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

Bernhard Hilpert
Joost Broekens
Leiden University
Leiden, Netherlands
{b.hilpert,d.j.broekens}@liacs.leidenuniv.nl

Muhan Hou
Kim Baraka
Vrije Universiteit Amsterdam
Amsterdam, Netherlands
{m.hou,k.baraka}@vu.nl

ABSTRACT

Reinforcement Learning (RL) agents often exhibit learning behaviors that are not intuitively interpretable by human observers, which can result in suboptimal feedback in collaborative teaching settings. 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 of human observers' understanding of the agent's learning process. A novel, observation-based paradigm to directly assess human inferences about agent learning was developed. In an exploratory interview study (N=9), we identify four core themes in human interpretations: Agent Goals, Knowledge, Decision Making, and Learning Mechanisms. A second confirmatory study (N=34) applied an expanded version of the paradigm across two tasks (navigation/manipulation) and two RL algorithms (tabular/function approximation). Analyses of 816 responses confirmed the reliability of the paradigm and refined the thematic framework, revealing how these themes evolve over time and interrelate. Our findings provide a human-centered understanding of how people make sense of agent learning, offering actionable insights for designing interpretable RL systems and improving transparency in Human-Robot Interaction.

KEYWORDS

Interactive RL, Human Robot-Interaction, Hybrid Intelligence

1 INTRODUCTION

Hybrid Intelligence (HI) refers to collaborative human–machine systems that combine strengths to accomplish tasks neither could perform alone [2]. A key example is Human-in-the-loop Reinforcement Learning (RL), where human teachers guide learning agents using signals such as feedback or demonstrations [1, 12, 28]. While humans can provide insight, expert knowledge, or common sense [11], feedback is often suboptimal—delayed [3], inconsistent [39], or based on misinterpretation of the agent's behavior [30]. These signals reflect the teacher's view of the agent's learning process—views which may be inaccurate or biased [10]. Prior work has only examined such inferences indirectly, via interaction patterns (see e.g. [10, 23, 30]) or simulated settings [16]. Moreover, existing paradigms confound observation with interaction: when humans actively teach, their expectations shape not only the feedback but also the agent's behavior, making it hard to isolate what they actually understand [23, 35]. 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 which aspects of learning do humans attend to when making sense of agent behavior.

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 takes a human-centered, bottom-up approach by introducing a novel observational paradigm specifically designed to elicit human inferences about RL agent behavior. Across two experiments — spanning different tasks and RL algorithms— a set of four recurring inference themes are identified and validated: Agent Goals, Knowledge, Decision Making, and Learning Mechanisms. A detailed analysis further examines how these themes evolve over time and contribute to structuring observers' understanding of agent learning.

2 BACKGROUND AND RELATED WORK

The mechanisms of Reinforcement Learning (RL) are inherently difficult for human observers to interpret from behavior alone, both due to algorithmic complexity [18] and the challenge of intuitively interpreting concepts like reward functions [19]. In interactive settings, a teacher's perception of an agent's behavior can recursively shape their interpretation of the learning process and thereby influence the course of the interaction [30]. For instance, Ho and colleagues demonstrate that teachers are strongly inclined to use rewards in the teaching process intuitively as a communicative signals rather than a reinforcement signal [22–24], indicating a persistent misalignment between users' assumptions and actual RL mechanisms. Although RL agents are designed to optimize cumulative rewards, the observed interaction pattern suggests that human teachers seem to interpret their actions as if the agent were responding to the communicative intent rather than reinforcement signals. Specifically, the researchers showed that the pattern persisted, even when it explicitly hindered agent learning progress [23]. The authors explained this

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.

through assumed differences in how teachers interpret the agent's state space, and that teachers seem to evaluate the agent by comparing its inferred policy to the teacher's own desired policy [23]. While such judgments may reflect a reasonable teaching strategy from the human perspective, the formation of positive reward cycles in the teaching process based on the resulting feedback [23] highlights a disconnect between human interpretations of learning and the underlying mechanisms of RL.

