Balancing Query Informativeness and Human Rationality in Reinforcement Learning from Human Feedback

Tsuheng Hsu
Aalto University
Espoo, Finland
tsu-heng.hsu@aalto.fi

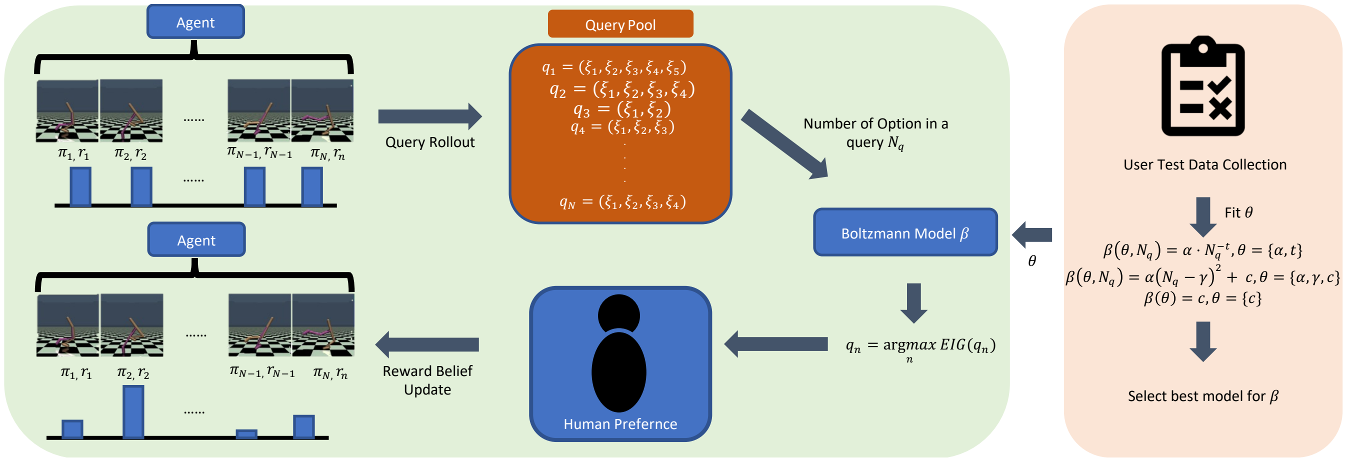

Agent
π_1, r_1
π_2, r_2
......
π_{N-1}, r_{N-1}
π_N, r_n
Query Rollout
Query Pool
$q_1 = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$
$q_2 = (\xi_1, \xi_2, \xi_3, \xi_4)$
$q_3 = (\xi_1, \xi_2)$
$q_4 = (\xi_1, \xi_2, \xi_3)$
$q_N = (\xi_1, \xi_2, \xi_3, \xi_4)$
Number of Option in a query N_q
Boltzmann Model β
θ
$q_n = \underset{n}{\mathrm{arg}max}\, EIG(q_n)$
Human Prefernce
Reward Belief Update
User Test Data Collection
Fit θ
$\beta(\theta, N_q) = \alpha \cdot N_q^{-t}, \theta = \{\alpha, t\}$
$\beta(\theta, N_q) = \alpha(N_q - \gamma)^2 + c, \theta = \{\alpha, \gamma, c\}$
$\beta(\theta) = c, \theta = \{c\}$
Select best model for β

Figure 1: Methodology of our study on query design in RLFH. (Left) An agent with prior knowledge of completing a task, e.g., walking, but without knowing human preference, generates rollouts forming a query pool, from which the queries are selected based on the EIG calculation, taking human rationality into account regarding the number of options in a query, and updates its belief in human preference. (Right) We perform a user study to capture human rationality regarding the number of options in a query, and fit, compare, and select the best hypothesis.

ABSTRACT

Reinforcement learning from human feedback (RLHF) often relies on preference queries, where humans guide a robot's behavior without hand-crafted reward functions. While pairwise comparisons are widely used due to their simplicity, they provide limited information per query. Queries with more options can, in theory, yield higher information, but also increase cognitive load and reduce decision reliability. In this work, we study how query informativeness interacts with bounded human rationality and how this trade-off influences reward preference learning. Through a small-scale user study, we model human choice behavior using a Boltzmann distribution and incorporate this model into an active learning framework, and showed that, when accounting for bounded human rationality, it enables more efficient learning under limited feedback, where short learning horizons make early stage learning particularly important.

