Modeling Variation in Human Feedback with User Inputs: An Exploratory Methodology

Jindan Huang
Tufts University
Medford, Massachusetts, USA
jindan.huang@tufts.edu

Reuben M. Aronson
Tufts University
Medford, Massachusetts, USA
reuben.aronson@tufts.edu

Elaine Schaertl Short
Tufts University
Medford, Massachusetts, USA
elaine.short@tufts.edu

ABSTRACT

To expedite the development process of interactive reinforcement learning (IntRL) algorithms, prior work often uses perfect oracles as simulated human teachers to furnish feedback signals. These oracles typically derive from ground-truth knowledge or optimal policies, providing dense and error-free feedback to a robot learner without delay. However, this machine-like feedback behavior fails to accurately represent the diverse patterns observed in human feedback, which may lead to unstable or unexpected algorithm performance in real-world human-robot interaction. To alleviate this limitation of oracles in oversimplifying user behavior, we propose a method for modeling variation in human feedback that can be applied to a standard oracle. We present a model with 5 dimensions of feedback variation identified in prior work. This model enables the modification of feedback outputs from perfect oracles to introduce more human-like features. We demonstrate how each model attribute can impact on the learning performance of an IntRL algorithm through a simulation experiment. We also conduct a proof-of-concept study to illustrate how our model can be populated from people in two ways. The modeling results intuitively present the feedback variation among participants and help to explain the mismatch between oracles and human teachers. Overall, our method is a promising step towards refining simulated oracles by incorporating insights from real users.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → *Modeling and simulation*.

KEYWORDS

human behavior modeling, reinforcement learning, human feedback, interactive robot learning, human-centered robotics

ACM Reference Format:
Jindan Huang, Reuben M. Aronson, and Elaine Schaertl Short. 2024. Modeling Variation in Human Feedback with User Inputs: An Exploratory Methodology. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3610977.3634925

This work is licensed under a Creative Commons Attribution International 4.0 License.

HRI '24, March 11–14, 2024, Boulder, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3634925

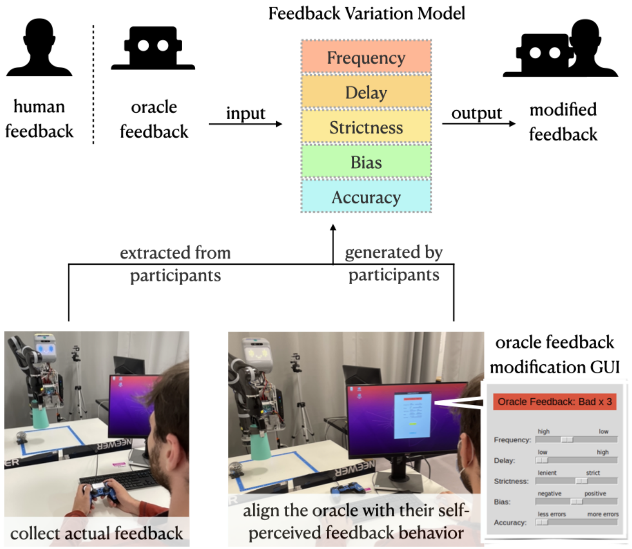


Figure 1: We propose a 5-dimensional model, which synthesizes the most representative feedback variation identified in the prior research, to categorize the gap between oracles and human teachers. The model can integrate with oracle feedback to produce modified feedback with human-like features and can be generated by working with participants.

1 INTRODUCTION

In human-centered robotics research, Interactive Reinforcement Learning (IntRL) is a commonly-used technique that enables efficient learning for intelligent robots by using both environmental observations and feedback from a human instructor. To quickly evaluate the design of IntRL algorithms and expedite the development process, researchers often use oracles, typically perfect oracles, to provide simulated feedback. These perfect oracles are generated from optimal policies or ground truth, delivering dense, instantaneous, and error-free feedback tailored to maximize the benefits for a robot learner.

However, this approach falls short in accurately modeling the heterogeneous feedback patterns exhibited by people. Prior work has shown that human teachers often respond to a robot in a delayed, stochastic and unreliable way [1], and can give different feedback in response to the same observation because of their unique personalities, preferences and experience [2]. Therefore, over-relying on perfect oracles may result in algorithm performance degradation or even failures during the transition from simulation to real-world environments, especially when perfect oracles are used in place of user studies; without evaluation with real users, we do not know if these algorithms will be robust to common sources of variation in human feedback.

In this paper, we aim to characterize the feedback disparities among real participants, and convey these variations to oracles (Fig. 1). This allows researchers to continue using oracles for rapid iteration in algorithm development, while ensuring that algorithms developed in this way are still valid in real-world deployment. To achieve this, we first formally examine the use of oracles in both the state-of-the-art and foundational interactive robot learning research. This is to gain a deeper understanding of the notable disparities that exist between simulated oracles and human instructors, particularly in terms of their feedback behavior. Building upon these insights and results from the literature outside of robotics, we propose a 5-dimensional model that consolidates five representative feedback variations: frequency, delay, strictness, bias, and accuracy. By mathematically defining each attribute, our model can be integrated with the output of a perfect oracle, augmenting the oracle with more human-like features. We demonstrate that the 5 dimensions of variation can influence learning in a simulation experiment. Lastly, we present a proof-of-concept user study to show that the model can be populated from interaction with users in two ways: both by extracting from real feedback data and by directly asking users to set model parameters that align the oracle's behavior more closely with their own.

The major contributions of this paper are:

(1) We conduct a literature review of the use of oracles in foundational IntRL papers and in cutting-edge robot learning publications over the last 3 years of 3 premier venues (HRI, CoRL, RSS), identifying the common sources of feedback discrepancies;

(2) To our knowledge, we are the first to synthesize multiple feedback dynamics into a unified model and mathematically formulate each dynamic in the context of binary feedback;

(3) We apply our model to modify the output of a perfect oracle, and explore the influence of modified feedback on a classic IntRL framework (Q-learning+TAMER) in an OpenAI Gym environment. The results offer valuable insights into how changes in parameter values for each feedback attribute affect the algorithm robustness;

(4) We introduce a mixed-methods approach in a user study to obtain two types of our feedback model with participants: extracted models and self-reported models. The results affirm the feasibility of collaborating with users to create these models and the effectiveness of our approach in understanding feedback disparities.