Further work indicates that these communicative assumptions about feedback generalize to Human-Human Interaction, where punishment and reward are also interpreted through inference [35]. More broadly, the Inferential Social Learning framework describes behavior interpretation as a process of "probabilistic inferences, guided by an intuitive understanding about how people plan and act" [17]. 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, research in Human-Agent Interaction has primarily studied teacher inferences indirectly, by analyzing interaction behaviors. The inferences (what teachers actually think about how an agent learns) have rarely been studied directly.

The present work addresses this gap by directly examining how human observers infer and interpret the learning processes of RL agents from behavior observation alone.

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 (n=3 female, n=6 male) were recruited. Five were bachelor students with RL exposure, four had no RL background.

3.1.2 Set-up & 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 [15]. For this, an exploratory qualitative interview setting was selected for this experiment, due to the suitability of this method for studying undirected perceptual data [38]. Participants observed an RL agent's learning process with a framing of being a teacher, but no possibility for intervention. Firstly, participants received a short introduction to RL together with an example video similar to the later stimuli and were asked to imagine helping the agent 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 [33] in order to elicit a verbal report that is as undirected, undisturbed and constant as possible [6]. Interviews (on average 45 minutes were conducted in person, audio-recorded; recordings were erased after an anonymized transcription A written consent for this procedure was obtained from the participants beforehand; after the experiment, they were debriefed.

3.1.3 Stimuli: The agent to be examined was a tabular Q-learning agent with temporal difference update. Stimuli consisted of rendered recordings of this agent solving the default 4x4 and 8x8 grids of the OpenAI gym toolkits' *FrozenLake* environment [8]. To support coherent perception of the learning process, but also assess whether chunking affects richness of observations, two presentation formats were used. One group viewed two full-length videos (entire learning process); the other viewed six shorter clips, each showing a third of the process. Participants were balanced across these conditions and RL backgrounds (2 per condition).

3.1.4 Measurements: Participants were instructed to explicitly report on the agent's learning process (rather than learning behavior) while watching the videos. With no validated questionnaire available, a pilot session with an RL expert helped refine question phrasing in line with Grounded Theory principles [15], sepcifically formulations that queried most sensibly the human inferences about the learning process. Based on their input, two open-ended questions were conceived: "*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?*" Furthermore, an additional question about teaching intentions reinforced the interactive teaching framing (results reported elsewhere).

3.1.5 Analysis: Data preparation: Audio recordings were transcribed and participant answers were pooled per video, to ensure full anonymization. Individual statements were then isolated to enable systematic analysis of the think-aloud data.

Qualitative Analysis: An iterative, three-step thematic analysis [7] was used to systematically cluster qualitative reports. First, the core meaning of each statement was extracted. Then, thematically similar statements were clustered and labeled. Finally, themes were reviewed and refined against the full dataset. This process of clustering and refinement was repeated three times to ensure stability.

3.2 Results

Data collection with the presented paradigm resulted in 264 isolated statements. Thematic analysis revealed four common emergent themes in participants' inferences about agent learning processes in the data, which together accounted for 127 (approximately 48%) of all statements.

Agent Goals: includes statements in which participants infer the agent's learning process is based on a set of goals. It was identified in 23 participant statements. 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 involved acquiring and using knowledge about the environment or task (36 mentions). Examples include "*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 infer how the agent takes decisions during the learning process, including individual actions, action sequences, and overall policies (50 mentions). Examples include "*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 infer mechanisms by which the agent updated its internal representations during the learning process (18 mentions). Examples include "*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 were mentioned too infrequently to be classified as separate themes. These include inferences about the agent's perception (e.g., vision), feelings, communicative signals, descriptions of the environment or its behavior, and general remarks about learning progress.

Participants in the split-up video condition produced approximately three times as many statements as those in the full-length video condition, suggesting that this format supports more differentiated and richer observer inferences in future applications of the paradigm.

3.3 Discussion

This experiment introduced a new observational paradigm to examine how human observers infer an agent's learning process from its behavior. A data-driven analysis revealed four recurring themes in these inferences — Agent Goals, Knowledge, Decision Making, and Learning Mechanisms — offering insight into how participants made sense of reinforcement learning agents 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. It reveals observer assumptions about the agent's understanding of the task structure, but could also reflect the observers' own interpretation of the task and environment. This is especially important, as every user approaches each task with their own set of assumptions, which can shape or even distort their interpretation of learning. For instance, inferring that the agent aims to reach the exit "as safely as possible," despite safety not being encoded in the reward structure. This highlights a broader challenge in human-agent interaction, where users project task goals that may misalign with an agent's actual goals.

