Clarifying Feature Overspecification in Reward Learning from State Corrections via Follow Up Questions

Ethan Villalovoz[1], Michelle Zhao[2], Henny Admoni[2], Reid Simmons[2]

***Abstract*— As autonomous systems, from robotic manipulators in homes to personal AI assistants on computers, become increasingly prevalent, aligning these robots with human stakeholders' preferences and values is a critical challenge. Current methodologies involve robots learning to emulate human intentions through interactive feedback, particularly from physical corrections provided by users. However, in dynamic real-world settings, users may prefer to modify the environment's state directly rather than teleoperating the robot. Our work investigates how robots can effectively learn from these online state corrections to enhance performance and align with human preferences across various domains. We leverage probabilistic models of human preference to enable Bayesian inference based on iterative state corrections. Recognizing that humans seek clarification when uncertain, our approach allows robots to prompt for guidance upon encountering significant uncertainty. We will evaluate our system's adaptive learning and proactive dialogue capabilities compared to a system without dialogue, measuring task completion efficiency and the reduction of error states. This research aims to develop robust, user-friendly autonomous systems that generalize learned behaviors and continuously improve through human interaction.**

I. INTRODUCTION

As autonomous systems become increasingly integrated into everyday life as in Figure 1, ensuring these systems align with human stakeholders' preferences and values is paramount. From robotic manipulators in homes to personal AI assistants on computers, the ability of these robots to learn and adapt to human intentions significantly impacts their utility and acceptance. Traditional methods for aligning robots with human preferences often involve direct teleoperation or kinesthetic teaching, which can be cumbersome and impractical in dynamic, real-world environments. This challenge necessitates novel approaches that enable robots to learn from more natural forms of human feedback, such as state corrections made directly by users.

Recent advancements in robotics and artificial intelligence have highlighted the importance of interactive feedback mechanisms, particularly physical human-robot interactions (pHRI), where users physically correct the robot's actions during task execution depicted in Figure 2 [1]. These interactions provide valuable insights into human preferences and intentions, allowing robots to refine their behaviors accordingly. However, current methodologies primarily treat

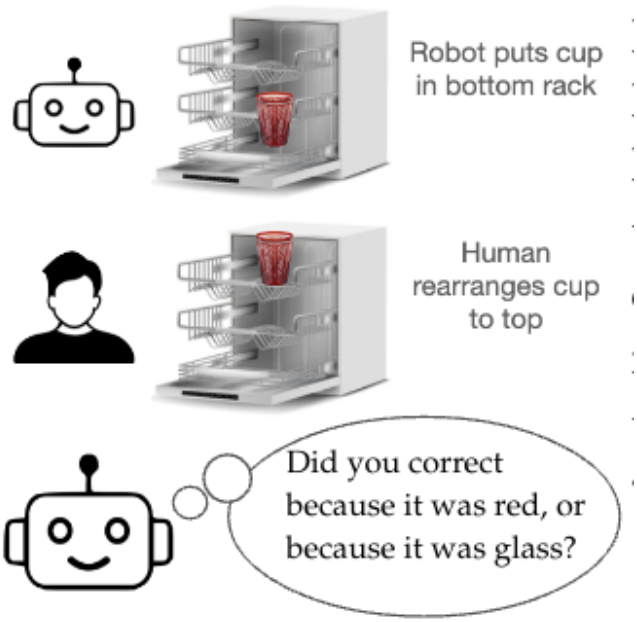

Fig. 1. In this scenario, the robot autonomously places a red cup in the bottom rack of a dishwasher. The human operator, preferring a different configuration, corrects the robot's action by moving the cup to the top rack. The robot then initiates a clarification dialogue, asking whether the correction was based on the cup's color (red) or material (glass). This figure illustrates the robot's proactive learning approach, which aims to understand and adapt to the human's underlying preferences by seeking specific feedback..

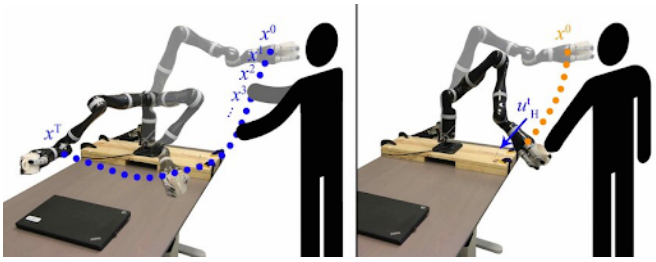

Fig. 2. Example of physical human-robot interaction (pHRI). In previous work, users physically corrected the robot's actions during task execution, enhancing collaboration and task performance. This figure illustrates the corrective actions performed by a human operator to guide the robot's movements in real-time [1].