2 BACKGROUND

Interactive Reinforcement Learning (IntRL), formally introduced in [3] as a branch of Reinforcement Learning (RL), allows a robot to interact not only with an environment but also with a human teacher. Compared to the traditional RL paradigm, IntRL algorithms incorporate a human-in-the-loop to obtain human prior knowledge, and have been proven to be effective for reducing required training time [4] and improving learning performance [5, 6]. Notably, IntRL can be very useful for some special conditions, such as preference learning tasks [7] and sparse-reward environments [8].

Existing IntRL algorithms typically use human feedback to augment *reward functions* [9–11], *policies* [12–14], and *exploration processes* [15–17]. The feedback can be collected from either a real participant or a simulated human (oracle). The idea of using simulated oracles can be traced back to the *Oz of Wizard* methodology [18] proposed by Steinfeld et al. in 2009, which aims to solve the impracticability of performing a large amount of user testing at every iteration of new technology development. Later work has proven that introducing simulated oracles is effective for shortening the development cycle of algorithms and providing useful insights in the early implementation stages [19, 20].

Nevertheless, researchers have also found that results with oracles do not accurately mirror real-world outcomes with human users, since simulated oracles are often generated as perfect oracles, oversimplifying the human feedback behavior [21, 22]. Individuals exhibit their own feedback patterns and variations in human feedback can lead to changes of an IntRL model's performance [23]. Although prior work has made attempts to add some human-like elements to their oracles, such as incorporating errors [24], delay [25] or reducing feedback frequency [26], those efforts often focus on isolated aspects of human feedback discrepancies and prescribe human behavior rather than validating it with actual users. As a result, the development of robust IntRL methods adaptable to feedback from diverse users remains an ongoing challenge.

A systematic understanding of the underlying causes behind the disparity between oracles and people is a preliminary and essential step to address this challenge, however, it appears to be absent in existing work. Therefore, in the next section, we undertake a literature review within the field of interactive robot learning to investigate how oracles are constructed and employed, and to identify the major factors contributing to the feedback divergence.

3 USE OF ORACLES IN THE ROBOT LEARNING LITERATURE

In this section, we delve into a more comprehensive and formal examination of prior research, with a specific focus on the use of oracles and the ways in which they diverge from human teachers. The findings of this literature review help us characterize human feedback discrepancies and motivate how we can mitigate the mismatch between oracles and persons. We select papers exclusively centered on robot learning from simulated and/or real human feedback. The form of feedback can be evaluative feedback, preference labels, and corrective demonstrations. The papers are drawn from two sources: 1) *formal search* on recent publications in premier venues to guarantee the inclusion of state-of-the-art work; and 2) *ad-hoc search* on Google Scholar to identify noteworthy examples that may not be present in the formal search.

For the formal search, we go through the proceedings of HRI, CoRL[1] and RSS[2] conferences over the last 3 years (2020-2022)[3] and we find 13 papers which satisfy our inclusion criteria. Additionally, we include 5 papers from our ad-hoc search, representing the foundational IntRL algorithms over the time period: TAMER [27], Policy Shaping (Advise) [26], SABL [28], COACH [29], PEBBLE [30]. Together, we study where and how the authors obtained the feedback for robots, how they created their oracles, what assumptions

[1] Conference on Robot Learning

[2] Robotics: Science and Systems

[3] We additionally examined the proceedings from HRI 2023 and RSS 2023, which were recently released at the time of our literature search.

they made when adopting oracles to simulate human feedback, and what challenges they encountered when working with human participants or transitioning from simulation to real-world testing.

Figure 2 illustrates the sources of feedback employed in the selected papers. Out of all 18 papers, 3 exclusively evaluate their algorithms using simulated feedback, 5 rely on feedback only from human teachers, and the remaining 10 papers combine feedback from both oracles and participants. We observe that a significant portion (73%) of the research includes oracles, highlighting their prevalent adoption in the IntRL studies. Upon closer examination of the design of oracles in these papers, a common pattern emerges. In all cases, the oracles are derived from either ground truth knowledge (heuristic functions) or optimal policies (fully-trained models). Most work uses a single perfect oracle that consistently delivers immediate and flawless feedback. However, one paper [24] adopts a dual-oracle approach. They incorporated both a perfect oracle and an imperfect oracle with 32.7% error rate to simulate a non-expert human teacher.

Interestingly, among all the work examined, 11 out of 18 (61%) papers acknowledged the discrepancies of feedback behavior between oracles and people. In each of those papers, the authors discussed one or two differences in terms of assumptions required for their research, challenges encountered during user studies, or recognized limitations. Specifically, some research mentioned ***the quality of human feedback*** does not consistently match that of a perfect oracle, as individuals might occasionally make mistakes [24, 31] and they may struggle with providing accurate feedback when robot movements are too subtle to discern [32] or when people themselves lack the necessary abilities [33]. Also, ***the timing of human feedback*** does not match the precision of perfect oracles, as individuals may omit providing feedback [26, 34, 35] or introduce delays in their feedback [27]. Furthermore, ***the feedback strategies of human teachers*** are not homogeneous, as individuals harbor diverse expectations on robot performance - tolerating suboptimal robot behavior [36], biasing to only encourage favorable actions or penalize undesirable ones [28], or extending their teaching objectives beyond mere task performance [30].

