# Modeling human learning of demonstration-based explanations for user-centric explainable AI

Suresh Kumaar Jayaraman
sureshkj@andrew.cmu.edu
Robotics Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Aaron Steinfeld
steinfeld@cmu.edu
Robotics Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Reid Simmons
rsimmons@andrew.cmu.edu
Robotics Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Henny Admoni
hadmoni@andrew.cmu.edu
Robotics Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

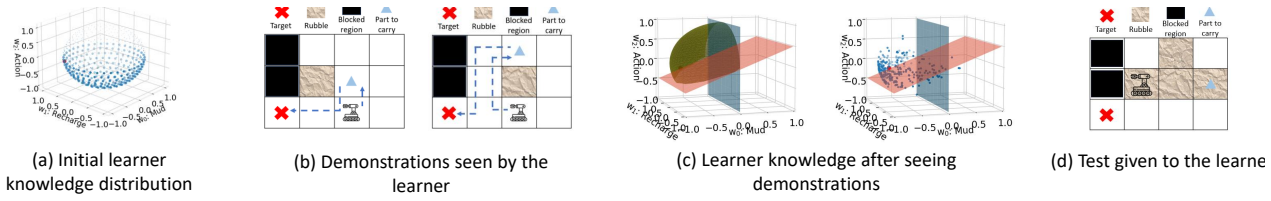


**Figure 1: A sample machine teaching sequence to explain a delivery robot's decision-making to a human learner. The robot's reward function has three features - mud, step cost, and battery. The learner starts with a prior knowledge that step cost is always negative and is shown demonstrations to learn the trade-offs between the features. Sometimes, the learner is provided with intermittent diagnostic tests to evaluate how well they have learned from the demonstrations. A simulated learner can sample a test response from the available distribution in machine teaching experiments for explainable decision-making.**

## ABSTRACT

Explainable reinforcement learning (XRL) aims to provide insights into the decision-making process of reinforcement learning (RL) agents, enabling humans to comprehend, trust, and collaborate with them. However, providing effective human-centric explanations requires collecting large amounts of human interaction data. In this paper, we propose a Bayesian inverse reinforcement learning model of "simulated learners" who infer the agent's reward function from demonstrations. We use a particle filter to represent and update the learner's beliefs based on the information conveyed by the demonstrations. We also introduce the concept of *understanding factor*, a parameter that captures the user's ability to learn from demonstrations and varies with feedback. We evaluate our learner model using a simulated delivery robot task and compare it with different teaching methods and learner types. We show that our model can simulate realistic human learning behavior and closely match the performance of actual human learners thus offering a novel and flexible way to design and evaluate user-centric XRL systems that can enhance user comprehension and trust in RL agents.

## CCS CONCEPTS

• **Human-centered computing** → **User models**.

## KEYWORDS

machine teaching, user-centric XAI, learner modeling, simulation

**ACM Reference Format:**
Suresh Kumaar Jayaraman, Aaron Steinfeld, Reid Simmons, and Henny Admoni. 2024. Modeling human learning of demonstration-based explanations for user-centric explainable AI. In *Proceedings of (HRI '24).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Reinforcement learning (RL) has shown promise in solving sequential decision-making tasks in various domains, but the opacity of RL models can hinder their practical implementation [19, 27]. Explainable reinforcement learning (XRL) seeks to provide insights into the decision-making process of RL agents, enabling humans to comprehend their actions, intervene when necessary, and ensure safety and reliability [8]. While explanations are a powerful way to enhance the transparency of AI decision-making processes, tailoring explanations to cater to the understanding, cognitive abilities, domain knowledge, and preferences of users is critical for their utility [11, 18].


*HRI '24, Mar 11–15, 2024, Boulder, CO*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

Explaining a machine's decisions to human users resembles a teaching-learning dynamic, with the XAI system playing the role of the teacher and humans as the students [15, 23]. Thus incorporating insights from cognitive science to grasp human thought processes and learning mechanisms is crucial for XAI developers to create systems that are not only informative but also user-friendly across diverse backgrounds and expertise levels. Constructing a comprehensive learner model involves delving into various cognitive aspects, such as analyzing performance, identifying misconceptions, and delineating goals and plans [23]. The learner model serves as the foundation for instructional decisions, facilitating understanding of user needs and enabling tailored adaptation [17].

