The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types

Gaurav R. Ghosal[1*], Matthew Zurek[2*], Daniel S. Brown[3], Anca D. Dragan[1]

[1]University of California, Berkeley
[2]University of Wisconsin, Madison
[3]University of Utah
{gauravrghosal,anca}@berkeley.edu, matthewzurek@cs.wisc.edu, dsbrown@cs.utah.edu

Abstract

When inferring reward functions from human behavior (be it demonstrations, comparisons, physical corrections, or e-stops), it has proven useful to model the human as making noisy-rational choices, with a "rationality coefficient" capturing how much noise or entropy we expect to see in the human behavior. Prior work typically sets the rationality level to a constant value, regardless of the type, or quality, of human feedback. However, in many settings, giving one type of feedback (e.g. a demonstration) may be much more difficult than a different type of feedback (e.g. answering a comparison query). Thus, we expect to see more or less noise depending on the type of human feedback. In this work, we advocate that grounding the rationality coefficient in real data for each feedback type, rather than assuming a default value, has a significant positive effect on reward learning. We test this in both simulated experiments and in a user study with real human feedback. We find that overestimating human rationality can have dire effects on reward learning accuracy and regret. We also find that fitting the rationality coefficient to human data enables better reward learning, even when the human deviates significantly from the noisy-rational choice model due to systematic biases. Further, we find that the rationality level affects the informativeness of each feedback type: surprisingly, demonstrations are not always the most informative—when the human acts very suboptimally, comparisons actually become more informative, even when the rationality level is the same for both. Ultimately, our results emphasize the importance and advantage of paying attention to the assumed human-rationality-level, especially when agents actively learn from multiple types of human feedback.

1 Introduction

Reward learning started from the inverse optimal control idea that we can recover the underlying objective when observing optimal behavior (Kalman 1964), and transitioned into AI with the introduction of inverse reinforcement learning (Ng, Russell et al. 2000). While initial research assumed optimal demonstrators (Ng, Russell et al. 2000; Ratliff, Bagnell, and Zinkevich 2006), the field quickly moved to the noisy-rational human model (Morgenstern and Von Neumann 1953): a number of simultaneous works, with different motivations, converged on a Bolzmann (maximum entropy) distribution, where the human actions are exponentially more probable the higher value they are (Baker, Tenenbaum, and Saxe 2007; Ramachandran and Amir 2007; Ziebart et al. 2008; Henry et al. 2010; Vasquez, Okal, and Arras 2014; Kretzschmar et al. 2016; Kitani et al. 2012; Wulfmeier, Ondruska, and Posner 2015; Brown and Niekum 2018; Christiano et al. 2017; Finn, Levine, and Abbeel 2016; Mainprice, Hayne, and Berenson 2015). Often this model would have a "rationality" coefficient β[1] meant to capture how good of an optimizer the human is—setting β to 0 would yield the uniform distribution capturing a random human, while $\beta \to \infty$ would put all the probability mass on optimal actions.

Inspired by the way economists look at preferences, the field then started looking beyond learning from demonstrations to learning from comparisons (Wirth et al. 2017; Christiano et al. 2017; Bıyık et al. 2020). The model was similar: still a Boltzmann distribution, but over two trajectories/actions, instead of over all possible trajectories/actions. Other researchers started looking at a deluge of feedback types: comparisons (Wirth et al. 2017), language (Matuszek et al. 2012), demonstrations (Ng, Russell et al. 2000), trajectory rankings (Brown et al. 2020), corrections (Bajcsy et al. 2017), critiques (Cui and Niekum 2018), e-stops (Hadfield-Menell et al. 2017a), binary feedback (Knox and Stone 2009), and proxy rewards (Hadfield-Menell et al. 2017b). Recently, it was shown that all of these can be interpreted as noisy-rational (Boltzmann) choices (Jeon, Milli, and Dragan 2020), opening the door to learning from all of these feedback types in combination, and even enabling robots to actively select feedback types.

Boltzmann-rationality's ability to unify different feedback types is useful, but the model comes with this one parameter, β, which begs the question: what should we set that to? Prior work often either omits β (implicitly setting it to 1) (Finn, Levine, and Abbeel 2016; Christiano et al. 2017; Ibarz et al. 2018) or sets it to a fixed, often heuristic, value across all feedback types (Ramachandran and Amir 2007; Shah et al. 2019; Bıyık et al. 2020; Jeon, Milli, and Dragan 2020). But demonstrations are sometimes easier or harder to give, depending on the task and the interface, suggesting that β should be adapted to the domain. And comparisons

*These authors contributed equally.

[1]Sometimes denoted by α and sometimes as the inverse ($1/\beta$).

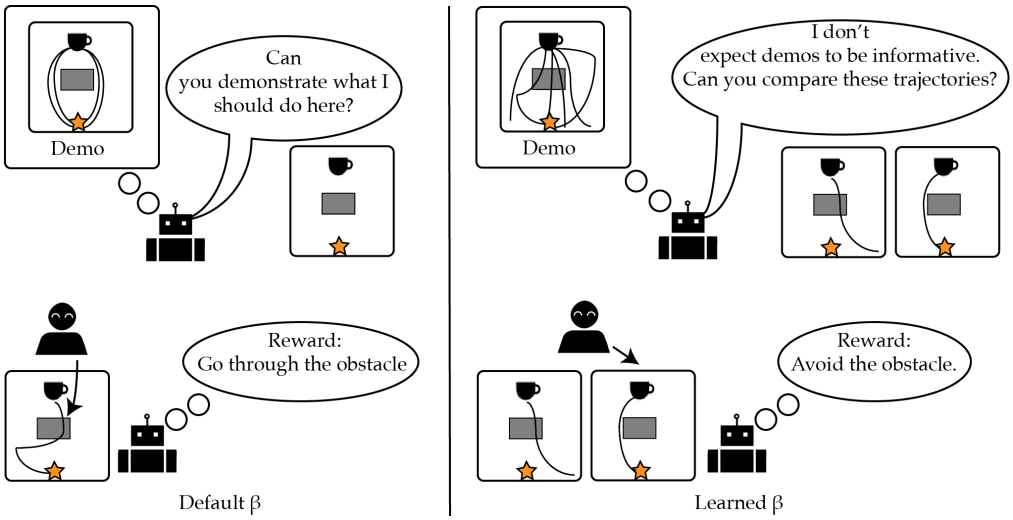


Figure 1: Benefits of Considering Human Rationality. We depict reward inference via active querying over multiple feedback types. In the Default β setting (left panel), the robot erroneously assumes that the human is rational in giving demonstrations and consequently infers the reward poorly. Under the Learned β setting (right panel), the robot anticipates the human's irrationality in demonstrations and is able to query for comparisons instead. As a result, the inferred reward more closely aligns with the human's intention.

might be much easier to answer than demonstrations, suggesting we should be using a higher β for the former. Our goal in this work is to answer the question: does this matter? Are there real benefits to grounding β in real data for each feedback type, or is it safe to stick to a default value?