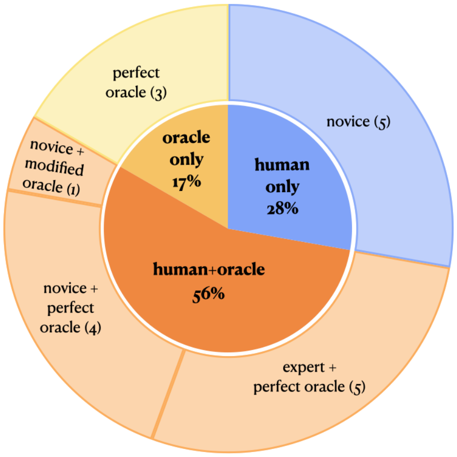


Figure 2: Usage of simulated oracles and participants in the interactive robot learning research we surveyed

Table 1: Feedback variations included in our model

Attribute	Definition	Papers Mentioned[4]
`frequency`	how often the teacher provides feedback	[26, 34, 35, 38–40]
`delay`	how long the teacher needs to react to the learner's action	[27, 41–43]
`strictness`	how willing the teacher is to accept suboptimal solutions	[36, 44, 45]
`bias`	how positive or negative the teacher's feedback is in general	[7, 28, 29, 46]
`accuracy`	how well the feedback reflects the actual performance	[24, 31, 47–49]

Although perfect oracles are commonly used, the heterogeneity of real participants has led researchers to realize many of the limitations of those oracles. This prompts the question of how we can enhance oracles to emulate human behavior more faithfully. Based on the considerations identified in this literature review, we formulate a model for modifying oracles. In the following sections, we will delve into the details of our model (Section 4), and demonstrate how it can effectively capture differences in real user feedback and involve users in the process of creating more realistic oracles (Section 5 and 6).

4 MODELING FEEDBACK VARIATION

In order to maintain the rapid iteration advantages offered by current oracles while addressing their tendency to oversimplify user behavior, one idea is to augment the oracles with feedback patterns that replicate human variability. Few works have explored the integration of imperfect oracles into simulation experiments, introducing errors or timing-related noises to modify the output of a traditional perfect oracle [24, 25, 37]. Using this approach, they effectively assessed their algorithm performance before the human-subject study and ensured the algorithm's robustness when deployed with non-expert participants. Inspired by the success of this oracle modification concept and with the goal of incorporating multiple representative human feedback variations, we introduce a model that categorizes 5 dimensions of feedback dynamics. Our model empowers us to adjust the behavior of a perfect oracle without the need for substantial recreation efforts.

Next, we will explain how we select our model attributes (Section 4.1), how our model can conceptually capture variation in human feedback and modify oracle feedback (Section 4.2), how the altered feedback can impact the robustness of IntRL algorithms (Section 4.3), and how we can obtain model parameters from and with participants (Section 5).

4.1 Model Attributes

We break down primary sources of human feedback variability identified in our literature review into 5 more detailed behavioral

[4] This includes the robot learning papers within our literature review as well as some machine learning and behavioral psychology papers outside of robotics.

features, and we integrate them as our model parameters (see Table 1). *Frequency* and *delay* characterize the timing of the feedback, while *strictness* and *bias* describe the teaching strategies employed by human teachers. Furthermore, *accuracy* reflects the quality of human feedback, indicating the presence of errors or misjudgments. These attributes collectively represent the prevalent feedback discrepancies observed in human teachers. They are grounded in human-robot interaction research and are closely associated with the development of IntRL algorithms. Most importantly, they are straightforward for us to explain and intuitive for non-expert participants to understand, since we hope to collect values of these attributes directly from participants themselves.

In our case, we use the model to study discrepancies in binary feedback (e.g. +1 for desirable robot actions, -1 for undesirable ones), as binary feedback is commonly used for interactive robot learning and is relatively simple to understand compared to other feedback types. However, our model is not limited to binary feedback; it can be extended to other feedback types based on the requirements of the specific learning problem.

4.2 Mathematical Formulation

We mathematically define each model attribute such that they can be used to construct modified oracles and so variation in human feedback can be categorized and described. Here, we introduce the formulation used in our work.

Notation. We model the learning environment as two separate processes: a sequence of robot actions parameterized by $i \in (0, \cdots, N-1)$, and a sequence of feedback instances parameterized by $j \in (0, \cdots, M-1)$. The robot in state x_i performs action a_i starting at time t_i and finishing after a duration d_i; a delay between actions requires that $t_i + d_i < t_{i+1}$. Separately, the teacher provides feedback $\phi_j \in \{-1, +1\}$ at time τ_j. To correlate feedback ϕ with actions a, we define the net feedback for an action f_i as the majority vote over all feedback given by the teacher corresponding to action a_i. Correlated feedback are those whose time τ_j falls between t_i, the beginning of action a_i, and $t_i + d_i + 1$, one second past the end of action a_i; this buffer incorporates delayed responses. The net feedback $f_i \in \{-1, 0, +1\}$ is -1 if there was more negative feedback than positive; +1 if there was more positive feedback than negative; and 0 if there was no feedback or there was an equal amount of positive and negative feedback provided.

Formulation. *Frequency* is calculated by the average amount of feedback assigned to per action:

$$\text{Frequency} = \frac{\#(f_i \neq 0)}{N}$$

Delay is the time between the teacher observing an action and providing feedback. We estimate this as the difference between each feedback time and the start time of the most recent action:

$$\text{Delay}_j = \tau_j - \max_{i,\ t_i < \tau_j} t_i$$

The total delay is found by taking the mean over all feedback delays. We adapt this to simulation by delaying oracle feedback for a set number of time steps. We note that this definition assumes that the feedback given by the teacher corresponds only to the most recent action, which may not always be the case. However, in the user study, we intend to know people's self-awareness of their own delay, which is more naturally measured in time since the most recent action. Furthermore, the robot used in our study has a relatively long action execution time (1.2 seconds), so most real teacher feedback was not delayed longer than the action duration.

Accuracy measures how well the feedback reflects the robot's actual performance. For each action, we determine if the feedback given is *correct* by comparing the observed action a_i with the optimal action $\hat{a}_i$ given by a fully-trained model. Feedback f_i was deemed correct if either $a_i = \hat{a}_i$ and $f_i = +1$ (true positive) or $a_i \neq \hat{a}_i$ and $f_i = -1$ (true negative). We estimate the overall accuracy by taking the ratio of the number of actions a_i that received correct feedback f_i divided by the total number of actions:

$$\text{Accuracy} = \frac{1}{N}\#(f_i \text{ correct})$$

This measures the probability that an action received correct feedback rather than either incorrect feedback or none at all. In other words, we define accuracy as the probability that a person or a modified oracle gives feedback consistent with a perfect oracle for each provided feedback.