While actual human learners undoubtedly offer valuable insights, there is a growing interest in using simulated learners in XAI as it provides several distinct advantages [4, 12, 24]. First, by leveraging simulated learners as test subjects, researchers can systematically assess the efficacy and user-friendliness of various explanation methods and interfaces within AI systems in an iterative way. Second, simulated learners enable the generation and testing of hypotheses regarding human learning processes−−a vital aspect of advancing XAI understanding—with minimal actual human data that are more difficult to collect. Through simulated experimentation, researchers can investigate how diverse factors such as prior knowledge, motivation, and feedback influence human learning outcomes and cognitive mechanisms.

In pedagogical literature, learners are frequently conceptualized as Bayesian Learners, leveraging probabilistic frameworks to represent their learning processes [6, 7]. This approach allows for the integration of prior knowledge with observed data to infer the learner's beliefs and update them accordingly. In explainable AI applications, [10] and [21] have extended this concept by modeling users as Bayesian Inverse Reinforcement Learners, aiming to elucidate how humans perceive and interpret the decisions made by AI systems. However, despite the theoretical advancements, there remains a notable gap in the validation of these learner models using empirical user data.

This work presents a simulated model of human learning specifically tailored for the context of machine teaching for explainable RL, adapted from [16]. Unlike [16], this model aims to simulate learner behavior by adaptively updating the particle filter update distribution based on the learning progress. Human learners are modeled as Bayesian inverse reinforcement learners utilizing a particle filter framework to approximate inference regarding robot reward weights. This approach not only offers a more accurate representation of human perception and learning. The model's versatility is demonstrated through its performance across various pedagogical explainability frameworks and its calibration with real user data enhances its validity, paving the way for more effective design and development of explainable AI systems.

## 2 BACKGROUND

**Markov Decision Process** We model the environment as a Markov Decision Process (MDP), given by the tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma, \mathcal{S}_\iota \rangle$, representing the state space, action space, transition function, reward function, discount factor, and initial state distribution respectively. An optimal trajectory $\xi^*$ is a sequence of $(s_i, a, s_i')$ tuples obtained by following the robot's optimal policy $\pi^*$. Similar to prior work [1], $R = \mathbf{w}^{*\top}\phi(s, a, s')$ is represented as a weighted linear combination of reward features. We define a group of MDPs that share $R, \mathcal{A}$, and $\gamma$ but differ in $T_i$, $\mathcal{S}_i$, and $S_i^0$, as a domain. Sharing the same $R$ ensures that all demonstrations within the domain support inference over a common $\mathbf{w}^*$. We use the MDP formulation to model an *item delivery* task where a robot is tasked with delivering an item through an environment that has rubble, blocked regions, and a battery recharge station (see Fig.1 (b)).

**Machine teaching for policies:** We adapt the machine teaching framework for policies [14] to select a set of demonstrations $\mathcal{D}$ of size $n$ that maximizes the similarity $\rho$ between optimal policy $\pi^*$ and the policy $\hat{\pi}$ recovered using a computational model $\mathcal{M}$ (e.g., IRL) on $\mathcal{D}$, $\arg\max_{\mathcal{D}\subset\Xi} \rho(\hat{\pi}(\mathcal{D}, \mathcal{M}), \pi^*)$ s.t. $|\mathcal{D}| = n$, where $\Xi$ is the set of all demonstrations of $\pi^*$ in a domain. Once $\mathbf{w}^*$ is approximated through IRL, this approach assumes that the learner can deduce $\pi^*$ by planning on the underlying MDP. Thus, the objective reduces to selecting demonstrations that are informative in conveying $\mathbf{w}^*$, which can be measured using behavior equivalence classes.