The emergence of Agent Knowledge indicates that participants inferred some form of internal representation of its past experiences or task structure. Unlike Goals, knowledge is a latent, less directly observable concept, yet participants appeared to attribute it readily. This indicates a more cognitively rich, anthropomorphized mental model of intelligence of the agent, using assumed knowledge to explain how learning unfolds. To what extent these inferences align with the agent's actual internal representations remains an open question.

The emergence of Agent Decision Making indicates that observers constructed a procedural view of the agent's learning process, beyond only structural components like goals and knowledge. Rather than seeing the agent as passively reacting to stimuli, participants infer it actively weighing options, selecting actions, or following strategies. This suggests that observers interpret learning not just as an internal update, but as an active process of deliberation — attributing a sense of control or strategy to the agent, similar to human-like decision making. This may hint at a potential over-attribution of deliberate strategy to the agent's processes.

The emergence of Learning Mechanisms suggests that observers inferred not just behavior, but processes of adaptation. This portrays agents as actively learning entities.

Taken together, the four themes suggest that observers employ an integrated, procedural view of agent learning. This framework appears to be anthropomorphized, aligning with prior findings in cognitive science [17, 23] and Human-Agent Interaction [5, 21, 27, 36] showing the attribution of human-like structure to make complex behavior more understandable.

Limitations and learned lessons Given the small sample size of the experiment and the exploratory nature of the paradigm, the findings should be viewed as preliminary. While the interview data provide useful insights into how participants interpret agent learning, their self-reported nature may only indirectly capture actual inference processes [4]. A larger-scale study is needed to validate these initial results in a more systematic and robust manner.

Secondly, while the collected data in the split-up condition proved to be rich, the resulting themes remain broad and undifferentiated. To develop and expand them, the next step requires refining the experimental paradigm with a more structured, 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 to facilitate understanding [37], but also due to the stimulus material (i.e. the human-like agent in FrozenLake) and the question formulation (i.e. referring to the agent's brain processes). 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

Building on the exploratory findings of Experiment 1, this second experiment aimed to confirm and extend those results through a larger-scale study. Specifically, it pursued three primary objectives:

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

To address these objectives, a larger-scale confirmatory questionnaire study was designed and implemented.

4.1 Methodology

In line with open research practices [13, 40], 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. Inclusion criteria required participants to be 18 years or older, with a minimum of 95% approval rate on Prolific and 100 completed tasks, and self-reported fluency in English, confirmed by country of residence. A total of 35 participants was recruited of which one was excluded due to inattentive responses, resulting in a final sample of $N = 34$ (26 Male, 8 Female, 0 Non-binary, self-assigned or undisclosed). The participants ranged from 19 to 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.

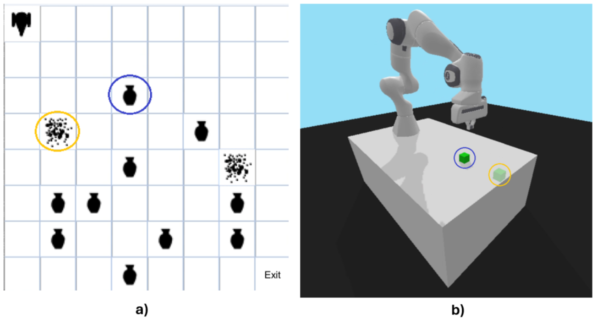

Figure 1: a): NT: The cleaning robot needs to clean the dirt patches (yellow) and avoid breaking the furniture (blue). b) MT: the assistive robot needs to push "medication" (blue) to the target position (yellow).