Strictness is measured by computing the normalized ranking of the observed action a_i among all possible actions that could have been performed in state x_i; this is possible since we assume the action set A is discrete. We assign the rank $r_i = 1$ if $a_i = \hat{a}_i$ is optimal, $r_i = 0$ if a_i is the worst action, and a value $\frac{k}{|A|-1}$ if it is the k-th from worst. We then compute strictness as:

$$\text{Strictness} = \frac{1}{2}\left(\operatorname*{mean}_{i,\ f_i=+1} r_i + \operatorname*{mean}_{i,\ f_i=-1} (1 - r_i)\right),$$

which is the average minimum ranking that an action must meet to warrant appropriate feedback. If the person is very strict, they will give positive feedback only to highly ranked actions and negative feedback otherwise, resulting in a strictness value close to 1.

Bias is measured by how much more often the user gives positive feedback than would be expected based on an optimal policy. Specifically, we compute the difference between the fraction of feedback that was positive and the fraction of actions that were optimal, then bound the number between 0 and 1:

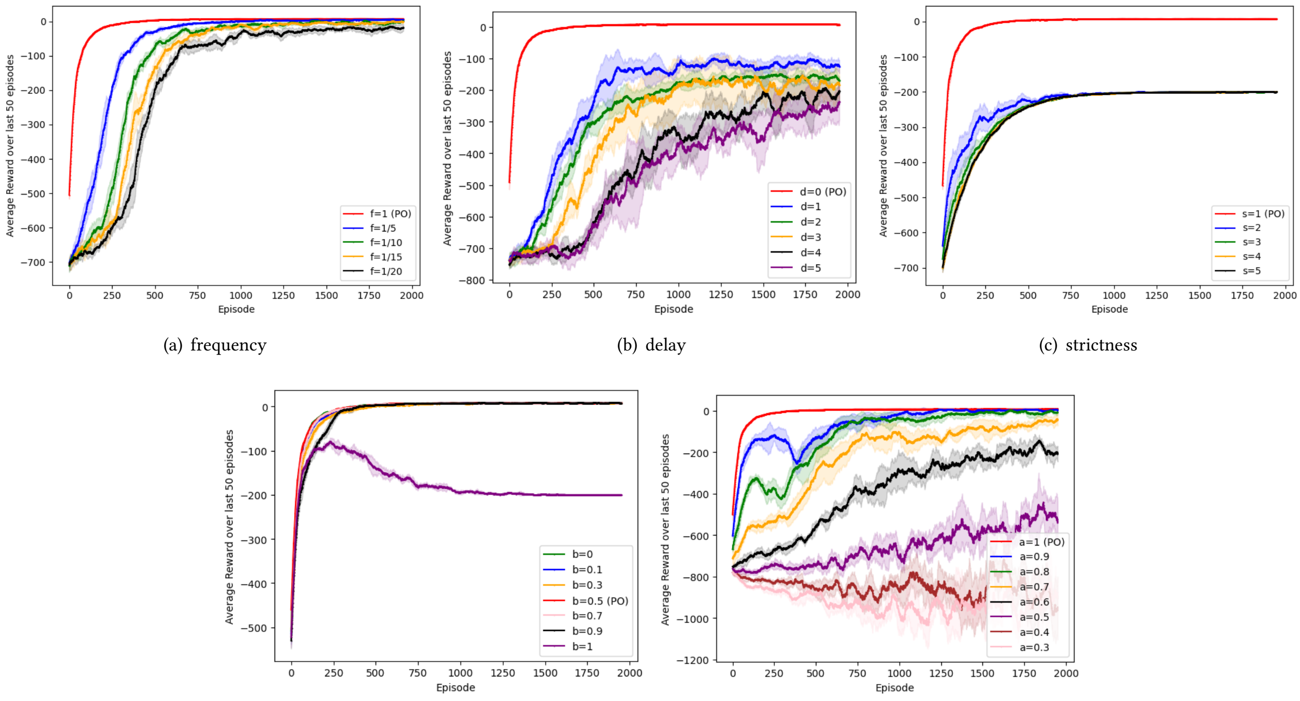
$$\text{Bias} = \frac{1}{2} + \frac{1}{2}\left(\frac{\#(f_i = +1)}{N} - \frac{\#(a_i = \hat{a}_i)}{N}\right)$$

If the person is biased towards giving negative feedback this value will be close to 0. If the person is biased towards giving positive feedback this value will be close to 1. When modifying oracle behavior, we formulate bias as the probability to skip providing negative or positive feedback depending on if the oracle is positive-biased or negative-biased respectively.

4.3 Effect of Model Parameters on Learning

Integrating our model with the output of a perfect oracle can produce modified feedback. In this section, we demonstrate that modified feedback can affect algorithm performance and potentially provide insights about its robustness. To do this, we ran a simulation experiment to examine the influence of model attributes on IntRL algorithms. We choose OpenAI Gym taxi-v3[5] as our testing environment. The task is to pick up and drop off a passenger in

[5] https://gymnasium.farama.org/environments/toy_text/taxi



(a) frequency

(b) delay

(c) strictness

(d) bias

(e) accuracy

Figure 3: Performance of Q-TAMER agent with different modified oracles, grouped by our model attributes. The red line in each subfigure denotes the learning curve of the agent with a perfect oracle (PO).

a grid-world map. We use Q learning + TAMER [47] as our IntRL algorithm because of TAMER's popularity and its capability to deal with feedback delay. We use a fully trained vanilla Q-learning model as our perfect oracle, which achieves the best average reward from the most recent 100 episodes to be 8.98.

Using the techniques outlined in Section 4.2, we modify the oracle to provide imperfect feedback to the learning agent. While real-world feedback variations often arise from a combination of feedback attributes, this section studies the impact of individual attributes on algorithm performance. Thus, we vary only one attribute per trial, keeping the other attributes fixed to match the settings of the perfect oracle. For each feedback attribute value, we repeat the training process 5 times, with 2000 episodes each time.

