:* Inferences describing the agent's understanding of its environment (89 mentions). For both settings, this included statements about the "*knowledge of the layout of the room*", certain elements (e.g. "*Once it has learned what furniture is*") and their location (e.g. "*location of the target and destination of the target*") and, additionally for the MT, its boundaries (e.g. "*It is using its knowledge on boundaries*").

(D) *Lack of Knowledge:* Statements in which observers explicitly attested a lack of knowledge to the agent (e.g. "*The robot currently does not seem to have enough knowledge*"), 43 mentions in total.

A total of twelve statements were classified as *Unclear*.

Decision Making (DM): Observer inferences regarding the agent's DM could be clustered into four categories.

(A) *Undirected DM:* Inferences in both settings about DM processes that are described as undirected (e.g. "*By trying to move about*"), random (e.g. "*mostly moves at random*"), explorative (e.g. "*The robot learns completely through discovery in this video*") or trial-and-error (e.g. "*It is making its decision through trial and error*"), 49 mentions in total.

(B) *Experience-based DM:* Inferences in both settings describing DM that repeats previous behavior or relies on previous experiences and assumed sensory feedback (65 mentions). For the NT, a strong focus on past mistakes seemed to prevail (e.g. "*due to already having experienced many mistakes i.e wrong moves, breaking vases*"), while for the MT, such statements included more general inferences (e.g. "*by following what it has learned overtime*"). For both tasks repetition (e.g. "*by deciding to repeat movements that have been successful*") and sensing (e.g. "*The robot is making its decisions based seeing the object there*") were inferred to be the basis of DM.

(C) *Expected Outcome-based DM:* Inferences that assume the DM to be based on expected outcomes (90 mentions). For both settings, this included statements such as DM to target specific subgoals (e.g. "*Once it cleans both spills, it makes the decision to go to the exit as quickly as possible.*"), perform systematic exploration (e.g. "*the robot is systematically learning the shape of the room and the objects within*") or DM based on planning or weighing of expected outcomes (e.g. "*Calculating the right path where the dirt is*")

(D) *Absence of DM:* Explicitly inferred absence of DM (e.g. "*I don't think the robot is making decisions*"), 28 mentions.

A total of ten statements were classified as *Unclear*.

Learning Mechanisms (LM): Observer inferences regarding the agent's LM could be clustered into 4 categories.

(A) *Learning by Exploring:* Inferences that assume the agent to be learning the task by exploring (59 mentions). This includes learning by trial-and-error and exploration behavior. It indicates a specific focus on learning through new experiences. For both settings, statements such as "*It learns through trial and error*", "*By exploring around*" and "*The robot learns by trying different patterns*" appeared.

(B) *Learning by Feedback:* Inferences that assume the agent in both settings to be learning by feedback it gets on its behavior (81 mentions), including general feedback (e.g. "*learning from the decisions it has been making throughout the whole learning process*"), reward and punishment (e.g. "*Via positive and negative feedback*"), but also statements about learning by repeating successful patterns

(e.g. "*by repeating its moves*") and practice (e.g. "*the robot learns by practicing*"). Further, this also encompasses learning from assumed sensory feedback (e.g. "*it learns by perception*").

(C) *Learning by Reasoning:* Inferences that assume the agent in both settings to be learning from processes of higher reasoning (e.g. "*It tries to detect the furniture which it breaks to understand where it is*"), internal representation (e.g. "*remembering where the vases, patches and exit are located*"), abstraction (e.g. "*I think this time it split the room into four quadrants and went from right to left learning where the obstacles were*") or from pieces of memory (e.g. "*storing the data of everything learnt [sic!]*"), 27 mentions.

(D) *Absence of Learning:* Statements in which observers explicitly denied the agent to engage in learning (e.g. "*I see no evidence of learning here*"), a total of 12 mentions.

Additionally, in eight cases observers explicitly attributed Machine Learning or RL mechanisms to the agent. Further, a total of ten statements were classified as *Unclear*.

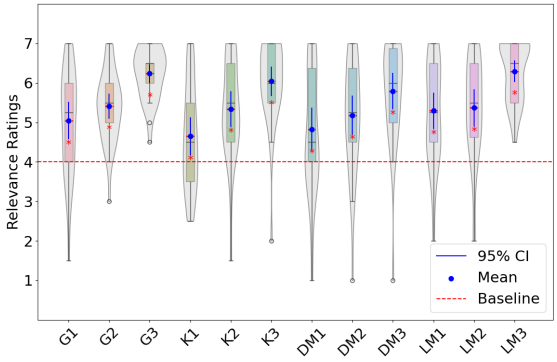

Figure 3: Relevance ratings across three time points for Goals (G), Knowledge (K), Decision Making (DM) & Learning Mechanisms (LM). The dashed line represents a neutral relevance rating of 4, red stars a significance testing against this value

4.3 Discussion

This experiment had 3 main goals:

(1) Validating the set of four common themes found in experiment 1 around which observer inferences about an agent's learning process is organized: Inferences about Goals, Knowledge, Decision Making and Learning Mechanisms.
(2) Adding granularity to these four common themes by examining the observer inferences in greater detail
(3) Generalizing the findings by expanding the experimental paradigm to a broader range of tasks and RL algorithms

The experimental paradigm in this study was designed to collect data in a direct way, unbiased by interaction and resulting expectations. The analysis not only revealed that participants rated the four common themes to be of high relevance for but even of generally increasing relevance over the course of the learning process, confirming the importance of these themes also for the subjective perception of observers. Furthermore, for each of the four common themes, participant reports were differentiated and numerous enough to classify at least 4 conceptual subcategories per theme. Lastly, the experimental paradigm was not only able to produce this differentiated picture in general. In addition, the four common themes including their subcategories could be found across two types of RL algorithms and tasks.

On a conceptual level, this experiment confirmed that observers seem to make sense of RL agent's learning processes along their interpretation of Agent Goals, Knowledge, Decision Making and Learning Mechanisms. These four themes seem to reflect a mental framework that observers apply to the agent and that helps them to make sense of agent learning behavior. Furthermore, assuming these four common emerging themes as a high level framework of understanding agent learning, this experiment also revealed subclusters of inferences that connect the four abstract themes with practical issues in the learning tasks at hand and demonstrate interdependencies between them. While the first experiment suggested that participant observations distinguish between what the agent intends to do and what it knows, the larger study revealed that both of these concepts had considerable overlap in their subclusters. Both of them included inferences about the task-description and outcomes to be achieved on the one hand, and execution- or procedure-related inferences on the other as well as very specific inferences of the agent understanding its environment, which seems to play a major role in observer inferences about RL. Their distribution of mentions across time confirms these interconnections (see Figure 4). However, while these themes strongly overlap on a conceptual level, the formulations of these inferences (e.g. what the agent *wants* to achieve vs. what it *knows* it should do) indicate, that the observers' way of applying these themes still intents a differentiated understanding. Furthermore, the subclusters being organized along outcomes, environment understanding and experience sampling across the four themes, indicates that the distinction between task-affordances and task execution seems to be superficial and suggests that while observers use these different themes to draw differentiated conceptual inferences about the agent learning, their overall perspective seems to be more integrated. This also makes sense on a theoretical level: In a learning setting as the one presented, any learning *per definitionem* reflects a goal evaluation and updates any type of knowledge, while decision making is in turn based on knowledge and thrives to achieve a goal, creating a next state that is to be evaluated. While the distinction between the concepts therefore remains on a descriptive level, this process of "translation" into distinct, anthropomorphized concepts may help inferences approaching agent learning in a human-understandable way. In other words, it can be suspected that the observational concepts that observers employ to describe their observations, serve as an anthropomorphizing explanatory framework for understanding the algorithmic RL process that the participants observed. While individual aspects of this such an interpretative framework have been reported before [24], this work offers a systematic framework of observer inferences about RL processes that indicates an integrated representation of a proactive and adaptive interaction partner in the observers. On a theoretical level, this underlines that the process of inferential social learning [16], that has been examined for Human Human-Interaction [29] and Agent Agent-Interaction [15] may also give useful insights into Human-in-the-loop RL - with positive [27] and negative [20] effects for interactive teaching settings.

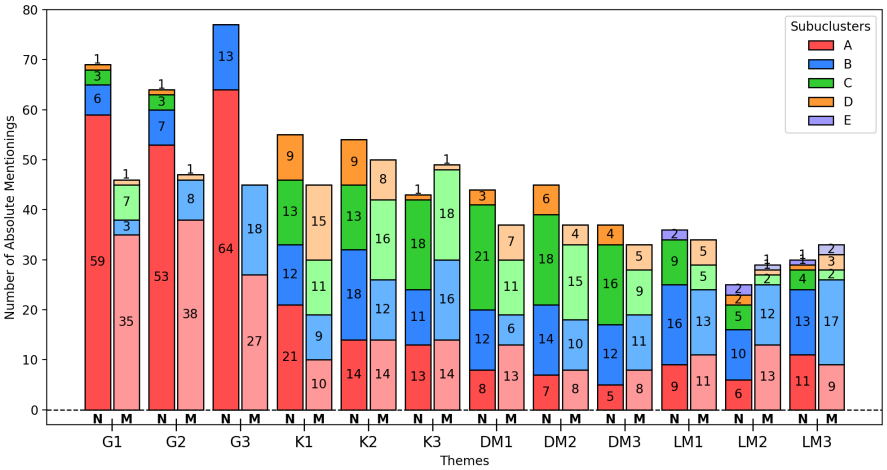

Figure 4: Themes + Subclusters across time split by task. Letters correspond to subclusters for each theme reported in section 4.2

5 IMPLICATIONS FOR FUTURE WORK

These results provide an important first step in understanding how observers perceive and interpret agent learning processes. However, the data was collected in deliberately observational experiments. Future work now has to determine whether these insights translate into an interactive setting, especially how this framework of common themes along which observer inferences about the learning process changes when teachers expectations about how teaching signals are received come into the mix. All in all, we caution researchers to interpret these results more like an approximate guideline rather than a perfectly accurate description of the creation of observer inferences.