We analyze this in both simulation and a user study. Our investigation begins in the single-feedback regime, where we isolate the key role played by β in *interpreting feedback*. Through empirical and theoretical analysis, we conclusively demonstrate that overestimating β is particularly harmful for reward inference and that inferring β benefits reward learning by avoiding this. Importantly, our results generalize across varying forms of biased and noisy behavior, beyond Boltzmann rationality. We find that many known systematic biases can be approximated by a noisy-rational model with a learned β, enabling better reward inference. We demonstrate this in simulation with specified biases as well as in a user study, where feedback contains arbitrary human biases.

We next study the importance of β when an agent actively learns from multiple feedback types. In this setting, we observe a new role played by β: in addition to controlling the reward inference, the estimate of β also affects the selection of which type of feedback should be queried. In particular, we see that β strongly affects the informativeness of different feedback types. Surprisingly, this is true even when there is a shared rationality level across feedback types: at low β values, we show that comparisons are more informative than demonstrations, while demonstrations gain an advantage at higher β values. As a result of β's role in both query selection and feedback interpretation, setting it correctly has a significant impact on performance and different settings of beta significantly change the queries selected by active learning. Notably, our findings show the insufficiency of relying on popular heuristics such as starting with demonstrations and fine-tuning with comparisons (Ibarz et al. 2018; Palan et al. 2019; Bıyık et al. 2020; Liu et al. 2023). Rather, we demonstrate that accounting for rationality is essential for active learning to uncover the feedback query that is truly most informative.

Overall, we contribute an analysis of the effects of estimating a human's rationality level on the quality of reward learning and demonstrate the importance of using the β parameter in a principled way over heuristic approaches. In particular, we show that setting β appropriately becomes increasingly crucial as we develop agents that actively learn from multiple types of human feedback. Our analysis can be summarized by the following practical findings:

1. Modeling human behavior with Boltzmann rationality provides benefits even in the face of harder to model systematic biases.
2. When accurate estimates of $\hat{\beta}$ cannot be found, one should err on the side of underestimation.
3. The success of active learning over multiple feedback types depends strongly on an accurate estimate of $\hat{\beta}$.
4. The most informative feedback type varies as a function of the human's rationality, even when the feedback types share a common rationality level.
5. It is possible to obtain good estimates of $\hat{\beta}$ by obtaining a small amount of calibration feedback (where the human optimizes a known reward).

2 Formulation

Preliminaries and Notation. We model the environment as a finite horizon Markov decision process (MDP) with states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, and transition dynamics $P(s' \mid s, a)$. The reward function $r : (S \times A) \to \mathbb{R}$ is initially unknown

to the robot but is communicated by a human through multiple forms of feedback, such as demonstrations of desired behavior, preference comparisons between trajectories, and corrective interventions.

Following prior work (Jeon, Milli, and Dragan 2020), we interpret these varying forms of feedback as a *reward-rational choice* over a (potentially implicit) choice set $\mathcal{C}$. In this work, we study a robot that *actively chooses* from multiple feedback types. To facilitate this, we model a query for human feedback as a Bayesian experimental design problem (Chaloner and Verdinelli 1995). We define feedback types as functions from designs to (choice set, grounding) pairs and active learning as the optimization over designs for information gain. For each possible feedback type, the robot has a choice over different possible experimental designs, $\mathcal{X}$, where the experiment design must be specified before the human can provide data. Given an experiment design $x \in \mathcal{X}$ and a human choice $c \in \mathcal{C}(x)$, we use the grounding function $\varphi(x, c)$ to ground the human feedback into the space of trajectories, Ξ. The grounded trajectory is interpreted to be Boltzmann rational under the human's reward function. A trajectory, ξ, is defined as a sequence of state-action pairs. We use $\xi_{i:j}$ to denote the sub-trajectory starting with the ith state-action pair and ending with the jth state-action pair.

2.1 Human Feedback Types

In this work, we study the three feedback types below. We refer the reader to Jeon et al. (Jeon, Milli, and Dragan 2020) for a discussion of how many other feedback types can be formalized similarly.

Demonstrations can be viewed as a sequence of explicit choices over actions conditioned on states.[2]

The design is all starting states: $\mathcal{X} = \mathcal{S}$ and the grounding function is identity. For a demonstration ξ starting from state s_0 we have the following observation model.

$$\begin{aligned} P(\xi \mid r, \beta) &= \prod_{(s_t, a_t) \in \xi} \pi_\beta(a_t \mid s_t) \\ &= \prod_{(s_t, a_t) \in \xi} \frac{\exp(\beta Q_t^{\text{soft}}(s_t, a_t \mid r))}{\sum_{b \in \mathcal{A}} \exp(\beta Q_t^{\text{soft}}(s_t, b \mid r))} \end{aligned} \tag{1}$$

where $Q_t^{\text{soft}}(s, a \mid r) = r(s, a) + \gamma \mathbb{E}_{s'}[V_{t+1}^{\text{soft}}(s')]$, and $V_t^{\text{soft}}(s) = \mathbb{E}_{a \sim \pi_\beta}[Q_t^{\text{soft}}(s, a) - \log \pi_\beta(a \mid s)]$ are the soft Q-function, and Value function, respectively (Kitani et al. 2012; Haarnoja et al. 2017), and π_β is the corresponding (time-dependent) policy.