Figure 3 illustrates the learning curves of the Q-TAMER agent grouped by feedback attributes. We found that frequency, delay and accuracy significantly affect the learning speed. Specifically, lower frequency, longer delay, and lower accuracy tend to result in slower improvement on the average reward. Changing feedback strictness results in a large disparity in learning outcomes between the perfect oracle and the modified ones, where the agent trained with a perfect oracle, which only provides positive feedback when the robot's action is also the best suggested by the oracle, performed significantly better than others. As the oracle becomes less strict and can accept actions that rank lower, the agent's performance deteriorates and eventually becomes unable to learn the task. Changing bias had surprisingly little influence on the agent's performance: only a completely positive-biased oracle (b=1) significantly hindered learning. We suspect this is due to the relatively low-dimensional discrete state space and large amount of allotted time for training. We note that early on in learning, within the first 250 episodes, bias had a much more varied effect on performance. In summary, each feedback attribute had an effect on learning performance in isolation, an effect we expect would be increased when multiple attributes are not consistent with a perfect oracle (as in the case with a human teacher). This suggests that truly robust algorithms need to be tested and developed with models that capture the ways human users vary in terms of these feedback attributes.

5 OBTAINING FEEDBACK VARIATION MODEL: A PROOF-OF-CONCEPT STUDY

In this section, we present a proof-of-concept study to illustrate the use of our model in capturing feedback disparities from participants. This study aims to shed light on three primary aspects: firstly, the variation in actual human feedback in relation to the parameters defined in our model; secondly, the divergent perceptions individuals hold about their feedback behavior when compared to a perfect oracle; and thirdly, the usability of our model for participants to tailor a perfect oracle to replicate their own feedback behavior.

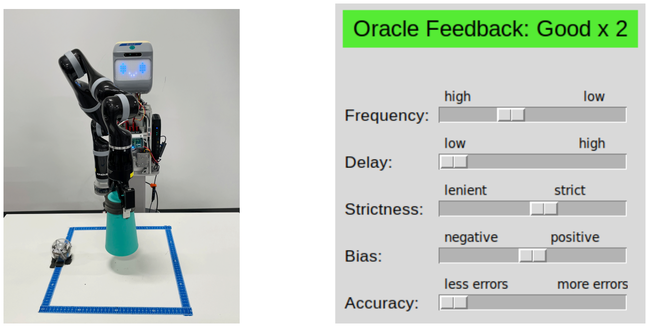
5.1 Experiment Setup

5.1.1 Environment. For the study, we implemented a robot catching environment. The environment includes a Kinova Gen2 arm holding a plastic cup and a Sphero BOLT robot remaining in place (Fig. 4a). The goal for the arm is to learn how to catch the Sphero (i.e. put the cup down over the Sphero). The arm knows if Sphero is caught based on data from Sphero's ambient light sensor. When the arm catches the Sphero or exceeds the maximum number of allotted time steps, an episode ends and the arm resets to a starting position. We model the environment as a Markov Decision Process (MDP) with action space A, state space S, transition function $T : (S, A) \rightarrow S$, and reward function R. A consists of 5 actions: catching (putting the cup down), moving forward, moving backward, moving left, and moving right. S is made up of the *arm end effector position* (p_x, p_y), and *distance between the end effector and Sphero* (d_x, d_y). The robot receives +100 reward if it successfully catches the Sphero and -100 reward for an unsuccessful catch attempt. The arm gets -1 reward after each step. We generated a perfect oracle for this environment which was subsequently integrated into our interactive system.

5.1.2 Oracle Modification GUI. Based on the feedback parameterization outlined in Section 4.2, we developed an interface that allows participants to view and modify the behavior of a simulated oracle as it provides feedback to a robot learner (Fig. 4b). The primary goal of this interface is to obtain people's perception of their feedback behavior (i.e. self-reported feedback model), which provides a user-centered perspective for generating more human-like oracles.

The interface includes a window displaying oracle feedback (e.g. the green area in Fig. 4b) and a set of slider UI elements, each of which controls a specific attribute in our feedback variation model. The values set through the sliders influence the visualization of the oracle feedback. By moving the sliders, participants can change the oracle's feedback-giving behavior to match their own self-perceived feedback-giving behavior. While users interact with the GUI, the robot performs the task repeatedly so that participants can compare the displayed feedback label with the real robot movements in real time and the current parameter settings.

To generate the online feedback display, we first trained a Q-learning agent on our robot catching environment, which achieves 90% catching rate over 30 consecutive episodes within 40 time steps. Then, feedback outputs of the fully-trained agent are modified in real time according to the parameter values specified in the GUI. The initial values of feedback attribute sliders are set to match a perfect oracle. Also, we set the minimum value of the frequency slider to be one feedback per action (1.2 second time gap between two actions, except the catching action, which takes longer than the other actions), because this is a common assumption when researchers use simulated oracles for IntRL algorithms.


Oracle Feedback: Good x 2

(a) (b)

Figure 4: (a) Environment (b) Oracle Modification GUI

5.2 Procedure

We conducted a within-subjects study and each experiment lasted ~1.5 hours. Each participant signed an informed consent form to confirm their eligibility (fluent English speaker, a United States resident, and at least 18 years old) and their permission to use recording devices and automatic transcription service. Participants continued to complete a brief survey collecting their demographics, technology background and previous robot experience. Next, participants went through the following 4 sessions in order:

Understanding teaching styles. Participants were asked to fill out the authoritative teaching questionnaire [50] to assess their general teaching styles. We then asked open-ended questions to know whether people would interact differently with a robot learner compared to a human student, and to understand their attitudes towards robots in general, including any positive, negative or neutral perceptions. This session helps us to identify high-level patterns that may relate to a teacher's feedback behavior.

Collecting human feedback. Participants were given a controller to provide binary feedback to the robot based on its performance, where they pressed "L1" for positive feedback and "R1" for negative feedback. Each participant had 10 minutes to get familiar with the experiment setup. Then, they evaluated 10 trials of the task (each made up of one of five recorded trajectories) for a total of 20 minutes of giving binary feedback. This provides insights into how each teacher *actually* provides feedback.

Modifying oracle feedback. We then proceeded to collect people's perception of their own feedback. Using our oracle modification GUI described in Section 5.1, participants were able to adjust the oracle's behavior. While observing the robot movements, they were encouraged to make the oracle behave in a manner similar to how they had given feedback in the last session. Participants could continue to modify oracle behavior until they were satisfied, and we recorded their final settings. This session allows us to analyze differences between a user's self-reported feedback behavior and their actual feedback behavior.