[1]School of Electrical Engineering & Computer Science, Washington State University, Pullman, Washington, USA, ethan.villalovoz@wsu.edu

[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, {mzhao2, rsimmons, hadmoni}@andrew.cmu.edu

these interactions as disturbances to be rejected rather than informative corrections to be learned from.

Our research addresses this gap by investigating how robots can effectively learn from online state corrections made by users, enhancing performance and aligning with human preferences across various domains. By leveraging probabilistic models of human preference, we enable Bayesian inference based on iterative state corrections, allowing robots to adaptively improve their behavior over time. Additionally, recognizing that humans seek clarification when uncertain, our approach incorporates proactive dialogue, enabling robots to prompt for guidance upon encountering significant uncertainty. This dual strategy of adaptive learning and proactive dialogue aims to develop robust, user-friendly autonomous systems capable of continuous improvement through human interaction.

II. RELATED WORK

The field of interactive robot learning has seen significant developments, with various approaches focusing on different types of feedback mechanisms. Imitation learning (IL) and reinforcement learning (RL) have been extensively studied, with IL demonstrating greater sample efficiency and practical utility in robotics due to its reliance on expert demonstrations. However, providing optimal state-action demonstrations can be challenging, leading researchers to explore alternative forms of feedback.

One such approach is learning from corrective feedback, where users provide adjustments during robot execution instead of full demonstrations. This method has proven effective in various studies, highlighting its potential for simplifying the human teaching process and improving robot learning efficiency. For example, Cakmak and Thomaz discussed the importance of robots asking good questions to clarify user intentions [2], while Jain et al. explored learning preferences for manipulation tasks from online coactive feedback [3]. Schmittle et al. proposed a meta-algorithm for learning from diverse corrective feedback types, emphasizing the need for flexible learning mechanisms that can handle different forms of noisy feedback from users [4].

Physical human-robot interaction (pHRI) represents another crucial area of research. Traditional pHRI approaches often treat human interaction forces as disturbances, aiming to either reject or comply with these forces without altering the robot's trajectory. However, recent studies suggest that these interactions can provide valuable information about human preferences and desired behaviors. Losey et al. formalized pHRI as a dynamical system where human corrections are seen as observations about the robot's objective function, allowing robots to learn and adapt in real-time [5]. Similarly, Hadfield-Menell et al. focused on cooperative inverse reinforcement learning, where the robot infers human preferences through interaction [6].

Incorporating human feedback into robot learning is further enriched by the concept of active learning, where robots actively seek guidance from users when uncertain about the correct action. This approach not only improves learning efficiency but also enhances user satisfaction by involving them in the training process. For instance, Daniel et al. highlighted the benefits of probabilistic inference for learning from human feedback [7].

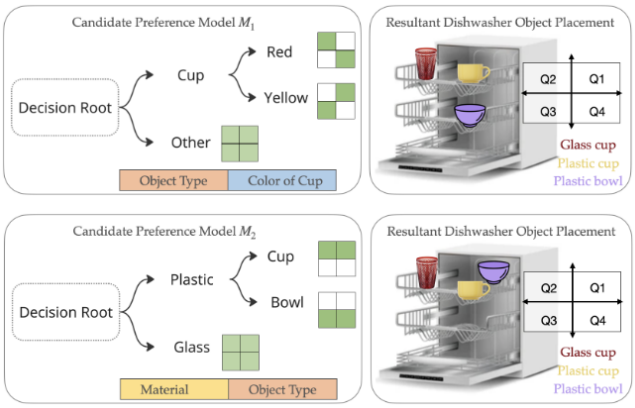

Fig. 3. Representation of candidate preference models (M1 and M2) that the robot uses to hypothesize about human preferences. The trees depict potential reward functions based on features like object type, color, and material. For instance, Model M1 assumes preferences based on the color of the cup, while Model M2 considers material and object type. The resultant object placement is shown in the dishwasher, guiding the robot's actions.