4.1.2 Expanded experimental paradigm: To test the generalizability of the findings in the first study, the experimental paradigm was expanded in two key ways. First, a second task involving a Deep Deterministic Policy Gradient (DDPG; an off-policy, actor-critic algorithm for continuous action spaces) agent was introduced, to assess whether human inferences about agent learning extend to function approximation settings. Second, to reduce anthropomorphism and increase ecological validity, both tasks were reframed as service robots performing everyday activities (see Figure 1).

In the navigation task (NT), the original tabular Q-learning agent operated in a modified 8x8 FrozenLake environment. Visual elements were adapted to depict a cleaning robot navigating a room: the elf was replaced with a conic robot icon (indicating orientation), ice tiles with white floor tiles, holes became vases, and the goal tile marked with an "exit" label. Two dynamic dirt patches were added, granting the agent an additional reward (+1) upon visit and disappearing afterwards, providing visible milestones to support observer inferences. The task logic and underlying algorithm remained unchanged.

The manipulation task (MT) featured a DDPG agent framed as an assistive robot pushing a box of medication across a table to a target zone (where a patient that has trouble reaching could more easily pick it up). This was implemented using a modified Push task from the PandaGym environment [14], with fixed object/target positions near the table's edge. The reward function combined distance-based shaping rewards with a large terminal reward (+1000) upon successful delivery.

4.1.3 Set-up & Procedure: Participants observed the RL agents' learning in a non-interactive, teaching framing. After recruitment via Prolific, they were directed to a study landing page outlining the procedure, ethical rights, and withdrawal options, followed by an informed consent form and commitment check. Next, participants then received a lay-level introduction to reinforcement learning concepts, designed to ensure baseline comprehension.[2] They were then presented with two task scenarios (see Section 4.1.2), each comprising three videos of an agent's learning trajectory (see Section 4.1.4), for a total of six videos. Following each video, participants completed a modular block of questions (see Section 4.1.5). The questionnaire took an average of 57 minutes to complete, after which, participants were redirected to Prolific and received compensation within 24 hours. This exploratory phase served to refine the observational paradigm and ground subsequent quantitative inquiry in authentically reported user perceptions. The study was approved by the university's ethics board.

4.1.4 Stimuli: To reflect genuine learning processes, stimulus videos were generated by training each agent to convergence and rendering roll-outs from saved checkpoint policies. Convergence was defined as five consecutive successful episodes. For the NT, the agent was trained for 10000 episodes, with checkpoints saved every 500 episodes. For the MT, the agent was trained for 100000 steps, with checkpoints saved every 1,000 steps. This yielded two full-length videos: 94 (NT) and 192 seconds (MT). Each source video was divided into

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

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

three equal-length clips representing the early, middle, and late stages of learning, resulting in six stimulus videos in total.

4.1.5 Measurements: To examine observer inferences about the agents' learning process in greater detail, a modular questionnaire was designed, combining qualitative and quantitative assessments. It was divided into two equal blocks, one per task scenario. Each block began with a briefing describing the agent's objectives and the environment's reward structure. This was followed by 11 question pairs—each comprising a qualitative and a corresponding quantitative query—presented after each stimulus video. The video stimulus at hand was available for rewatching above each question pair. The first item served as a manipulation check: participants rated, in percent, how much they believed the agent had learned the task, to assess perceived learning progress (LP). A baseline qualitative prompt—*"What do you think is happening in this video?"*—followed, ensuring participants had engaged with the stimulus. Each of the four inference themes (Goals, Knowledge, Decision Making, and Learning Mechanisms) was then assessed using a 7-point relevance rating (1 = *irrelevant*, 7 = *very relevant*), followed by a theme-specific open-ended question. For example: *"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, reinforcing the interactive framing of the setting. This block of questions was repeated for each of the six videos. The questionnaire concluded with demographic questions assessing age, gender, and yes/no questions on pet ownership and having children.

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

4.2 Results

This experiment examined observers' inferences about agent learning processes across two task scenarios and three learning stages. Subjective ratings on perceived learning progress (LP), theme relevance over time, and open-ended responses linked to the four common inference themes were collected.