Comparisons are a choice between two trajectories. Thus, the possible designs are $\mathcal{X} = \Xi^2$, all pairs of trajectories and the grounding function maps to the preferred trajectory. The likelihood that the human prefers trajectory A over B is given by the Bradley-Luce-Shepherd rule (Bradley and Terry 1952; Christiano et al. 2017):

$$P(\xi_A \mid r, \beta) = \frac{\exp\left(\beta \cdot r(\xi_A)\right)}{\exp\left(\beta \cdot r(\xi_A)\right) + \exp\left(\beta \cdot r(\xi_B)\right)} \tag{2}$$

E-stops represent the intervention of a human telling the robot to stop rather than continue its trajectory. We assume that the human is able to observe the robot's planned trajectory and then selects a desired stopping point t at which point the episode terminates. Thus, the space of possible designs is $\mathcal{X} = \Xi$, all trajectories. The choice set is the stopping time t, and the grounding function is the sub-trajectory $\xi_{0:t}$. Given a robot trajectory ξ, we have the following likelihood function for the human's choice $c_h = t$:

$$P(t \mid \xi, r, \beta) = \frac{\exp(\beta \cdot r(\xi_{0:t}))}{\sum_{k=0}^{T} \exp(\beta \cdot r(\xi_{0:k}))}. \tag{3}$$

2.2 Estimating a Human's Rationality Level from Data

Rather that assuming a known or constant value for the rationality coefficient, we study the effect of learning an estimate of the human's rationality level, $\hat{\beta}$, from human data. As a vehicle for our analysis, we consider access to a separate calibration phase where we present the human with a known, calibration reward function, r', and then ask them to provide feedback (e.g., demonstrations, comparisons, e-stops) with respect to this reward function. The benefit of this calibration is that given human feedback that corresponds to a known reward function, we can find $\hat{\beta}$ that maximizes the log-likelihood of Eq. (3):

$$\begin{aligned} \hat{\beta} = &\arg\max_{\beta} \beta \cdot \mathbb{E}_{\xi \sim \varphi(x, c_h)}[r(\xi)] \\ &- \log \sum_{c \in \mathcal{C}} \exp\left(\beta \cdot \mathbb{E}_{\xi \sim \varphi(x,c)}[r(\xi)]\right), \end{aligned} \tag{4}$$

since c_h is collected from the human during calibration and r', x, C, φ are constant and known to the robot. This approach intentionally favors simplicity over real-world practicality—our focus is on assessing the importance of having a good model of the human's rationality level—in practice, one could also fit $\hat{\beta}$ on human data optimizing an unknown r by marginalizing over r.

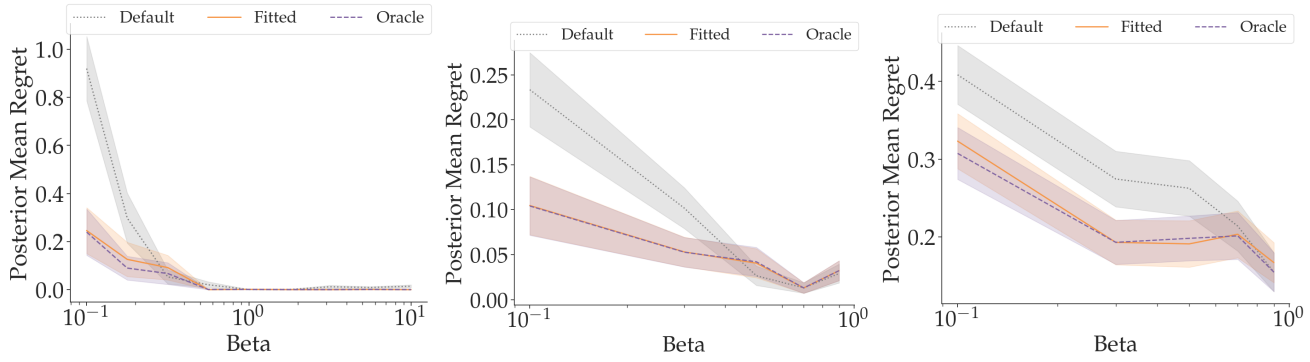
2.3 Active Learning over Feedback Types

We consider the scenario in which the robot can actively query the most informative feedback given its current belief over a parameterized reward function. We can cast this as the problem of selecting a design $\mathcal{X}$ which optimizes the expected information gain over the possible human feedback induced by $\mathcal{X}$. Concretely, this can be written as the following optimization problem

$$\max_{x \in \mathcal{X}} \mathbb{E}_{c_h \sim P(c_h \mid x)} \left[D_{KL}\big(P(\theta \mid c_h, x) \| P(\theta)\big) \right], \tag{5}$$

in which we consider $P(\theta)$ to be our prior distribution over the reward function, r_θ, parameterized by θ.

[2] Jeon et al. (Jeon, Milli, and Dragan 2020) model human demonstrations as choices over all possible trajectories; however, with stochastic dynamics, human actions are conditioned on observed state transitions and the human cannot pre-select a specific trajectory.



(a) Demonstrations (b) Comparisons (c) E-Stops

Figure 2: Reward Inference on Simulated Boltzmann-Rational Feedback. We compare the rationality-fitting (Fitted) method which learns $\hat{\beta}$ from data to a Default method which assumes $\hat{\beta} = 1$ and an Oracle method which sets $\hat{\beta} = \beta^*$. Results show that Fitted (our approach) significantly outperforms Default, achieving similar performance to Oracle. Note that due to the overlap between fitted (orange) and oracle (purple) both curves appear red.

3 Effect of $\hat{\beta}$ When Learning from a Single Feedback Type

In this section, we test the following hypothesis:

H1: *Reward inference with a fitted beta will perform better than reward inference with a default beta across different feedback types and different forms of human irrationality.*

We test hypothesis **H1** over different forms of simulated feedback: first Boltzmann-rational and then systematically biased. In Section 5, we test **H1** on real human data.

Metrics. We measure reward inference performance via the normalized regret from optimizing the posterior mean reward i.e., the difference measured in the ground truth reward between optimizing the ground truth reward vs optimizing the posterior mean inferred reward. In Appendix C, we evaluate other reward inference metrics, which show similar trends.