**Behavior equivalence class:** *The behavioral equivalence class (BEC)* of the optimal policy $\pi^*$ is the set of reward functions under which $\pi^*$ is optimal. For a reward function that is a weighted linear combination of features, the BEC of a demonstration (trajectory) $\xi^*$ of $\pi^*$ is the intersection of half-spaces [3] formed by the exact IRL equation [20]

$$\mathrm{BEC}(\xi^*|\pi^*) := \mathbf{w}^\top \left(\mu_{\pi^*}^{(s,a)} - \mu_{\pi^*}^{(s,b)}\right) \geq 0, \forall (s,a) \in \xi^*, b \in \mathcal{A}. \quad (1)$$

where $\mu_{\pi^*}^{(s,a)} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi\,(s_t) \mid \pi, s_0 = s, a_0 = a\right]$ is the vector of reward feature counts accrued from taking action $a$ in $s$, then following $\pi^*$ after. Any demonstration can be converted into a set of constraints on $\mathbf{w}$ using (1), with each constraint being a *knowledge component/concept (KC)* [13] that captures a facet of the reward function (e.g., tradeoffs between the underlying reward features). Consider the item delivery domain, which has binary reward features $\phi$ = [*traversed rubble*, *battery recharged*, *action taken*]. In practice, we require $||\mathbf{w}^*||_2 = 1$ to bypass both the scale invariance of IRL and the degenerate all-zero reward function. If no prior knowledge is assumed, the potential belief space on reward weights would uniformly span the surface of the $n-1$ sphere ($n$ is the number of domain features) due to the $L^2$ norm constraint on $\mathbf{w}^*$. We instead assume that learners begins with a prior that action weight is negative (e.g. favoring shortest path, see Fig. 1 (a)).

## 3 PARTICLE FILTER LEARNER MODEL

Drawing inspiration from [24], we conceptualize the learner (user) as a representation derived from the teacher's (robot's) model of the learner proposed in [16].

### 3.1 Robot task and teaching framework

Consider that the robot's task is to deliver items to an emergency response team. The robot has limited maneuverability over rubble, and limited range, and may prefer to recharge when possible. These capabilities and preferences (i.e. its decision-making) must be taught to the team quickly because of the time-sensitive situation.

We briefly discuss the teaching framework this model is evaluated for. For more details, readers are encouraged to refer to [16].

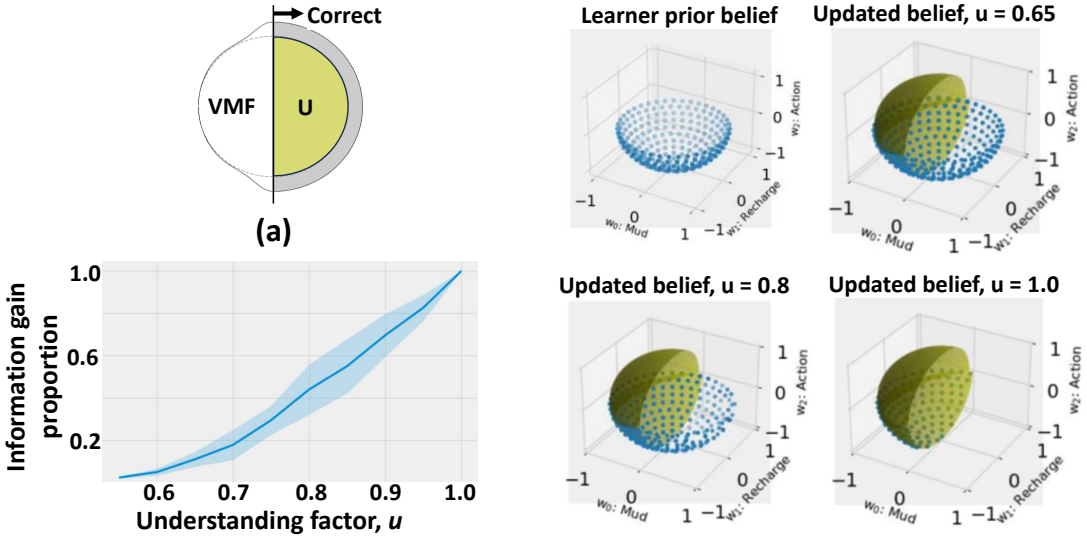


**Figure 2: (a) The custom probability density function (pdf) for updating particle weights based on a constraint generated. The probability mass on the uniform side is operationalized as the *understanding factor* of simulated learners. (b) Variation in learning dynamics with understanding factor, $u$. The information gain ratio of the learner monotonically increases with $u$. (c) All learners started with the same prior, i.e. step cost is negative. The particle filters are updated based on the custom pdfs for each $u$. The distribution gets more and more concentrated with increasing $u$.**

Using the pedagogical principle of scaffolding, the algorithm selects individual knowledge components/concepts (KCs) that incrementally increase in information across an increasing subset of features. For example, the KCs could incrementally teach the bounds on the cost of traveling through rubble given the step cost, followed by bounds on the reward for recharging given the step cost, and then trade-offs between these three. The demonstrations are selected based on the KCs.