Reflecting. We conducted a retrospective interview to gather more in-depth information on their experience in the prior sessions. We asked open-ended questions related to their feedback strategy, such as how they decided when to give positive or negative feedback, and their thoughts when modifying the oracle, such as how they perceived themselves and quantified each feedback attribute. We also asked for their opinions about the study interface design.

6 RESULTS

6.1 Participants

The study was approved by the university Institutional Review Board. We recruited 24 participants (16 females, 8 males; aged 18-34) from the campus, and they were compensated $35 for participating in the study. 10 out of 24 participants were from non-STEM majors. 95% of the participants had no prior experience with robots or only little experience with non-industrial robots (e.g. vacuum robots).

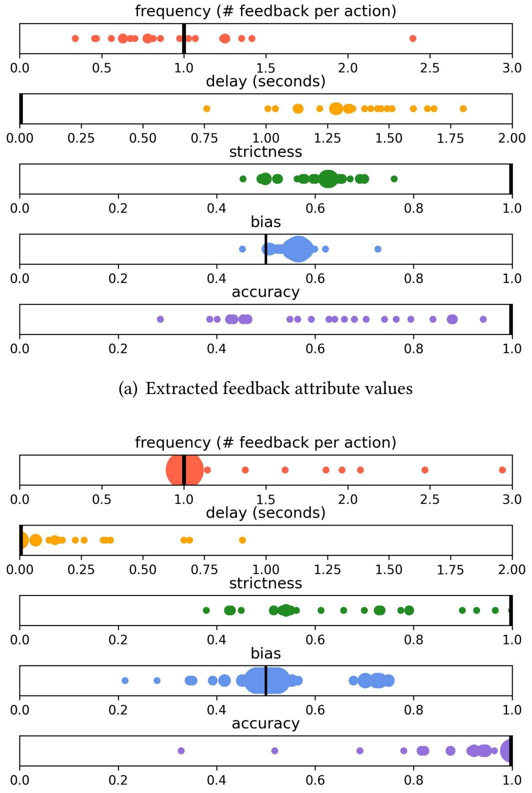
Two participants were excluded for not following study instructions. We used the data from the remaining 22 participants for analysis.

6.2 Modeling Results & Analysis

6.2.1 User's feedback differs from a perfect oracle, varying among individuals. To analyze feedback variations across people, for each participant, we used their feedback data to extract a model of their actual feedback, following the approach mentioned in Section 4.2.

Figure 5(a) visualizes the extracted values from each participant, grouped by model parameters. The results clearly illustrate that people do not behave like a perfect oracle in general. 51% of participants did not give feedback to every action, highlighting the high likelihood of human teachers giving less frequent feedback than oracles. None of the participants had zero delay: they required time to process the robot's movements before responding. The accuracy data reveals that the human feedback did not provide the same quality as the perfect oracle, likely because people had their own teaching criteria and objectives.

Moreover, we found the parameters reflecting the feedback strategies (strictness, bias) exhibited less variation across people than the parameters associated with the timing and quality of feedback (frequency, delay, accuracy). Specifically, 90% of participants displayed a slight positive bias. Also, participants generally appeared to be more lenient than a perfect oracle, with a notable concentration in the 50%-70% strictness range. This could be attributed to the fact that, unlike oracles, individuals often recognize multiple ways to solve a given task and may take into account social factors such as trying to be kind to the robot [51].



(a) Extracted feedback attribute values

(b) Self-reported feedback attribute values

Figure 5: Extracted and self-reported feedback attribute values. The size of each blob represents the number of participants who chose that value (within 0.01). The black vertical line indicates the setting of a perfect oracle.

6.2.2 Users' perception of their feedback also differs from a perfect oracle, and varies among individuals. Figure 5(b) shows the parameter values participants selected for generating oracles that mimic their own feedback behavior. We noticed a unimodal distribution for frequency, delay and accuracy. Specifically, 13 out of 22 participants chose the lowest frequency value, indicating a single feedback signal was given per action. While this aligns with a perfect oracle, this was also the minimum frequency value participants could choose due to the system design. As five participants mentioned during the post-study interview, they might have preferred an even lower value if it were available. Like with frequency, the data from delay and accuracy were heavily skewed. 7 participants believed they had very low delay (≤ 0.01s) and 8 perceived themselves to have very high accuracy (≥ 0.99). This demonstrates that people perceive their feedback behavior to be somewhat similar to that of a perfect oracle in terms of delay and feedback, albeit not identical.

Furthermore, we observed a bimodal distribution of strategy-related attribute values. Participants predominantly perceived themselves as either balanced teachers, providing a mix of positive and negative feedback, or as reward-focused teachers, offering more positive feedback. They also saw themselves as somewhat strict but less so than a perfect oracle, with values centering around 55% and 75% strictness. It is worth noting that this parameter may be task-dependent. In our case, participants could evaluate robot performance by observing the distance between the cup and Sphero, making it quite intuitive for them to judge whether an action was desirable or not.

6.2.3 Comparison between the extracted model and the self-perceived model. To examine how well participants parameterized their feedback behavior, we compared the parameter values of their actual feedback model (Fig. 5a) with their self-reported ones (Fig. 5b). To control for slight differences when applying our feedback model for oracle modification and attribute extraction, we adopted Spearman's correlation test rather than doing a direct comparison. We did not run the test on frequency data because some participants chose the minimum frequency but perceived their frequency lower than the minimum value they can report. We found participants were able to estimate their bias well, as the extracted bias values and the reported ones had a significant positive correlation ($\rho = 0.634, p = .002$), but we did not observe statistically significant results for the other attributes (delay: $\rho = -0.154, p = 0.494$; strictness: $\rho = -0.011, p = 0.962$; accuracy: $\rho = 0.332, p = 0.131$). The result indicates that while participants were aware of the relationship between feedback attributes and their behavior, they were not always precise in quantifying them.