Experimental Design. We follow the same experimental design when fitting $\hat{\beta}$ and performing reward inference for both simulated and real human feedback. When fitting $\hat{\beta}$, we use 4 randomly chosen calibration reward functions and query the human for feedback 5 times for each calibration reward. When performing reward inference we query the human for feedback 5 times and then perform reward inference using the previously calibrated $\hat{\beta}$. When analyzing learning from individual feedback types, we randomly sample designs. In Sec. 4 we examine actively selecting designs.

Experimental Domains. Our simulation experiments take place in a suite of discrete gridworld navigation environments. Reward functions, r_θ, are parameterized by a linear combination of features that indicate the color of each grid-cell (see Appendix B for more details). In order to perform exact Bayesian inference, we discretize the reward space by 1000 points. Additionally, in Appendix K, we provide evidence of similar results in a self-driving car domain.

3.1 Learning from Simulated Boltzmann-Rational Feedback

Results. We assess the importance of beta fitting for demonstrations, comparisons, and e-stops by running reward inference on simulated Boltzmann-rational feedback generated with different β values. We compare reward inference using the fitted $\hat{\beta}$ (*Fitted*) to a default method that sets $\hat{\beta} = 1$ (*Default*) and an oracle method that has access to the ground truth β value (*Oracle*). The results in Fig. 2 demonstrate that when feedback is highly sub-optimal, *Fitted* results in significantly better inference than *Default*, and performs comparably to *Oracle*. These observations support **H1**—using a learned value for $\hat{\beta}$ improves performance, especially in cases where the human acts more noisily.

Remark: Under-estimating β is better than over-estimating it. In Fig. 2, we observe an asymmetry between the settings of over- and under-estimating β. We see that while over-estimation results in poorer performance, under-estimation does not harm reward inference performance as much. In what follows, we present some intuition for this phenomenon. In particular, we show that using a lower $\hat{\beta}$ is risk-averse when the human is suboptimal but is still a good choice even when the person is optimal, whereas a high value of $\hat{\beta}$ leads to poor reward inference when the human is suboptimal. To simplify notation, we define $r(c) \triangleq \mathbb{E}_{\xi \sim \varphi(x,c)}[r(\xi)]$ for $c \in \mathcal{C}$.

Proposition 1. *If the human is optimal, then r^* is an MLE estimate for any value of $\hat{\beta} \in [0, \infty)$.*

Proof. An optimal human (i.e., $\beta = \infty$) never makes mistakes. Thus, $r^*(c_h) \geq r^*(c), \forall c \in \mathcal{C}$ and

$$r^* \in \arg\max_r e^{r(c_h)} = \arg\max_r e^{\hat{\beta} \cdot r(c_h)} = \arg\max_r P(c_h | r, \hat{\beta}). \tag{6}$$

Thus, r^* is an MLE estimate given c_h, $\forall \hat{\beta} \in [0, \infty)$. □

Even though the MLE reward may not change, the shape of the posterior distribution over r is strongly influenced by the choice of $\hat{\beta}$. When the human is suboptimal, we want to have robots that hedge their bets, rather than becoming overly confident in their estimate of the true reward function. Prior work proposes the Shannon entropy over the robot's belief $P(r \mid c_h)$ as a quantitative measure of the robot's confidence (Jonnavittula and Losey 2021). Using this definition, we present the following result.

Proposition 2. *The robot becomes more (over-)confident as $\hat{\beta}$ increases.*

Proof. If $\hat{\beta} = 0$ and we have a uniform prior, then we have a uniform belief distribution over r, resulting in maximum entropy and risk-averse behavior. As $\hat{\beta}$ increases, the posterior distribution concentrates on only a small number of reward functions, resulting in lower entropy and less risk-aversion. To see this note that

$$\begin{aligned} P(r \mid c_h, \hat{\beta}) \propto & \frac{\exp(\hat{\beta} \cdot r(c_h))}{\sum_{c \in \mathcal{C}} \exp(\hat{\beta} \cdot r(c))} P(r) \\ = & \frac{1}{1 + \sum_{c \neq c_h} \exp(\hat{\beta} \cdot (r(c) - r(c_h)))} P(r) \end{aligned} \tag{7}$$

As $\hat{\beta} \to \infty$, we have $\exp(\hat{\beta} \cdot (r(c) - r(c_h))) \to 0$ if $r(c) < r(c_h)$ and $\exp(\hat{\beta} \cdot (r(c) - r(c_h))) \to \infty$ if $r(c) > r(c_h)$. Thus, $P(r \mid c_h, \hat{\beta}) \to 0$ as $\hat{\beta} \to \infty$ if c_h does not maximize $r(c)$ and $P(r \mid c_h, \hat{\beta}) \propto P(r)$ if c_h uniquely maximizes $r(c)$. If c_h is a non-unique maximizer of $r(\phi(x, c))$, then we have $P(r \mid c_h, \hat{\beta}) \propto P(r)/|\{c : r(c) = r(c_h)\}|$. □

The above result shows that, as $\hat{\beta}$ increases, the Shannon entropy decreases and the robot places high probability on a smaller set of reward functions, thereby behaving very confidently about its estimate of the reward. Finally, we have the following result for when the human makes a sub-optimal feedback choice.

Proposition 3. *If the human makes a suboptimal feedback choice, the likelihood of the true reward, r^*, decreases exponentially as $\hat{\beta}$ increases.*

Proof. For a suboptimal choice c_h, $\exists c^*$ such that $r^*(c_h) < r^*(c^*)$ and

$$\begin{aligned} P(c_h | r^*, \hat{\beta}) &= \frac{\exp\left(\hat{\beta} r^*(c_h)\right)}{\sum_c \exp\left(\hat{\beta} r^*(c)\right)} \\ &\leq \frac{\exp\left(\hat{\beta} r^*(c_h)\right)}{\exp\left(\hat{\beta} r^*(c^*)\right)} = \exp\left(\hat{\beta}(r^*(c_h) - r^*(c^*))\right) \end{aligned} \tag{8}$$

By assumption $r^*(c_h) - r^*(c^*) < 0$. Therefore, the likelihood decreases exponentially as $\hat{\beta}$ increases. □

This analysis supports our empirical findings and shows overestimating $\hat{\beta}$ can have negative consequences since it makes the robot overconfident in the human data, potentially overfitting to mistakes. On the other hand, using a lower $\hat{\beta}$ leads to more risk-averse behavior (beneficial when the human is suboptimal), while still being optimal (under a uniform prior) when learning from perfectly-rational humans.