### 3.2 Operationalization of the learner model

We model learner belief about the robot's reward weights using a particle filter, adapting the modeling approach from [16] used for the teacher's model of the learner. Each particle represents a potential belief about the robot's reward function and the particle weights are updated in a Bayesian manner based on constraints conveyed through demonstrations. The constraints correspond to the knowledge gained from demonstrations. The particle filter update for a demonstration is shown in Fig. 1.

We use a custom probability distribution, $p(x_t|y_t)$, (refer Fig. 2(a)) to update particle weights after seeing a demonstration. This distribution is a combination of a uniform distribution for the correct half-space of the constraint (indicating that any particle lying in this space is equally valid for the demonstration) and a von Mises-Fisher distribution for the incorrect half-space (indicating that particles farther away from the constraint are exponentially less likely to have generated the demonstration).

Learning from examples depends on the analogical reasoning ability of individuals [5, 25]. Analogical reasoning is a cognitive process where individuals use analogies, or comparisons between different objects, concepts, or situations, to understand or solve problems. It involves recognizing similarities between two or more things that may be superficially different but share underlying commonalities. For machine teaching of robot policy, this involves observing patterns of robot behaviors in some situations and generalizing them to similar situations [3, 15].

We parametrize this ability to understand from demonstrations as the understanding factor, $u$. It is a measure of how much users can understand the underlying constraints from the demonstrations seen and is defined as the probability mass on the uniform (correct) side of the custom distribution of the underlying constraint (see Fig. 2(a)). Its effect is that of a modifying factor on the information gain in the demonstration that is translated to the information gain of the particle filter, $IG_{pf} = f(u, IG_{demo})$.

A higher understanding factor implies that learners assign more weights to particles on the correct side of the constraints. Fig. 2 shows the change in distribution after seeing the same demonstration for various understanding factors and the associated information gain ratio. We calculate information gain ratio as the change in entropy of the particles before and after seeing the demonstration to the entropy change when the understanding factor is '1'.

## 4 LEARNER MODEL PERFORMANCE

### 4.1 User Study

As proposed in [16], the authors conducted an online user study to evaluate the effects of different teaching frameworks - open loop and closed loop. Open loop consisted of only a series of demonstrations for various KCs of increasing difficulty and a set of final tests to evaluate their understanding of the robot's reward. Closed loop additionally provided diagnostic tests for each KC to test the learner's knowledge of a KC intermittently. The algorithm provides feedback on whether the learners got the diagnostic tests right or wrong and also provides remedial demonstrations and tests when