However, the results of this work encourage a human-first approach that reveal not only areas of interest for the human observer, but also specific learning related categories towards which transparency and agent-teacher communication methods could be tailored. This could be done by specifically designing methods that try to add transparency into a human understanding of agent goals, knowledge, decision making and learning mechanisms. Another way to utilize these findings is to specifically identify needs for transparency and better communication by targeting areas where human understanding and agent processes diverge. One example could be the distinction of the "artificial" separation of experience-based decision making from outcome-based decision making learning about the environment. On the other hand, a deeper understanding of the teachers' perspective on agent functioning, could potentially organize the analysis of how the teachers think their teaching signal will be processed, which may play a role in choice of teaching strategy or feedback signal. As a next step, these findings could inform how transparency mechanisms and agent-teacher communication methods should be designed (e.g. [8]). While there is a considerable amount of recent work on making RL and interactive robot learning more transparent [26] and human understandable (e.g. [17, 22]), most of these methods are founded in the architecture and functioning of the learning algorithms. A concrete example could be to use the insights about a teacher's expectation of what kind of goals it is trying to achieve to prevent teachers creating positive reward cycles in the teaching interaction [20] by specifically correct this misalignment in terms that align with observer inferences about agent learning processes. Overall, transparency mechanisms should be designed along the lines of observer inferences to which the insights in this work could contribute.

6 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 leveraging joint human and machine intelligence and working towards Hybrid intelligence.

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] Paul Atkinson and David Silverman. 1997. Kundera's Immortality: The interview society and the invention of the self. *Qualitative inquiry* 3, 3 (1997), 304–325.

[5] Ted Boren and Judith Ramey. 2000. Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication* 43, 3 (2000), 261–278.

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

[7] G Brockman. 2016. OpenAI Gym. *arXiv preprint arXiv:1606.01540* (2016).

[8] Joost Broekens and Mohamed Chetouani. 2019. Towards transparent robot learning through TDRL-based emotional expressions. *IEEE Transactions on Affective Computing* 12, 2 (2019), 352–362.

[9] 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).

[10] 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.

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

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

[13] Quentin Gallouédec, Nicolas Cazin, Emmanuel Dellandréa, and Liming Chen. 2021. panda-gym: Open-Source Goal-Conditioned Environments for Robotic Learning. *4th Robot Learning Workshop: Self-Supervised and Lifelong Learning at NeurIPS* (2021).

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

[15] 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).

[16] Hyowon Gweon. 2021. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in cognitive sciences* 25, 10 (2021), 896–910.

[17] Soheil Habibian, Antonio Alvarez Valdivia, Laura H Blumenschein, and Dylan P Losey. 2023. A review of communicating robot learning during human-robot interaction. *arXiv preprint arXiv:2312.00948* (2023).

[18] Donald Joseph Hejna III and Dorsa Sadigh. 2023. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*. PMLR, 2014–2025.

[19] Mark K Ho. [n.d.]. Teaching Agents with Evaluative Feedback: Communication versus Reward. ([n. d.]).

[20] 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.

[21] Mark K Ho, Michael L Littman, Fiery Cushman, and Joseph L Austerweil. 2015. Teaching with rewards and punishments: Reinforcement or communication?. In *CogSci*, Vol. 3. 3.

[22] Muhan Hou, Koen Hindriks, AE Eiben, and Kim Baraka. 2024. " Give Me an Example Like This": Episodic Active Reinforcement Learning from Demonstrations. *arXiv preprint arXiv:2406.03069* (2024).

[23] 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.

[24] Hana Kopecka, Jose Such, and Michael Luck. 2024. The role of perception, acceptance, and cognition in the usefulness of robot explanations. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 7868–7876.

[25] 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.

[26] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2024. Explainable reinforcement learning: A survey and comparative review. *Comput. Surveys* 56, 7 (2024), 1–36.

[27] Sara Moussawi, Marios Koufaris, and Raquel Benbunan-Fich. 2021. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets* 31, 2 (2021), 343–364.

[28] Sanghee Oh and B Wildemuth. 2009. Think-aloud protocols. *Applications of social research methods to questions in information and library science* (2009), 178–188.

[29] 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.

[30] Tanja Schneeberger, Mirella Scholtes, Bernhard Hilpert, Markus Langer, and Patrick Gebhard. 2019. Can social agents elicit shame as humans do?. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 164–170.

[31] David Silverman. 2017. How was it for you? The Interview Society and the irresistible rise of the (poorly analyzed) interview. *Qualitative research* 17, 2 (2017), 144–158.

[32] 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.

[33] 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