3.2 Learning from Simulated Biased Feedback

The results in Section 3.1 demonstrate the importance of modeling a human's rationality level when the human actually is Boltzmann-rational. But of course, human behavior can suffer from systematic biases and irrationalities, not just Boltzmann (ir)rationality (Evans, Stuhlmüller, and Goodman 2016; Alanqary et al. 2021). In this section, we study whether we can use the model of Boltzmann rationality, with a learned rationality level, to improve reward inference in the presence of biases commonly exhibited in human behavior (Guan et al. 2015; Do, Rupert, and Wolford 2008; Sharot et al. 2007; Thompson 1999). **H1** hypothesizes that inferring beta can help compensate for these unmodelled aspects of human behavior and therefore lead to better reward inference, in particular when the impact of the bias is not consistent across feedback types. We first evaluate this hypothesis by generating simulated feedback with various biases unknown at reward-inference time. Later we evaluate this hypothesis with a user study.

Types of Simulated Bias. We study several different models of human biases. Following (Chan, Critch, and Dragan 2021), we formalize each bias as a particular modification to the standard Bellman update, resulting in a modified value function which we use to determine the resulting policy and simulate choices from a biased human. We assume that the person is Boltzmann rational under their biased value function. Thus, β remains a parameter, and we set it to 1 in all cases; however, the presence of the bias means that the human is actually *not* Boltzmann β-rational for any $\beta > 0$.

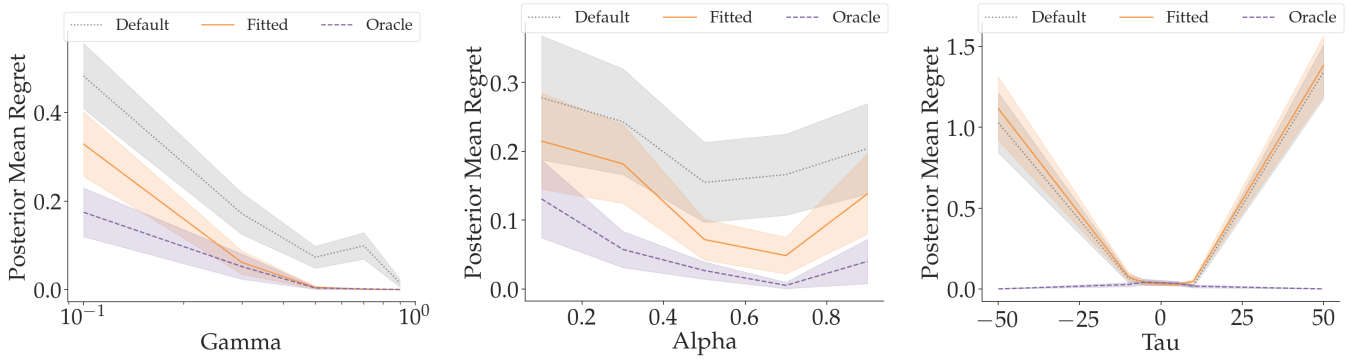
Myopia Bias: Humans sometimes demonstrate myopic behavior, concentrating on immediate rewards without evaluating the longer-term impacts of their actions (Guan et al. 2015; Grüne-Yanoff 2015). We simulate myopic human feedback by changing the discount factor $\gamma \in [0, 1]$ and then providing Boltzmann-rational feedback with respect to the value function computed using this discount factor.

Extremal Bias: Humans sometimes pay attention to high-intensity aspects of an experience, at the exclusion of lower-intensity events (Do, Rupert, and Wolford 2008). We model this behavior using a modified Bellman update

$$V_{i+1}(s) = \sum_{s' \in S} P(s'|s, a) \max(r(s, a), (1-\alpha) r(s, a) + \alpha V_i(s')), \tag{9}$$

where $\alpha \in [0, 1]$. As $\alpha \to 1$ the human seeks to maximize the maximum reward obtained at any point within a trajectory. As $\alpha \to 0$, the human maximizes immediate reward.

Optimism/Pessimism Bias: Humans can sometimes over- or under-estimate the likelihood of experiencing a good or bad event (Sharot et al. 2007). We simulate this bias by changing the transition function that the biased human uses for planning to reflect the fact that the human believes that the



(a) Myopia (b) Extremal (c) Optimism

Figure 3: Reward Inference on Simulated Biased Feedback. We compare the β-fitting method (Fitted) introduced in this work with a Default method which assumes $\beta = 1$ and an Oracle method that performs reward inference with access to a perfect model of the biased human. Fitted shows improvements over Default for myopia and extremal biases, but shows no improvement for the optimism bias.

likelihood of an outcome depends on its value:

$$\tilde{P}(s' \mid s, a) \propto P(s' \mid s, a) \cdot \exp\left(\tau \cdot (r(s, a) + \gamma V_i(s'))\right), \tag{10}$$

where $\tau \in \mathbb{R}$. As τ increases (decreases) the human becomes more optimistic (pessimistic).

Results. We study three variants of reward inference: (1) *Default* assumes $\hat{\beta} = 1$ and performs inference under the corresponding Boltzmann-rational model of feedback, (2) *Fitted* first fits the rationality parameter $\hat{\beta}$ and then performs reward inference using the $\hat{\beta}$-Boltzmann rational model, and (3) *Oracle* performs reward inference using the true model of the bias. We present the results for demonstrations in Fig. 3 and refer the reader to Appendix D for results on comparisons and e-stops.