Our research builds on these foundational works by integrating probabilistic models of human preference with Bayesian inference to learn from state corrections. Additionally, we incorporate proactive dialogue to prompt for user guidance during moments of high uncertainty, thereby combining the strengths of passive learning from corrections with active learning through dialogue. This integrated approach aims to develop autonomous systems that are not only adaptive and efficient but also intuitive and responsive to user needs (Figure 4).

III. APPROACH: OVERSPECIFICATION FROM STATE CORRECTIONS

Our approach represents the reward function as a reward hypothesis space Θ, which consists of multiple hierarchical reward functions (Figure 3). Each function is structured as a decision tree, where each tree encodes different hypotheses about human preferences based on features such as object color, material, and type. The trees are flexible, allowing customization at each level to represent specific preferences by incorporating some or all of these features.

We formulate our approach as a two-agent Markov Decision Process (MDP). Each iteration of the MDP involves the following steps: observing the interaction between the human and the robot, determining the appropriate action for the robot, performing a Bayesian update based on the human's corrections. This iterative process allows the robot to learn and refine its reward model through continuous interaction with the human agent. Mathematical notation is adopted from [6].

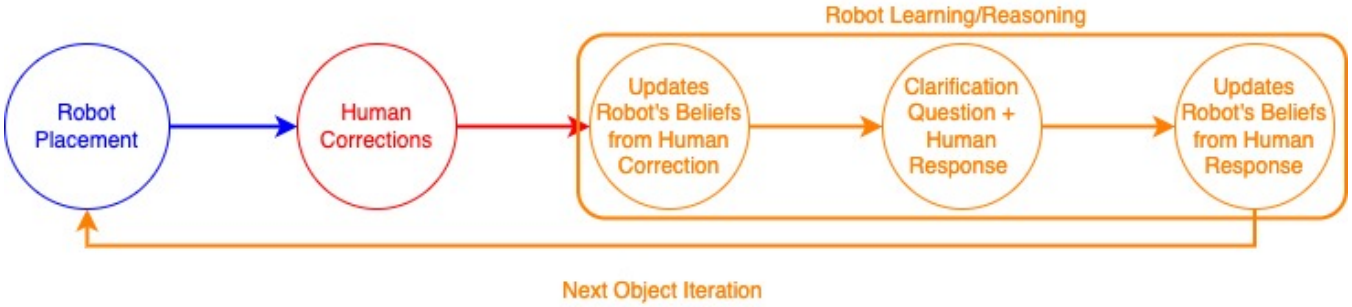

Fig. 4. The interactive workflow between the robot and human, detailing the iterative process through each object in the scene. The robot initially places an object, followed by human corrections. The robot updates its beliefs based on these corrections and may ask clarification questions. The human's response further refines the robot's understanding, which is iteratively improved through this collaborative learning and reasoning process.

A. Interaction MDP

To formalize the robot learning from human corrections, we model the interaction as a Markov Decision Process (MDP). The MDP consists of:

- **States** ($\mathcal{S}$): The set of all possible states representing the environment, including the position and attributes of objects in the dishwasher.
- **Robot Actions** ($\mathcal{A}^{\mathbf{R}}$): The set of possible actions the Robot can take to move the object.
- **Human Actions** ($\mathcal{A}^{\mathbf{H}}$): The set of possible actions the Human can take to correct the robot's placements. These are the same actions within $\mathcal{A}^{\mathbf{R}}$.
- **Reward Hypothesis Space** (Θ): A discrete set of trees representing different reward functions that model human preferences.
- **Transition Function** ($T(\cdot|\cdot,\cdot,\cdot)$): The probabilistic model of transitioning from one state to another given the human's correction.
- **Initial State Distribution** ($P_0(\cdot,\cdot)$): The probability distribution over initial states.

B. Bayesian Update Given Corrected State

When the human corrects the robot's action, the robot updates its belief about the true reward function using Bayesian inference [8], [9]. The update process involves:

1) **Observation**: The robot observes the corrected state (S_2) and compares it to the initial state (S_0) and its own placement (S_1).
2) **Case Analysis**:
   - $S_2 \neq S_1$: The robot's placement differs from the human's correction.
   - $S_2 = S_1$ and $S_0 \neq S_1$: The robot's placement matches the human's correction, but not the initial state.
   - $S_2 \neq S_0$: The human's correction differs from the initial state.
3) **Update Rule**: The robot updates its belief about the reward function (θ_i) using the following Bayesian update rule [9]:

$$P(\theta_i|S_h > S_r) = \frac{P(\theta_i) \cdot P(S_h > S_r|\theta_i)}{\sum_{\theta_j \in \Theta} P(\theta_j) \cdot P(S_h > S_r|\theta_j)}$$

where S_h is the human-corrected state and S_r is the robot's placement. The Bradley-Terry model will be used to represent $P(S_h > S_r|\theta_i)$ as [10]:

$$P(S_h > S_r|\theta_i) = \frac{e^{\beta S_h}}{e^{\beta S_r} + e^{\beta S_h}}$$

IV. Clarification Questions

In addition to learning from state corrections, our approach incorporates proactive dialogue through clarification questions. When the robot encounters significant uncertainty about the correct action, it prompts the human for guidance.