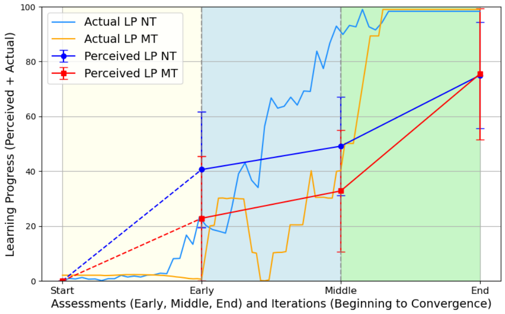

Figure 2: Perceived vs. Actual Learning Progress

As a first step, perceived LP was assessed to confirm that participants recognized agent improvement, validating subsequent inference analysis. Figure 2 shows, that perceived LP increased steadily over time in both tasks. Early in the learning, observers rated the NT agent as learning more quickly; later, this pattern reversed, with the MT agent rated as showing stronger progress. This trajectory closely mirrored the agents' actual LP curves based on 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, relevance ratings showed an increasing trend over time, suggesting growing salience of these themes as learning progressed.

Finally, the qualitative analysis examined thematic clusters of observer inferences within Goals (G), Knowledge (K), Decision Making (DM) & Learning Mechanisms (LM) (for an overview see Table 1). Figure 4 presents the distribution of subcluster mentions by task and learning phase. From 816 participant responses, a total of 1,105 individual statements were extracted and coded, reflecting instances where responses contained multiple distinct inferences.

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 aimed to improve or optimize its task performance (55 mentions). For both tasks, this included specific statements about efficiency (e.g. "*to clean the room efficiently*") or speed (e.g. "*the robot wants to push the box quick*").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*

Goals	Knowledge	Decision Making	Learning Mechanisms
Outcome-related Goals	Outcome Knowledge	Undirected DM	Learning by Exploring
Execution-related Goals	Procedural Knowledge	Experience-based DM	Learning by Feedback
Learning (about Environment)	Knowledge about Environment	Expected Outcome-based DM	Learning by Reasoning
Lack of Goals	Lack of Knowledge	Absence of DM	Absence of Learning

Table 1: Four common inference themes and their respective subclusters

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 knowledge on how to move (e.g. "*Being able to stretch*"), implement behavior (e.g. "*It knows that it needs more deliberate and softer pushes and to stay down on the table more*"), 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 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 (e.g. "*The robot currently does not seem to have enough knowledge*"), 43 mentions.

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*.

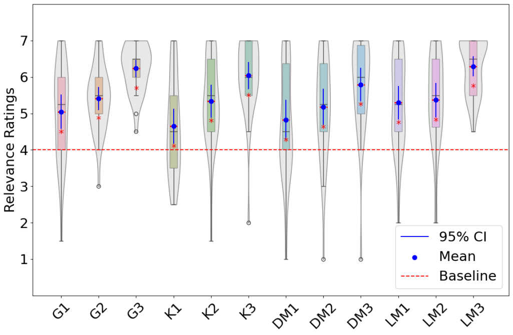

Figure 3: Relevance ratings across three time points for Goals (G), Knowledge (K), Decision Making (DM), and Learning Mechanisms (LM). Dashed line: neutral rating of 4; red stars: significant deviation from neutrality

4.3 Discussion

This experiment had 3 main goals: Validating the four common themes, found in experiment 1, adding granularity by examining the observer inferences in greater detail and generalizing the framework across multiple RL tasks and algorithms. The experimental paradigm in this study was designed to collect data in a direct way, unbiased by interaction and resulting expectations. Participants rated all four inference themes as highly relevant to the agent's learning process, with perceived relevance generally increasing across learning stages. This supports their importance not only conceptually, but also in terms of how observers subjectively make sense of agent behavior, which validates the thematic structure identified in Experiment 1. Participant responses were differentiated enough to identify at least four conceptual subclusters within each of the four themes. The findings generalized across both task types (navigation, manipulation) and RL algorithms (tabular, function approximation), further supporting the robustness of the experimental paradigm for eliciting observer inferences.