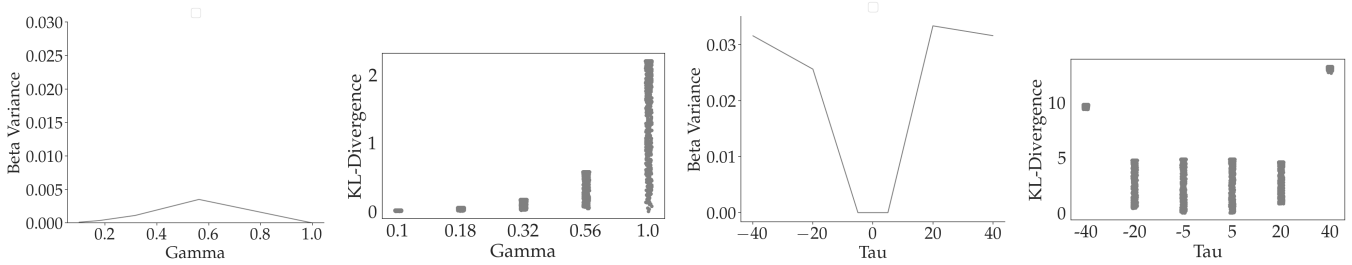
Our findings in Fig. 3 provide evidence for **H1**: using a learned value for $\hat{\beta}$ overall results in lower regret than using a default value. However, we find that there is often a large gap between the performance of Fitted and Oracle. This demonstrates that while there is utility in estimating the rationality level of the human, it is not always possible to accurately model systematically biased behavior using a tuned Boltzmann rationality model. In particular, for the Optimism bias, we find that both Fitted and Default perform similarly, and that as the τ parameter diverges from 0 (diverging from Boltzmann rationality) regret increases.

Understanding the Success and Failure of Beta Fitting for Biased Human Feedback. To understand when β-fitting improves reward inference, we study $\hat{\beta}$-generalization and quality of fit on the Myopia and Optimism biases. We consider the infinite data limit of maximum likelihood estimation, which, as shown in Appendix F, can be calculated from the biased demonstration policy via an adapted policy evaluation technique.

In Fig. 4 (a) and (c), we show the variance of the MLE $\hat{\beta}$ over different reward functions at various levels of Myopia and Optimism bias, respectively. A lower variance in this experiment implies that $\hat{\beta}$ generalizes well across different reward functions for that bias setting. The uniformly low variances for the myopia bias suggest that the fitted $\hat{\beta}$ remains consistent over different reward functions, while the higher variances for the optimism bias suggest that the fitted rationality coefficient, $\hat{\beta}$, is less transferable across different reward functions.

In Fig. 4 (b) and (d), we fixed the ground truth reward θ and show scatter plots of the KL-Divergence between the biased policy on θ and the soft-optimal policies for a large sample of other rewards, θ'. We generate these scatter plots for different bias settings. We observe that the Myopia bias has some rewards with low magnitude KL-Divergences to the biased policy, indicating that the biased policy can be fit well by Boltzmann rationality. On the other hand, for some settings of optimism bias (such as $\tau = \pm 40$), the magnitude of the KL-Divergence is uniformly high, indicating that Boltzmann rationality cannot fit the biased policy well with any reward function. Interestingly, neither a low beta variance nor a low KL-Divergence ensures that β-fitting can recover the true reward. For example, when $\gamma = 0.1$ both β variance and KL-divergence are close to 0, yet fitted performs worse than oracle.

In Fig. 4, we see that under some bias settings, such as $\tau = \pm 40$ or $\gamma = 0.1$, all the soft-optimal policies for the sampled rewards (θ') are roughly equidistant from the biased policy (forming a tight cluster in the scatter plot). In these settings, β-fitting fails since the true reward cannot be uniquely identified—all rewards appear to model the biased behavior equally well. Comparing Fig. 4 (b) and (d), we see this situation can arise both when the biased behavior can (in the case of myopia $\gamma = 0.1$) and cannot (in the case of optimism $\tau = \pm 40$) be modeled well by the Boltzmann distribution, as measured by the scale of the KL-Divergences. We leave further analysis of which biases preserve reward identifiability to future work.



(a) Myopia Beta (b) Myopia KL (c) Optimism Beta (d) Optimism KL

Figure 4: Quality of Fit and Generalization of Beta for Biased Demonstrations. In (a) and (c), we compare the variance of $\hat{\beta}$ over all reward functions in our discretization, when fitting on simulated Myopic and Optimistic demonstration behavior. This indicates how well $\hat{\beta}$ generalizes across reward functions. In (b) and (d), for a fixed reward function θ, we show a scatter plot of the KL-Divergence between the soft-optimal policy for every other reward function θ' and the Myopic or Optimistic policy on θ. This indicates how well the respective biases are modelled by Boltzmann-rationality and how identifiable the ground truth reward (θ) is.

4 Effect of $\hat{\beta}$ When Actively Learning over Multiple Feedback Types

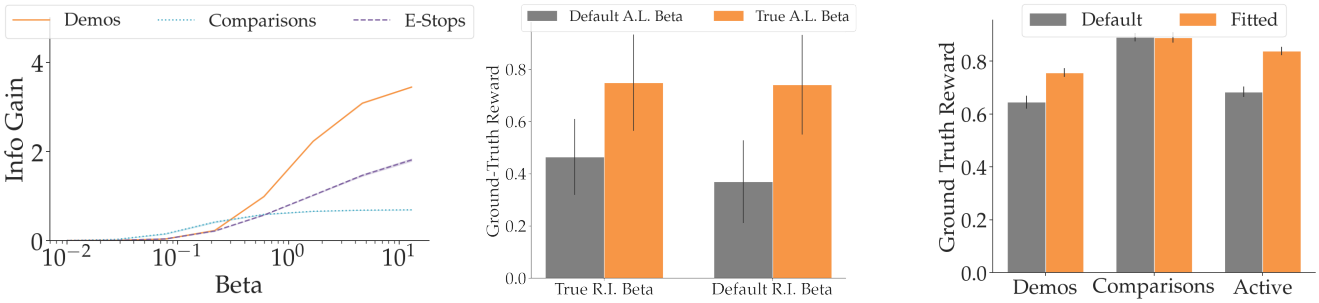
In the previous section, we observed fitting β is beneficial when learning passively from a single feedback type. In this section, we consider allowing the robot to actively select the feedback it receives from a set of multiple feedback types, taking into account its current belief over the reward function. When performing this selection, an additional notion of *feedback informativeness* plays a key role. We first investigate the interaction between rationality and feedback informativeness, followed by an analysis of the overall reward inference performance in the active learning setting. Ultimately, we test the following hypothesis:

***H2**: Active learning that decides what feedback to ask for will perform better with a fitted beta for each type than with a default beta.*

Rationality and Feedback Informativeness. We first examine the information gain provided by the different feedback types when the human is equally rational across all-types. In this case, it may appear intuitive that demonstrations would uniformly provide the most information gain because they represent an implicit choice over *all* possible trajectories. However, our results in Fig. 5a reveal that the most valuable human feedback type is a function of the common rationality parameter β. While demonstrations do provide the most information when the human is highly rational, comparisons gain an advantage when querying a more irrational human.