Our post-study data further explains this phenomenon. Participants were requested to list the feedback attributes they found intuitive to comprehend and those they could conveniently adjust

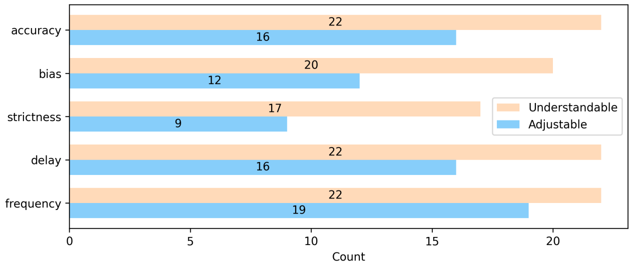


Figure 6: The number of participants who identified each attribute as easy to understand ("Understandable") or easy to adjust ("Adjustable"). There were 22 participants total.

using the oracle modification GUI. We tallied the number of participants who identified each attribute as easy to understand or adjust and present the results in Figure 6. We found that while all feedback attributes were generally intuitive for participants to understand, participants were not always able to report them precisely. Specifically, they found it a little bit harder to adjust strategy-related attributes (bias and strictness) than other attributes. This may be because people are familiar with conceptually describing their strategy but are less familiar with parameterizing it (e.g., P8: "*I think I am positive-biased but did not pay attention to how biased I am when giving feedback*"). This may also stem from the complex and evolving strategies that some participants were trying to communicate through the model (e.g, P6: "*Initially I would tolerate wrong catch actions and allow the robot to explore, but then [when the robot can catch better] I gave more bad feedback to push it to catch faster*").

7 DISCUSSION & CONCLUSION

In this paper, we propose a five-dimensional feedback model that can be used to modify the output of a "perfect" oracle to better reflect common dimensions of variation in human feedback. Our approach provides a means to better describe the robustness of IntRL algorithms when exposed to human-like feedback. The findings in Section 4.3 demonstrate that varying feedback along our model attributes affects learning performance. Those results can be very helpful for rapid prototyping of more robust algorithms.

Our study verifies that our model can be populated from users through two ways: by extracting parameters from actual user feedback, and by having users set the values directly. Both methods enable algorithm designers to take into account the perspectives and abilities of real-world users, even in the early stages of algorithm development where repeated user studies are impractical. The combination of these two methods also offers valuable insights into the origins of the gap between oracles and human instructors. For example, when both the extracted and self-reported values of a model parameter deviate significantly from the settings of a perfect oracle, this implies the fundamental dissimilarity between people's conceptions of teaching robots and the design principles underpinning perfect oracles for improving robot learning.

The analysis performed in Section 6 shows substantial individual variation in feedback behavior, and that users give feedback that does not exactly match the parameters of a perfect oracle. Users' self-reported feedback also does not exactly match their extracted behavior. While precise quantification is difficult for users, we expect that interacting with users to populate the model can allow them to use the model to communicate how they think of their teaching and what they feel was important about their teaching strategies. For example, how users set the accuracy parameter might be used to understand self-efficacy in teaching, and settings of the bias and strictness parameters may reveal differences in teachers' strategies between scenarios (e.g., a school setting vs. a industry setting) or between cultures (e.g., the US vs. Japan). Though further research is needed, our method has the potential to support communication between researchers and users about teaching styles/strategies, and assists researchers to be explicit about the assumptions they make when modeling human teaching.

Limitations & Future Work. Our work mainly investigates discrepancies of binary evaluative feedback. Given that different ways to interact with robots can result in different human teaching behavior [52, 53], we recognize our study results may not generalize to other feedback types, such as natural language feedback. Additionally, we only focus on modeling feedback discrepancies among individuals not the instabilities within an individual's behavior. As we found in the user study, people might change their feedback patterns over time to adapt to robot learning performance. Future work may explore how to incorporate this internal inconsistency to our existing model, such that the refined model can increase the similarity between simulated oracles and human teachers, leading to the development of more robust IntRL algorithms. Finally, while we are able to show that the parameters of our model have an effect on learning, it is outside the scope of this work to develop novel algorithms that optimize performance relative to the model and verify whether the algorithm results in improved performance with human teachers, especially non-experts. Our hope is that this work spurs future efforts in such a direction; with a growing interest in human-in-the-loop learning methods, ensuring that such methods are robust to real user behavior is critical.

Conclusion. This paper introduces a novel user-engaged methodology for modeling variation in human feedback. We consolidate five common feedback discrepancies identified in previous work into a unified model and define mathematical formulations for each model attribute. With the help of those formulations, we successfully derive the model from both on-the-fly human feedback data and participants' self-perception of their feedback behavior. Our modeling results intuitively describe the gap between oracles and individuals, and help to explain the underlying causes of this gap. Rather than replacing human teachers with simulated oracles or relying solely on human studies for algorithm development, our methodology offers a promising path towards enhancing simulated oracles by integrating insights from real user behavior, contributing to the development of robust IntRL algorithms.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation (grant No. IIS-2132887). We thank the AABL Lab at Tufts and especially Isaac Sheidlower for their feedback and help.

REFERENCES

[1] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin-ichi Maeda. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748*, 2018.

[2] Andras Kupcsik, David Hsu, and Wee Sun Lee. Learning dynamic robot-to-human object handover from human feedback. In *Robotics research*, pages 161–176. Springer, 2018.

[3] Andrea Lockerd Thomaz, Guy Hoffman, and Cynthia Breazeal. Real-time interactive reinforcement learning for robots. In *AAAI 2005 workshop on human comprehensible machine learning*, 2005.

[4] Rachit Dubey, Pulkit Agrawal, Deepak Pathak, Thomas L Griffiths, and Alexei A Efros. Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*, 2018.

[5] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45, 2003.

[6] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[7] Paul Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017.

[8] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. A review on interactive reinforcement learning from human social feedback. *IEEE Access*, 8:120757–120765, 2020.

[9] Eric Wiewiora, Garrison W Cottrell, and Charles Elkan. Principled methods for advising reinforcement learning agents. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 792–799, 2003.

[10] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

[11] Hang Yu, Reuben M Aronson, Katherine H Allen, and Elaine Schaertl Short. From "thumbs up" to "10 out of 10": Reconsidering scalar feedback in interactive reinforcement learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4121–4128. IEEE, 2023.

[12] Thomas Cederborg, Ishaan Grover, Charles L Isbell Jr, and Andrea Lockerd Thomaz. Policy shaping with human teachers. In *IJCAI*, pages 3366–3372, 2015.