1 INTRODUCTION

In standard RL, an explicit reward signal is required for exploration and exploitation, which is hard to define in a real-world environment or human-centered domain, as various factors have to be taken into account beyond the immediate reward signal, such as negative side effects, reward hacking, and safe exploration [1]. These challenges are pronounced in the domain of robotics due to the difficulty of hand-crafting reward functions that capture nuances of real-world interactions and the possibility of reward misspecification [19].

Learning from human feedback has been successfully widely applied in various domains such as large language models (LLMs) and robotics [31, 37, 41]. In particular, reinforcement learning from human feedback (RLHF) leverages preference signals for reward learning that guide reinforcement learning (RL) algorithms. RLFH provides a flexible way to align agent behavior with implicit goal or reward signals directly from human feedback, as providing preference judgments such as pairwise comparison [14] or ranking [10, 42] is more intuitive for humans than attempting to define reward signals by hand that encodes all complex intended behaviors. While RLFH has been successfully presented, human feedback is inherently limited and expensive, raising the problem of how to learn from it effectively.

Given the limited availability of human feedback, a core process of RLFH involves generating and selecting queries for the agent to learn human preferences efficiently. Query selection is typically guided by an *acquisition function*, which evaluates the informativeness of candidate queries based on the agent's current uncertainty. Such uncertainty is usually measured in the Bayesian setup, such as *volume removal (VR)* [47] or *expected information gain (EIG)* [22]. Such informativeness itself can be further interpreted in several dimensions, such as the number of options in a query or their complexity. For example, from an information-theoretic perspective, increasing the number of options generally increases entropy and thus the potential informativeness of a query.

However, human feedback is inevitably influenced by the inconsistencies, biases, and bounded rationality of humans, as well as other factors that can impact their preferences. Therefore, while Bayesian acquisition functions suggest that queries with more candidate options or higher complexity can provide richer information about human preferences, this will make the queries more challenging for humans to respond to. This effect is well-studied in psychology, such as Hick's law, which states that the response time of humans increases logarithmically with the number of available choices [21], which can further reduce the quality of feedback as discussed in [24].

Although various studies have explored different strategies [7, 8, 42, 50] for selecting the most informative queries that reduce human cognitive loads, most work is still limited to pairwise preference queries, where humans are asked to choose between only two candidate options. In contrast, other work has pursued richer preference modeling, moving beyond binary choices to listwise approaches [10, 11, 13, 38, 39], though these methods often still rely on pairwise comparisons to construct the underlying ranked lists. Therefore, in this study, we investigate the trade-off that arises when increasing informativeness within a query. While more informative queries promise richer signals for RLFH, they also introduce greater cognitive demands, which can affect the reliability of human feedback and, in turn, learning efficiency. In particular, we consider the number of options in a query as our approach to varying informativeness.

To address this, we adopt a Boltzmann rationality model [35] to capture stochastic choice behavior from humans. Specifically, we consider different assumptions for modeling the inverse temperature parameter β, which controls the level of human rationality, capturing the relation of the level of rationality with the number of options. Through a small-scale user study, we fit and compare these models against real human choice data to identify which best explains observed behavior. Finally, we integrate the fitted Boltzmann model into an active learning framework and conduct simulation experiments, showing how incorporating bounded rationality into query selection strategies changes reward preference learning dynamics compared to standard approaches.

The contributions of the paper:

- Modeling bounded human rationality as a function of query informativeness, with the number of options serving as a concrete instantiation.
- Incorporating the fitted rationality model into an active learning framework for reward preference learning using EIG.
- Empirical evaluation of the method in locomotion tasks, demonstrating its effectiveness under limited human feedback.

2 RELATED WORK

2.1 Reward Learning from Preferences and Active Query Selection

RLFH has been employed in robotics to learn complex behavior without handcrafting rewards. Unlike normal inverse reinforcement learning (IRL) [40, 46], which infers a reward function from expert demonstrations, RLFH approaches with human preference feedback. Early work like [30] learns to predict human reinforcement signals in a supervised setup, guiding the agent to choose actions that maximize predicted signals. Later work, such as [14, 23, 47], introduced a scalable standard framework for reward learning and training deep reinforcement learning agents directly from pairwise preference queries. Further work improves reward learning with a similar framework with meta-learning [20], pre-training [32], or learning reward features [10, 27].

A mutual problem in human preference-based learning is that human time is expensive; therefore, various methods and works have been explored in the active learning part of RLHF to improve efficiency [4, 16, 26, 34]. However, deep learning–based RLHF frameworks usually optimize a point estimate of the reward function, meaning that uncertainty is only implicit and difficult to compute directly. Since informativeness is fundamentally tied to reducing uncertainty about the underlying reward, this creates challenges for principled query selection.