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. Together, they seem to reflect a coherent teacher mental model, consistent with prior work in Human-Robot Interaction [5, 34]. While individual aspects have been reported before [29], this work offers a systematic framework of observer inferences about RL processes that indicates an integrated representation of a proactive and adaptive interaction partner. The first experiment suggested that participant observations distinguish between what the agent intends to do and what it knows. Notably, they reflect a meaningful distinction between structural assumptions (e.g., goals and knowledge) and procedural ones (e.g., decision making and learning mechanisms). However, the larger study revealed considerable overlap: participants made both task-related and execution-based inferences within each theme, attributing to the agent a detailed understanding of its environment. Furthermore, building on these four common themes as a high-level framework of understanding agent learning, this experiment also revealed subclusters of inferences that ground each theme in the specific demands of the learning tasks and demonstrate interdependencies between them. The distribution of these inferences over time supports a mental model that is both interconnected and dynamic, with distinct yet evolving components (see Figure 4). On a theoretical level, this integrated framing makes sense: learning inherently updates knowledge, supports decision making, and reflects goal evaluation. While these functions are distinct descriptively, the observer's use of anthropomorphized categories may serve as an explanatory bridge, making algorithmic learning processes more understandable. This aligns with social learning theories, suggesting that the inferential frameworks humans use to understand other minds [17], previously studied in Human-Human [35] and Agent-Agent settings [16], may also shape their interpretations of artificial agents. When these inferences misalign with an agent's actual learning mechanisms, they can lead to mismatches in expectations and behavior—potentially resulting in counterproductive teaching strategies [23]. This highlights the importance of designing systems that anticipate and adapt to human assumptions.

5 IMPLICATIONS FOR FUTURE WORK

These results provide an important first step understanding how observers perceive and interpret agent learning processes. While the thematic analysis was based on subjective coding, it followed a rigorous, iterative method with high internal stability. Nonetheless, future replications should incorporate inter-rater reliability to further strengthen the framework. Although the study included both RL experts and non-experts, exploratory comparison did not suggest major differences in theme salience. Yet this may warrant systematic analysis in future work. Further, the data was collected in deliberately

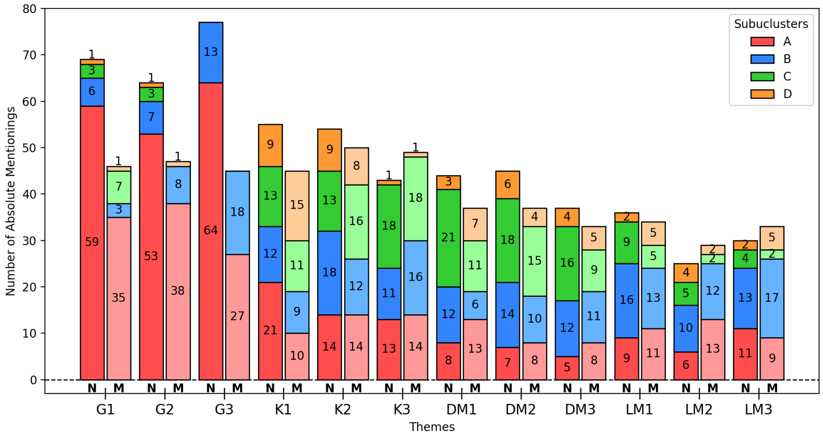

Figure 4: Themes+Subclusters across time split by task. Letters correspond to subclusters for each theme, see 4.2

observational experiments. All in all, researchers should interpret the present results as an approximate guideline for understanding observer inferences, not a definitive model.

A key next step is to validate this framework in more complex, real-world HRI settings where observer inferences may interact with task realism, embodiment, and time constraints. It remains to be seen how these themes shift when users move from passive observation to active teaching, especially as their expectations about how agents receive and use teaching signals come into play. Connecting observer inferences with actual teaching strategies could reveal systematic patterns and help anticipate breakdowns in communication.

Future analyses could link observer inferences more directly to the agent's internal learning state or policy structure, to assess where alignment or divergence occurs. An important extension of this work could be to formalize a metric of human-centric interpretability grounded in the four-theme framework, enabling direct evaluation of agent transparency from a user perspective. Further, the framework allows modeling the users' current agent interpretation to facilitate real-time perspective alignment [31]. These insights also offer concrete implications for system design. The identified subclusters suggest specific areas where users tend to form strong—but potentially inaccurate—assumptions. For instance, distinguishing experience-based from outcome-based decision making could clarify misunderstandings about agent behavior.