A. Question Types

We categorize clarification questions into three types:

- **Environment States**: Questions about the current state of the environment. Example: "Is the cup placed correctly?"
- **Features**: Questions about the features relevant to the correction. Example: "Is the reason for placing the cup here because of its color, type, or material?"
- **Hypothesis**: Questions about the inferred preferences. Example: "Do you prefer objects to be placed closer to the center?"

B. Implementation

To enhance the robot's understanding of human preferences, we implemented a mechanism that generates clarification questions based on the current state, features of the objects, and the robot's hypothesis space. This approach allows the robot to iteratively refine its reward model by querying the human about which features were most relevant in their corrections.

The clarification questions are designed to focus on specific attributes of the object in question, such as color, material, or type. For example, after the robot has placed a 'yellow glass cup' and received a correction from the human, it might ask:

> *"For the recent 'yellow glass cup', which features (color, type, material) were relevant to its placement?"*

This process is implemented as follows:

1) **Identification of Relevant Features**: For each hypothesis in the reward hypothesis space Θ, the robot identifies the features that are relevant to the decision-making process for the current object. This involves analyzing the decision trees associated with each hypothesis to determine which attributes (color, material, type) are critical for that particular hypothesis.
2) **Human Query**: The robot then asks the human to specify which features were relevant to the object's placement. The human provides a response indicating the relevant features, such as color, type, or material.
3) **Belief Update**: Based on the human's response, the robot updates its beliefs about the hypotheses in Θ. This belief update is performed using a Bayesian approach. Specifically, the posterior probability of each hypothesis $\theta_i \in \Theta$ given the human's response is calculated as:

$$P(\theta_i \mid \text{response}) = \frac{P(\text{response} \mid \theta_i) \cdot P(\theta_i)}{\sum_j P(\text{response} \mid \theta_j) \cdot P(\theta_j)} \quad (1)$$

Here, $P(\theta_i)$ is the prior belief in hypothesis θ_i, and $P(\text{response} \mid \theta_i)$ is the likelihood of observing the human's response if θ_i were the correct hypothesis.
4) **Likelihood Assignment**: The likelihood $P(\text{response} \mid \theta_i)$ is determined based on whether the features identified in hypothesis θ_i match the human's response. If they match, the likelihood is set to a high value, $\alpha = 0.8$, representing a high probability that the hypothesis is correct given the response. If they do not match, the likelihood is set to $1 - \alpha = 0.2$.
5) **Normalization**: The updated beliefs are normalized so that they sum to 1, ensuring a valid probability distribution over the hypotheses.

This method ensures that the robot's model of human preferences is continuously refined through a process of interactive clarification and learning, leading to more accurate and customized reward functions.

C. Simulation Demo

To demonstrate our approach, we developed a simulation that showcases the iterative learning process of our robot. This simulation involves a robotic agent tasked with organizing items in a simulated dishwasher environment in Figure 9, interacting with a human user to refine its understanding of human preferences through state corrections and clarification questions.

In our simulation, the robot begins by selecting a random reward function from the predefined hypothesis space, which serves as the ground truth for human preferences. The simulation proceeds with the following iterative workflow:

1) **Initial Placement**: The robot places an object (e.g., a yellow glass cup) in one of the quadrants of the dishwasher (Figure 5).
2) **Human Correction**: The human user observes the placement and provides corrections if the object is not placed according to their preference. These corrections are made by moving the object to the desired location within the dishwasher (Figure 5).
3) **Bayesian Update**: Upon receiving a correction, the robot updates its belief about the true reward function using the Bayesian update rule. This update incorporates the human's correction into the robot's understanding, refining its model of human preferences (Figure 6).
4) **Clarification Questions**: If the robot encounters significant uncertainty about the human's preference, it generates clarification questions. These questions are designed to probe the relevant features (e.g., color, type, material) or the reasoning behind the correction (Figure 7). For example, the robot might ask, "Is the reason for placing the cup here because of its color, type, or material?"
5) **Human Response**: The human user answers the clarification questions, providing additional information that helps the robot further refine its reward model (Figure 8).
6) **Iteration**: The process repeats for each subsequent object, with the robot continuously improving its understanding of human preferences through this interactive learning and reasoning process.