[13] Samantha Krening and Karen M Feigh. Newtonian action advice: Integrating human verbal instruction with reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 720–727, 2019.

[14] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[15] Chao Yu, Tianpei Yang, Wenxuan Zhu, Guangliang Li, et al. Learning shaping strategies in human-in-the-loop interactive reinforcement learning. *arXiv preprint arXiv:1811.04272*, 2018.

[16] W Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a robot via human feedback: A case study. In *International Conference on Social Robotics*, pages 460–470. Springer, 2013.

[17] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara Grosz. Interactive teaching strategies for agent training. In *In Proceedings of IJCAI 2016*, 2016.

[18] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. The oz of wizard: simulating the human for interaction research. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 101–108, 2009.

[19] Francisco Cruz, Johannes Twiefel, Sven Magg, Cornelius Weber, and Stefan Wermter. Interactive reinforcement learning through speech guidance in a domestic scenario. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2015.

[20] Isaac Sheidlower, Allison Moore, and Elaine Short. Keeping humans in the loop: Teaching via feedback in continuous action space environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 863–870. IEEE, 2022.

[21] Samantha Krening and Karen M Feigh. Interaction algorithm effect on human experience with reinforcement learning. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(2):1–22, 2018.

[22] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. An evaluation methodology for interactive reinforcement learning with simulated users. *Biomimetics*, 6(1):13, 2021.

[23] Christian Arzate Cruz and Takeo Igarashi. A survey on interactive reinforcement learning: Design principles and open challenges. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 1195–1209, 2020.

[24] Guan Wang, Carl Trimbach, Jun Ki Lee, Mark K Ho, and Michael L Littman. Teaching a robot tasks of arbitrary complexity via human feedback. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 649–657, 2020.

[25] Siddharth Reddy, Anca D Dragan, and Sergey Levine. Shared autonomy via deep reinforcement learning. *arXiv preprint arXiv:1802.01744*, 2018.

[26] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. Georgia Institute of Technology, 2013.

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

[28] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, 30(1):30–59, 2016.

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

[30] Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

[31] Jake Brawer, Debasmita Ghose, Kate Candon, Meiying Qin, Alessandro Roncone, Marynel Vázquez, and Brian Scassellati. Interactive policy shaping for human-robot collaboration with transparent matrix overlays. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 525–533, 2023.

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

[33] Ryan Hoque, Lawrence Yunliang Chen, Satvik Sharma, Karthik Dharmarajan, Brijen Thananjeyan, Pieter Abbeel, and Ken Goldberg. Fleet-dagger: Interactive robot fleet learning with scalable human supervision. In *Conference on Robot Learning*, pages 368–380. PMLR, 2023.

[34] Ruohan Zhang, Dhruva Bansal, Yilun Hao, Ayano Hiranaka, Jialu Gao, Chen Wang, Roberto Martín-Martín, Li Fei-Fei, and Jiajun Wu. A dual representation framework for robot learning with human guidance. In *Conference on Robot Learning*, pages 738–750. PMLR, 2023.

[35] Jonathan Spencer, Sanjiban Choudhury, Matthew Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. In *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020.

[36] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. Inquire: Interactive querying for user-aware informative reasoning. In *6th Annual Conference on Robot Learning*, 2022.

[37] Carlos Celemin and Jens Kober. Knowledge-and ambiguity-aware robot learning from corrective and evaluative feedback. *Neural Computing and Applications*, pages 1–19, 2023.

[38] Charles Isbell, Christian R Shelton, Michael Kearns, Satinder Singh, and Peter Stone. A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents*, pages 377–384, 2001.

[39] Daniel Harnack, Julie Pivin-Bachler, and Nicolás Navarro-Guerrero. Quantifying the effect of feedback frequency in interactive reinforcement learning for robotic tasks. *arXiv preprint arXiv:2207.09845*, 2022.

[40] Angel Ayala, Claudio Henríquez, and Francisco Cruz. Reinforcement learning using continuous states and interactive feedback. In *Proceedings of the 2nd International Conference on Applications of Intelligent Systems*, pages 1–5, 2019.

[41] Charles Isbell and Christian Shelton. Cobot: A social reinforcement learning agent. *Advances in neural information processing systems*, 14, 2001.

[42] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*, 2019.

[43] Patrick M Pilarski, Michael R Dawson, Thomas Degris, Farbod Fahimi, Jason P Carey, and Richard S Sutton. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *2011 IEEE international conference on rehabilitation robotics*, pages 1–7. IEEE, 2011.

[44] Raymond G Miltenberger. *Behavior modification: Principles and procedures*. Cengage Learning, 2015.

[45] Matthew E Taylor and AI Borealis. Improving reinforcement learning with human input. In *IJCAI*, pages 5724–5728, 2018.

[46] Andrea L Thomaz and Cynthia Breazeal. Experiments in socially guided exploration: Lessons learned in building robots that learn with and without human teachers. *Connection Science*, 20(2-3):91–110, 2008.

[47] W Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th IEEE International Conference on Development and Learning*, pages 292–297. IEEE, 2008.

[48] Taylor A Kessler Faulkner and Andrea Thomaz. Interactive reinforcement learning from imperfect teachers. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 577–579, 2021.

[49] Zhiyu Lin, Brent Harrison, Aaron Keech, and Mark O Riedl. Explore, exploit or listen: Combining human feedback and policy model to speed up deep reinforcement learning in 3d worlds. *arXiv preprint arXiv:1709.03969*, 2017.

[50] Sigrun K Ertesvåg. Measuring authoritative teaching. *Teaching and Teacher Education*, 27(1):51–61, 2011.

[51] Kerstin Fischer. Interpersonal variation in understanding robots as social actors. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 53–60, 2011.

[52] Gianpaolo Maggi, Elena Dell'Aquila, Ilenia Cucciniello, and Silvia Rossi. "don't get distracted!": the role of social robots' interaction style on users' cognitive performance, acceptance, and non-compliant behavior. *International Journal of Social Robotics*, pages 1–13, 2020.

[53] Pallavi Koppol, Henny Admoni, and Reid G Simmons. Interaction considerations in learning from humans. In *IJCAI*, pages 283–291, 2021.