While there is a considerable amount of recent work on making RL and interactive robot learning more explainable [32] and human understandable (e.g. [18, 25, 26]), most of these methods are founded in the architecture and functioning of the learning algorithms. The results of this work encourage a human-first approach, that could inform the design of transparency mechanisms that anticipate how users conceptualize agent learning, such as interface cues aligned with inferred goals or decision strategies. This could be done by specifically designing transparency methods around user-inferred concepts—such as Goals, Knowledge, Decision Making, and Learning Mechanisms to better align system behavior with human expectations. A concrete example could be to use the insights about a teacher's expectation of goals to prevent teachers creating positive reward cycles in the teaching interaction [23] by specifically correcting this misalignment in terms that align with observer inferences about agent learning processes. This opens up new possibilities for human-centered interfaces, explainability mechanisms and agent-teacher communication methods (e.g. [9, 20]). 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 investigated how humans infer RL agents' learning processes through observation alone. To this end, we introduced a novel experimental paradigm and applied it across two tasks and RL algorithms. The results revealed a coherent framework of observer inferences structured around four common inference themes, with a detailed structure of interrelated subclusters, along the lines of which observer inferences about the agent learning process are organized: Agent Goals, Knowledge, Decision Making, and Learning Mechanisms. These findings offer a data-driven foundation for designing more transparent and interpretable learning agents. By aligning agent communication and behavior with these human inference patterns, future systems can foster better understanding, reduce misalignment, and enable more effective collaboration and synergy between human teachers and learning agents. Taken together, this work marks a step toward more human-centered RL systems and contributes to the broader vision of Hybrid Intelligence, combining human and machine intelligence to solve tasks neither can tackle alone.

ACKNOWLEDGMENTS

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

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] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

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

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

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

[9] 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.

[10] 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).

[11] 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.

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

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

[14] 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).

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

[16] 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).

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

[18] 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).

[19] 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.

[20] Bernhard Hilpert. 2024. Closing the Teacher-Learner Loop: The Role of Affective Signals in Interactive RL. In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 97–101.

[21] Bernhard Hilpert, P Gebhard, and T Schneeberger. 2021. Employing virtual agents for building trust in driving automation: A qualitative pilot study. In *Workshop on Robo-Identity: Artificial identity and multi embodiment at the 16th International Conference on Human-Robot Interaction (HRI'21)*.

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

[23] 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.

[24] 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.

[25] Muhan Hou, Koen Hindriks, AE Eiben, and Kim Baraka. 2023. Shaping Imbalance into Balance: Active Robot Guidance of Human Teachers for Better Learning from Demonstrations. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1737–1744.

[26] 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).

[27] Diana Immel, Bernhard Hilpert, Patricia Schwarz, Andreas Hein, Patrick Gebhard, Simon Barton, René Hurlemann, et al. 2025. Patients' and Health Care Professionals' Expectations of Virtual Therapeutic Agents in Outpatient Aftercare: Qualitative Survey Study. *JMIR Formative Research* 9, 1 (2025), e59527.

[28] 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.

[29] 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.

[30] 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.

[31] Ramira VAN DER MEULEN and Rineke VERBRUGGEb Max VAN DUIJN. 2024. Common Ground Provides a Mental Shortcut in Agent-Agent Interaction. *HHAI 2024: HYBRID HUMAN AI SYSTEMS FOR THE SOCIAL GOOD* (2024), 281.

[32] 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.

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

[34] Phillip Richter, Heiko Wersing, and Anna-Lisa Vollmer. 2024. Reducing Mental Model Mismatch with Intention-Based Feedback in Human-Robot Teaching. In *Proceedings of the 12th International Conference on Human-Agent Interaction*. 399–401.

[35] 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.

[36] Tanja Schneeberger, Anna Lea Reinwarth, Robin Wensky, Manuel Silvio Anglet, Patrick Gebhard, and Janet Wessler. 2023. Fast Friends: Generating Interpersonal Closeness between Humans and Socially Interactive Agents. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–8.

[37] 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.

[38] 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.

[39] 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.

[40] 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