The simulation aims to demonstrate the effectiveness of our dual strategy, combining adaptive learning from state corrections and proactive dialogue through clarification questions. This iterative process is designed to enhance the robot's performance in aligning with human preferences, ultimately

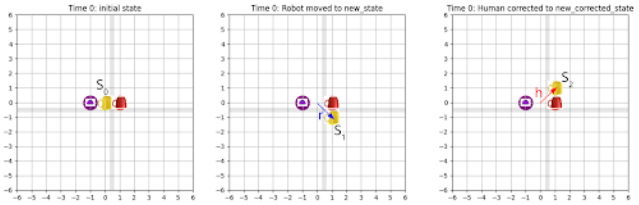

Fig. 5. The sequence of images illustrates the process: The left panel shows the initial state where the objects are unplaced S_0. The center panel shows the robot placing the first object, a yellow glass cup labeled S_1. The right panel shows the human correcting the position of the placed object to a new, corrected stateS_2.

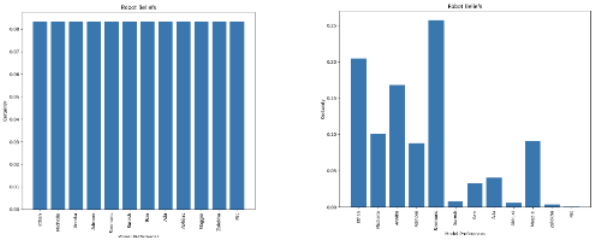

Fig. 6. The bar graphs depict the robot's confidence in various models of the true reward function (representing human preferences) before and after human state corrections. The left graph shows the initial state with uniform confidence across models. The right graph shows the updated state, where the robot's beliefs have become more refined. Despite the improvements, the model representing the true reward (labeled "Ethan") is not the most certain model according to the robot's updated beliefs.

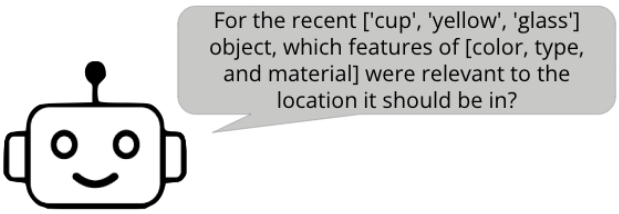

Fig. 7. The follow-up question posed to validate the reasoning behind the corrections made by the human. The question requests the identification of relevant features that justify the corrections, ensuring that the human's reasoning aligns with the intended outcomes.

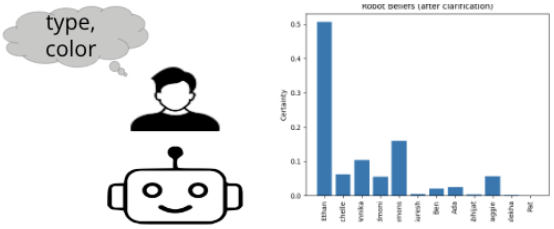

Fig. 8. The robot's understanding of the true reward function, representing human preferences, improves markedly after posing a clarification question. This process leads to the identification of the true reward model that accurately reflects human preferences.

leading to more efficient and user-friendly autonomous systems.

V. Experiments and Results

To evaluate our approach, we conducted a series of experiments in a simulated dishwasher loading scenario. Our experimental setup includes:

- **Environment**: A simulated dishwasher with a grid representation (Figure 9).
- **Objects**: Various kitchen items with different attributes (color, type, material).
- **Participants**: Human users providing state corrections and answering clarification questions.

A. Metrics

We used the following metrics to assess the performance of our system:

- **Task Completion Time**: The time taken to correctly place all items in the dishwasher.
- **Number of Corrections**: The number of corrections made by the user.
- **Clarification Questions**: The frequency and types of clarification questions asked.
- **User Feedback**: Qualitative feedback from participants regarding the system's usability and effectiveness.