In contrast to this, Bayesian approaches such as [45] instead learns a posterior distribution of reward parameters instead of a point estimate, providing a natural basis for uncertainty-aware query selection. Within this framework, later work adopts Bayesian acquisition functions such as VR [2, 3, 47] and EIG [7, 8, 33, 36] as criteria for selecting queries. Although VR performed well compared to earlier heuristics, [7, 8] showed that due to its non-convexity, it will lead to an informative local optima instead of a global one, suggesting that EIG often provides more reliable query selection compared to VR; therefore, this study will focus on EIG of queries.

2.2 Human Factor in Preference Learning

Most previous work assumes humans are completely rational when providing feedback; however, as studied in psychology and economics, human rationality can be affected by various factors, such as the number of options [21, 24] or the ordering [17]. Modeling human choice or preference behavior has always been a challenging field; the most common method is with a probability model, such as the Boltzmann rationality model [35] or the Bradley-Terry model [9]. The latter has been widely applied in the majority of work based on pairwise comparison preference learning (see, e.g. [14, 23, 32, 47]) without taking human rationality into account, while [12, 18, 28] pointed out the importance of modeling human rationality or uncertainty, even if modeling is approximated, it can still significantly improve reward learning under human irrationality, and will lead to worse outcomes if overestimate the rationality (e.g. complete ration). We show this by modeling human irrationality regarding the informativeness of queries, capturing how increased query informativeness reduces decision reliability, and examining the importance of taking this irrationality into account in preference learning.

2.3 From Human-Friendly to Informative Queries

To improve the efficiency of reward learning from human feedback under human irrationality, a line of work focused on improving the selection of queries, making the selected query informative and more human-friendly, and therefore reducing cognitive load and mitigating errors. [8] focused on fusing human uncertainty when encountering hard-to-answer queries into EIG calculation, enabling robots to ask easy queries for humans. [43] takes the effect of contextual changes in the querying process into account, grouping similar queries to minimize the mental effort of humans.

While these approaches focus on reducing cognitive load, easing the effect of overestimating human rationality, other lines of research aim to increase the information content of each query or to find a more efficient approach. [7, 42] integrate demonstration into human preference learning, building more informative queries based on prior knowledge learned from demonstration. [6] models the query selection in batches of 2-10 pairwise queries, allowing each query to be individually informative. [13] approach with a similar approach by learning a reward from a ranking list constructed by pairwise comparison from human feedback.

Taken together, these two directions highlighted a fundamental tension in preference-based reward learning from human feedback. Despite these efforts, the vast majority of approaches remain fundamentally based on pairwise queries, extending them with demonstrations, batching, or listwise aggregation. What is still unclear is how increasing the informativeness of queries interacts with bounded human rationality and how it further affects the efficiency of reward learning in limited human feedback. This raises our central research question: *How does human bounded rationality, as influenced by query information (number of options, complexity), affect the efficiency of reward preference learning in RLHF?*

3 PROBLEM SETUP

We consider an agent in a Markov decision process represented with the tuple: $\langle \mathcal{S}, \mathcal{A}, T, r \rangle$ where $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively, and T is the transition function such that when action $a_t \in \mathcal{A}$ is taken at $s_t \in \mathcal{S}$ at time step t, the next state is $s_{t+1} \sim T(\cdot \mid s_t, a_t)$. The reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ maps state action at timestep t to a reward value, R denotes the cumulative reward over a trajectory ξ: $R(\xi) = \sum_{(s,a)\in\xi} r(s,a)$.

In our setting, the underlying true reward function is not available. Instead, we assume access to human feedback in the form of preference labels. Similar to prior works [39], given the desired task and one or several robot trajectories human selects the most similar one to the task. Then the goal objective of the work is to learn the reward function that best aligns with collected human preferences.

4 METHOD

4.1 Reward Learning From Human Feedback

Unlike most prior work, which adopts pairwise comparison feedback, we model the human feedback in a selection from a multiple-option query. Given a query pool of $Q = \{q_1 \dots q_N\}$, where each query $q_i = \{\xi_1 \dots \xi_K\}$ for $q_i \in Q$, the human prefers a trajectory with the highest cumulative rewards under the target reward $R^*(\xi) = \sum_{(s,a)\in\xi} r^*(s,a)$,

$$\xi_k \succ \xi_j \iff R^*(\xi_k) \geq R^*(\xi_j), \quad \forall j \neq k, \xi_j \in q_i. \tag{1}$$

However, as discussed, human choices are not always rational; Equation (1) thus corresponds to the oracle case of an entirely rational annotator. Therefore, we can see it as a special case of reward-rational choice [25], where human choices can be modeled as:

$$p(\xi_k \mid q_i, \beta, r) = \frac{exp\left(\beta R(\xi_k; r)\right)}{\sum_{j=1}^{K} exp(\beta R(\xi_j; r))}, \quad r \sim p(r), \tag{2}$$

where β expresses how close the human is to complete rationality ($\beta = 0$ corresponds to random choice and $\beta \rightarrow +\infty$ indicates complete rational choice).