In Appendix E we further explore this surprising finding in a toy environment. We construct a toy reward inference environment with a finite set of reward functions and choices. The structure of the reward function and the choice set of this environment means that, given a ground truth reward, only a subset of the choices will be sensitive to correctly identifying this reward. Intuitively, this means that there will be many "uninformative" choices in the choice set. An uninformative choice will result in poor information gain, as the posterior reward distribution will remain largely unchanged. We model *demonstrations* as a feedback query where the user may choose any choice from the entire choice set. As $\beta \to 0$, demonstrations become increasingly noisy and the human converges to choosing uniformly from all choices, yielding a higher probability of making an uninformative choice and reducing the expected information gain. On the other hand, *comparisons* restrict the choice set to two elements. Thus, it is possible to construct a comparison query which eliminates uninformative choices and therefore has a higher expected information gain. This analysis confirms the trend we see in Fig. 5a, where the rationality coefficient β has a strong influence on the informativeness of different feedback types.

Importance of Beta Fitting for Active Reward Learning. In practice, a human is likely to have have varying degrees of rationality across feedback types. Intuitively, β plays two roles in this setting: it affects the kinds of queries that are selected as well as the interpretation of the response. In order to gain a more complete understanding of the impact of β in this regime, we seek to disentangle the relative importance of these roles. We study the reward inference performance of four variants of active learning, where each of the 2 steps (query selection and reward inference) has either the correct or default β. In Fig. 5, we show these results for one choice of relative rationalities ($\beta_{\text{demo}} = 0.1$, $\beta_{\text{comp}} = 10, \beta_{\text{estop}} = 1$) and in Appendix I, we consider the case where demonstrations are the most rational but all feedback rationalities are overestimated by default. We observe that the relative importance of having the correct β for query selection vs. reward inference varies significantly between these cases. When comparisons have a much higher rationality than demonstrations (shown in Fig. 5), the value of β used for active learning plays a significant role in quality of reward inference, because when comparisons are selected, the default β underestimates rationality. On the other hand, Appendix I shows that when the default values of β are all overestimates and demonstrations are the most rational, then the β used for reward inference has a significant effect on performance. Ultimately, our results reveal an in-



(a) Info Gain vs. Rationality Level (b) Active Learning Ablation (c) User Study

Figure 5: Learning Rewards from Multiple Feedback Types. (a) Given the assumption that the human is β rational for all three feedback types, this figure displays the information gain of the most informative design for each feedback type. We observe that one feedback type does not uniformly dominate across all rationality levels, but that the most informative feedback type is a function of human rationality. (b) We study the effect of using a misspecified β for active learning (A.L.) and reward inference (R.I.). We find that having access to the correct β is particularly important for active learning in this setting. (c) Results from a user study show that β-fitting helps when learning from demonstrations (designed to be hard) and is no worse for comparisons (designed to be easy) and is especially beneficial when performing active learning over both demonstrations and comparisons.

tricate dependency of the optimal active learning strategy on the human rationality and underscores the insufficiency of generic heuristics and the importance of calibrating to individual biases and noise levels.

5 User Study

We conducted a small-scale user study of $N = 7$ users (aged 21-58, mean = 28) in order to test the effect of conducting β-fitting on real-world human irrationality. The user study took place in the same grid-world navigation setup as the simulated experiments and each user provided a set of 5 comparisons and 5 demonstrations for each of 5 reward functions. In our setting, the interface for demonstrations was more challenging than that for comparisons, due to the presence of slippery dynamics in the demonstrations control interface (see Appendix J for more details). For each reward function, we tested the reward inference by using the data corresponding to the 4 other reward functions to fit $\hat{\beta}$ and then running reward inference using the individual feedback types, as well as active selection from both. The comparison of performance between using $\beta = 1$ (Default) and $\beta = \hat{\beta}$ (Fitted) are shown in Figure 4(c). We observe that the results validate both **H1** and **H2**. β-fitting on demonstrations results in better performance than using the default β and we observe a particularly large benefit from fitting beta in the active learning setting. For comparisons, we observed that the users were able to perform close-to-optimally, which lessened the importance of modeling the rationality level.

6 Discussion

Summary: In this work, we examine the importance of modeling the level of human rationality when learning from multiple kinds of human feedback. We demonstrate the importance of utilizing the correct rationality coefficient in cases where the human is Boltzmann-rational (with an unknown rationality level), as well in cases where the human is *not* Boltzmann-rational, but is systematically biased. Finally, we demonstrate that β-fitting is especially important when performing active learning: in a user study we find that active queries based on learned rationality levels significantly outperform an active learning baseline that uses a uniform, default level of rationality across feedback types.

Limitations and Future Work: Our contribution is studying the importance of having an estimate of β (the human's level of rationality), but how exactly to get that remains an open question—our experiments use calibration data, assuming that we can "incept" a calibration reward function into a human's head and then ask them to provide feedback. We note that this type of calibration approach has been shown to work well in some settings, such as humans interacting with a driving simulator (Schrum et al. 2022); however, in other settings, this type of calibration may be difficult, and future work includes studying β-fitting techniques that do not require providing the human with an explicit calibration reward function. Furthermore, while we study the benefits of β-fitting on actively learning from multiple feedback types, we have only modeled the rationality level of each feedback type, ignoring the query cost in terms of cognitive burden and time required. Future work includes learning models of the cognitive burden per user and per feedback type, incorporating cognitive and feedback-time costs into active learning, and analyzing β-fitting in more domains.