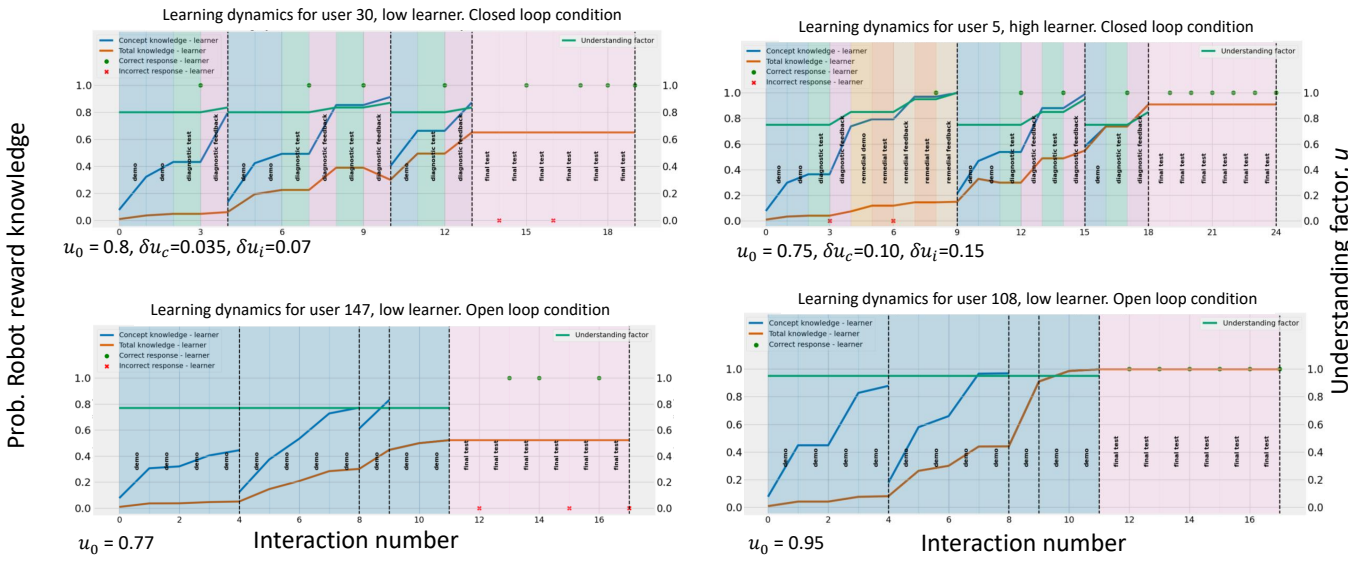


**Figure 3: Learner model simulated for specific user trials. The parameters of initial understanding factor $u_0$ and change in understanding factor, $\delta u_i$, and $\delta u_c$ were tuned to get the prob. robot reward knowledge close to the final test performance that six tests in total. By tuning these understanding factor parameters, we can get diverse learning behaviors of low and high learners and also able to handle various teaching paradigms, i.e. open loop without feedback and closed loop with feedback.**

the learners got the intermittent diagnostic tests wrong until the learner demonstrates concept mastery. These teaching frameworks are inspired by the pedagogical principles of providing teacher feedback [2, 26] and testing effect [22].

## 4.2 Model performance comparison to real learner behavior

We evaluate our learner model in two dimensions — type of learner, and teaching framework. Learners have different cognitive capabilities and understand at different levels the same information provided. Furthermore, the teaching process is likely to be an adaptive and varied process, catered to the specific learner. Thus our model of the learner should be able to encompass a wide variety of learner abilities in different dynamic teaching contexts.

We evaluate how our learner model can be utilized to model various observed human learning behaviors in the user study data for different teaching frameworks and different learning abilities. We operationalize the understanding factor, $u = u_0$, as the individual's ability to learn from demonstrations. Pedagogical literature suggests that people get better at learning a concept when they get feedback and are repeatedly exposed to the concept [9, 26]. So, to incorporate the effect of feedback, we define the understanding factor for teaching frameworks that receive feedback as,

$$u_t = u_{t-1} + \Delta u_{t-1}, \text{ where}$$

$$\Delta u_{t-1} = \begin{cases} \delta u_c & \text{if test is correct at time } t-1 \\ \delta u_i & \text{if test is incorrect at time } t-1 \end{cases} \quad (2)$$

$u_t$ resets to $u_0$, which is the base ability of the learner to learn from demonstrations for each new concept as we assume the concepts to be independent. Thus the improvement in $u$ due to feedback is contained within the specific concept / KC.

Fig. 3 shows simulations of learner behavior for various categories of learners and different teaching frameworks. We evaluate the measure, probability of robot reward knowledge as the sum of weights of particles within the correct BEC area of the robot reward. The parameter values of $u$, $\delta u_c$, & $\delta u_i$ were manually tuned to match the observed performance. By carefully choosing the parameters, the learner models can simulate close to the observed performance of actual human learners while seeing the same demonstrations and feedback. This demonstrates the model's applicability not only to model different types of learners but also for various teaching frameworks with different learning dynamics. Further, as seen in the top right of Fig. 3, the learner is also to capture complicated learning behaviors that had trouble learning the first concept but performed very well after that.

## 5 CONCLUSION

In this work, we explore a Bayesian Inverse Reinforcement Learner model that can simulate realistic human learning behavior for machine teaching experiments. We were able to show that the simulated learning behavior based on the proposed model closely matches the observed final performance of actual human learners for several types of learners and teaching frameworks and is even able to model complex learning dynamics. With the availability of actual data, the model can be used to identify the distribution of the understanding factor parameters that best capture the learning dynamics of each type of learner under different teaching contexts. These in turn can be used to sample the simulated learners for various explainability experiments and using actual human learners for fine-tuning, drastically reducing the requirements for actual human data.

## ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research award N00014-181-2503.

I want to thank Michael Lee from Carnegie Mellon University for the valuable discussions and for providing the user study data.

## REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *ICML*.

[2] David Boud and Elizabeth Molloy. 2013. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in higher education* 38, 6 (2013), 698–712.

[3] Daniel S Brown and Scott Niekum. 2019. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *AAAI*.

[4] Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. 2022. Use-case-grounded simulations for explanation evaluation. *Advances in Neural Information Processing Systems* 35 (2022), 1764–1775.

[5] Andrea Cheshire, Linden J Ball, and CN Lewis. 2005. Self-explanation, feedback and the development of analogical reasoning skills: Microgenetic evidence for a metacognitive processing account. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society, ed. BG Bara, L. Barsalou & M. Bucciarelli.* 435–41.

[6] Konstantina Chrysafiadi and Maria Virvou. 2013. Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications* 40, 11 (2013), 4715–4729.

[7] Cristina Conati, Abigail Gertner, and Kurt Vanlehn. 2002. Using Bayesian networks to manage uncertainty in student modeling. *User modeling and user-adapted interaction* 12 (2002), 371–417.

[8] Mica R Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human factors* 59, 1 (2017), 5–27.

[9] Lisa K Fazio, Barbie J Huelser, Aaron Johnson, and Elizabeth J Marsh. 2010. Receiving right/wrong feedback: Consequences for learning. *Memory* 18, 3 (2010), 335–350.

[10] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. 2019. Enabling robots to communicate their objectives. *Autonomous Robots* (2019).

[11] Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, and Adish Singla. 2019. Interactive teaching algorithms for inverse reinforcement learning. *arXiv preprint arXiv:1905.11867* (2019).

[12] Tanja Käser and Giora Alexandron. 2023. Simulated learners in educational technology: A systematic literature review and a turing-like test. *International Journal of Artificial Intelligence in Education* (2023), 1–41.

[13] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[14] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Exploring Computational User Models for Agent Policy Summarization. In *International Joint Conference on Artificial Intelligence*.

[15] Michael S Lee, Henny Admoni, and Reid Simmons. 2022. Reasoning about Counterfactuals to Improve Human Inverse Reinforcement Learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9140–9147.

[16] Michael S Lee, Henny Admoni, and Reid Simmons. 2023. Closed-loop Reasoning about Counterfactuals to Improve Policy Transparency. In *International Conference o Machine Learning (ICML) Workshop on Counterfactuals in Minds and Machines*.

[17] Nan Li, William W Cohen, Kenneth R Koedinger, and Noboru Matsuda. 2011. A machine learning approach for automatic student model discovery.. In *Edm*. ERIC, 31–40.

[18] Francisco S Melo and Manuel Lopes. 2021. Teaching Multiple Inverse Reinforcement Learners. *Frontiers in Artificial Intelligence* 4 (2021), 625183.

[19] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2022. A survey of explainable reinforcement learning. *arXiv preprint arXiv:2202.08434* (2022).

[20] Andrew Y Ng and Stuart Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *International Conf. on Machine Learning*.

[21] Peizhu Qian and Vaibhav Unhelkar. 2022. Evaluating the Role of Interactivity on Improving Transparency in Autonomous Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1083–1091.

[22] Henry L Roediger III and Jeffrey D Karpicke. 2006. The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science* 1, 3 (2006), 181–210.

[23] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[24] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. 2021. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artificial Intelligence* 301 (2021), 103570.

[25] Claire E Stevenson, Wilma CM Resing, and Mandy N Froma. 2009. Analogical reasoning skill acquisition with self-explanation in 7-8 year olds: Does feedback help? *Educational and Child Psychology* 26, 3 (2009), 6.

[26] Marieke Thurlings, Marjan Vermeulen, Theo Bastiaens, and Sjef Stijnen. 2013. Understanding feedback: A learning theory perspective. *Educational Research Review* 9 (2013), 1–15.

[27] Lindsay Wells and Tomasz Bednarz. 2021. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in artificial intelligence* 4 (2021), 550030.