B. Results

As the time of this writing we do not have any results to display for our work. This will be done at a future date to illustrate the effectiveness of our work in a real user participants.

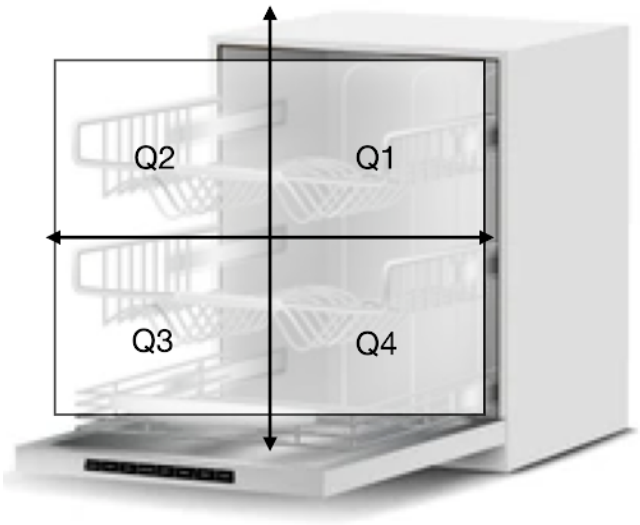

Fig. 9. The interior of a dishwasher divided into a four-quadrant grid. The x-axis delineates the border between the top and bottom rack sections, resulting in four quadrants (Q1, Q2, Q3, Q4). This segmentation is utilized to analyze and optimize the placement and organization of items within the dishwasher.

VI. Conclusion and Future Work

In this paper, we presented a novel approach for enhancing robot learning from human state corrections through Bayesian inference and proactive dialogue. Our system leverages clarification questions to address uncertainties and align robot actions with human preferences effectively.

Our work demonstrate the potential of this approach in improving task performance and user satisfaction. Moving forward, we plan to:

- Refine the clarification question generation process using more advanced language models.
- Conduct a comprehensive user study to validate our findings in real-world scenarios.
- Explore the application of our approach in other domains requiring close human-robot collaboration.

By continuing to develop and validate these methods, we aim to create autonomous systems that are not only efficient and adaptive but also deeply aligned with human values and preferences.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to Michelle Zhao, Dr. Henny Admoni, and Dr. Reid Simmons for allowing me to be a part of the Robotics Institute at Carnegie Mellon University, specifically within the HARP (Human And Robot Partners) Laboratory and the RASL (Reliable Autonomous Systems Laboratory). Their guidance and support have been invaluable to my research and professional development.

A special thanks to Rachel Burcin, RISS Co-director, and Dr. John M. Dolan, RISS Director, for the opportunity to participate in the Robotics Institute Summer Scholars (RISS) Program. This experience has been instrumental in advancing my knowledge and passion for robotics.

REFERENCES

[1] A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan, "Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 835–854, 2020.

[2] M. Cakmak and A. L. Thomaz, "Designing robot learners that ask good questions," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 17–24.

[3] A. Jain, S. Sharma, T. Joachims, and A. Saxena, "Learning preferences for manipulation tasks from online coactive feedback," *The International Journal of Robotics Research*, vol. 34, no. 10, pp. 1296–1313, 2015.

[4] M. Schmittle, S. Choudhury, and S. S. Srinivasa, "Learning online from corrective feedback: A meta-algorithm for robotics," *arXiv preprint arXiv:2104.01021*, 2021.

[5] D. P. Losey, A. Bajcsy, M. K. O'Malley, and A. D. Dragan, "Physical interaction as communication: Learning robot objectives online from human corrections," *The International Journal of Robotics Research*, vol. 41, no. 1, pp. 20–44, 2022.

[6] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," *Advances in neural information processing systems*, vol. 29, 2016.

[7] C. Daniel, H. Van Hoof, J. Peters, and G. Neumann, "Probabilistic inference for determining options in reinforcement learning," *Machine Learning*, vol. 104, pp. 337–357, 2016.

[8] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning." in *IJCAI*, vol. 7, 2007, pp. 2586–2591.

[9] A. Bobu, A. Bajcsy, J. F. Fisac, and A. D. Dragan, "Learning under misspecified objective spaces," in *Conference on Robot Learning*. PMLR, 2018, pp. 796–805.

[10] S. University, "STATS 200: Introduction to Statistical Inference Lecture 24 — The Bradley-Terry model," https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/Lecture24.pdf, 2016, [Online; accessed 06-August-2024].