Ethics Statement

Our work seeks to approximate potentially biased behavior with noisily rational behavior. This could have negative societal impacts if it leads a robot to incorrectly infer human intent, especially in safety critical settings. While our results show that β-fitting is useful, we caution against simply forcing robots to view all human behavior as β-rational—using more nuanced and sophisticated models of human bias and irrationality is an important area of future work.

Acknowledgments

We thank the members of the InterACT lab for helpful discussion and advice. This work was supported by the ONR Young Investigator Program (YIP).

References

Alanqary, A.; Lin, G. Z.; Le, J.; Zhi-Xuan, T.; Mansinghka, V. K.; and Tenenbaum, J. B. 2021. Modeling the mistakes of boundedly rational agents within a Bayesian theory of mind. *arXiv preprint arXiv:2106.13249*.

Bajcsy, A.; Losey, D. P.; O'Malley, M. K.; and Dragan, A. D. 2017. Learning robot objectives from physical human interaction. In *Conference on Robot Learning*, 217–226. PMLR.

Baker, C. L.; Tenenbaum, J. B.; and Saxe, R. R. 2007. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

Bıyık, E.; Losey, D. P.; Palan, M.; Landolfi, N. C.; Shevchuk, G.; and Sadigh, D. 2020. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *arXiv preprint arXiv:2006.14091*.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Brown, D.; Coleman, R.; Srinivasan, R.; and Niekum, S. 2020. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, 1165–1177. PMLR.

Brown, D.; and Niekum, S. 2018. Efficient probabilistic performance bounds for inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Chaloner, K.; and Verdinelli, I. 1995. Bayesian experimental design: A review. *Statistical Science*, 273–304.

Chan, L.; Critch, A.; and Dragan, A. 2021. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cui, Y.; and Niekum, S. 2018. Active reward learning from critiques. In *2018 IEEE international conference on robotics and automation (ICRA)*, 6907–6914. IEEE.

Do, A. M.; Rupert, A. V.; and Wolford, G. 2008. Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic bulletin & review*, 15(1): 96–98.

Evans, O.; Stuhlmüller, A.; and Goodman, N. D. 2016. Learning the Preferences of Ignorant, Inconsistent Agents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 323–329. AAAI Press.

Finn, C.; Levine, S.; and Abbeel, P. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, 49–58. PMLR.

Grüne-Yanoff, T. 2015. Models of temporal discounting 1937–2000: An interdisciplinary exchange between economics and psychology. *Science in context*, 28(4): 675–713.

Guan, S.; Cheng, L.; Fan, Y.; and Li, X. 2015. Myopic decisions under negative emotions correlate with altered time perception. *Frontiers in Psychology*, 6: 468.

Haarnoja, T.; Tang, H.; Abbeel, P.; and Levine, S. 2017. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, 1352–1361. PMLR.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2017a. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Hadfield-Menell, D.; Milli, S.; Abbeel, P.; Russell, S. J.; and Dragan, A. 2017b. Inverse reward design. *Advances in neural information processing systems*, 30.

Henry, P.; Vollmer, C.; Ferris, B.; and Fox, D. 2010. Learning to navigate through crowded environments. In *2010 IEEE International Conference on Robotics and Automation*, 981–986. IEEE.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human preferences and demonstrations in Atari. *arXiv preprint arXiv:1811.06521*.

Jeon, H. J.; Milli, S.; and Dragan, A. D. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *arXiv preprint arXiv:2002.04833*.

Jonnavittula, A.; and Losey, D. P. 2021. I know what you meant: Learning human objectives by (under) estimating their choice set. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2747–2753. IEEE.

Kalman, R. E. 1964. When is a linear control system optimal? *Transactions ASME, Journal Basic Engineering., 86:51–60,*.

Kitani, K. M.; Ziebart, B. D.; Bagnell, J. A.; and Hebert, M. 2012. Activity forecasting. In *European conference on computer vision*, 201–214. Springer.

Knox, W. B.; and Stone, P. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*, 9–16.

Kretzschmar, H.; Spies, M.; Sprunk, C.; and Burgard, W. 2016. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11): 1289–1307.

Liu, Y.; Datta, G.; Novoseller, E.; and Brown, D. S. 2023. Efficient Preference-Based Reinforcement Learning Using Learned Dynamics Models. In *International Conference on Robotics and Automation (ICRA)*. IEEE.

Mainprice, J.; Hayne, R.; and Berenson, D. 2015. Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 885–892. IEEE.

Matuszek, C.; FitzGerald, N.; Zettlemoyer, L.; Bo, L.; and Fox, D. 2012. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*.

Morgenstern, O.; and Von Neumann, J. 1953. *Theory of games and economic behavior*. Princeton university press.

Ng, A. Y.; Russell, S. J.; et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, 2.

Palan, M.; Shevchuk, G.; Charles Landolfi, N.; and Sadigh, D. 2019. Learning Reward Functions by Integrating Human Demonstrations and Preferences. In *Robotics: Science and Systems*.

Ramachandran, D.; and Amir, E. 2007. Bayesian Inverse Reinforcement Learning. In *IJCAI*, volume 7, 2586–2591.

Ratliff, N. D.; Bagnell, J. A.; and Zinkevich, M. A. 2006. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, 729–736.

Schrum, M. L.; Hedlund-Botti, E.; Moorman, N.; and Gombolay, M. C. 2022. Mind meld: Personalized meta-learning for robot-centric imitation learning. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 157–165. IEEE.

Shah, R.; Gundotra, N.; Abbeel, P.; and Dragan, A. 2019. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, 5670–5679. PMLR.

Sharot, T.; Riccardi, A. M.; Raio, C. M.; and Phelps, E. A. 2007. Neural mechanisms mediating optimism bias. *Nature*, 450(7166): 102–105.

Thompson, S. C. 1999. Illusions of control: How we overestimate our personal influence. *Current Directions in Psychological Science*, 8(6): 187–190.

Vasquez, D.; Okal, B.; and Arras, K. O. 2014. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1341–1346. IEEE.

Wirth, C.; Akrour, R.; Neumann, G.; Fürnkranz, J.; et al. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*.

Wulfmeier, M.; Ondruska, P.; and Posner, I. 2015. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; Dey, A. K.; et al. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, 1433–1438. Chicago, IL, USA.