The task of reward learning is then to infer a posterior distribution over the true reward function with human feedback. Formally, after observing that the human selects ξ_k from query q_i, we update our belief via Bayes' rule:

$$p(r \mid \xi_k, q_i, \beta) \;\propto\; p(\xi_k \mid q_i, \beta, r)\; p(r). \tag{3}$$

As more correct feedback is provided, the posterior mass will concentrate around the true reward.

4.2 Human Rationality

The β in the Boltzmann rationality model decides the rationality level of the human; to capture the relation of it with the number of options in a query, we model the β with three hypotheses based on the study from [21, 24], which as number of options increases in a query, the probability mass of the rational choice will be shared across the options:

(1) Exponetial: $\beta(\theta, N_{qi}) = \alpha \cdot N_{qi}^{-t}, \quad \theta = \{\alpha, t\}$
(2) Quadratic: $\beta(\theta, N_{qi}) = \alpha(N_{qi} - \gamma)^2 + c, \quad \theta = \{\alpha, \gamma, c\}$
(3) Constant: $\beta(\theta) = c, \quad \theta = \{c\}$

where N_{qi} indicates the number of options in a specific query. Then, with a given human feedback dataset $D = \{(q_1, \xi_{k1}) \dots (q_N, \xi_{kN})\}$, where $\xi_{ki}, i \in [1, N]$ represent the response of human prefering option k in query n, the goal is to find the parameters that maximize the likelihood of the data for the following equation under a known reward r:

$$\prod_{i=1}^{N} p(\xi_{ki} | q_i, \theta, N_{qi}, r) = \prod_{i=1}^{N} \frac{exp\left(\beta(\theta, N_{qi}) R(\xi_{ki}; r)\right)}{\sum_{j=1}^{N_{qi}} exp(\beta(\theta, N_q qi) R(\xi_{ki}; r))} \tag{4}$$

4.3 Query Selection

During query selection, we would want to select the queries that will reduce the uncertainty the most after feedback is given. This can be achieved by selecting queries with the highest *information gain* by solving:

$$\begin{aligned} q_i^* &= \arg\max_{q_i} I(r; \xi_{ki} \mid q_i, \theta, N_{qi}) \\ &= \arg\max_{q_i} \left[H(r) - H(r \mid \xi_{ki}, q_i, \theta, N_{qi})\right] \end{aligned} \tag{5}$$

where I is the mutual information and H is the Shannon's information entropy [15]. Since the actual feedback ξ_{ki} is not known during query selection, we instead maximize the EIG[5] by using the marginal predictive distribution of the feedback ξ_{ki}:

$$q_i^* = \arg\max_{q_i} \mathbb{E}_{p(\xi_{ki} | q_i, \theta, N_q)} \left[I(r; \xi_{ki} \mid q_i, \theta, N_{qi})\right] \tag{6}$$

$$= \arg\max_{q_i} \mathrm{EIG}_{p(r)}(q_i) \tag{7}$$

In the previous section, we defined the problem of reward learning from multi-option human feedback in a Bayesian framework. We now describe how these formulations are instantiated in practice. Specifically, we explain how we construct the reward prior, perform posterior updates with Boltzmann likelihoods, and approximate expected information gain for query selection.

4.4 Human Rationality Fitting

The rationality parameter β is modeled as a function of query size N_q according to the hypotheses introduced in Section 4.2. We estimate the corresponding parameter by maximizing likelihood estimation (MLE) using the observed human response dataset $D = \{(q_1, \xi_{k1}) \dots (q_N, \xi_{kN})\}$:

$$\text{MLE}(\hat{\theta}) = \arg\max_{\theta} \sum_{i=1}^{N} \log p(\xi_{ki} | q_i, \theta, N_q, r). \tag{8}$$

We optimize the parameters with the Adam optimizer [29]. The dataset collection and user study setup are detailed in Section 5.1.

4.5 EIG Calculation

From Equation (7), we use a Rao-Blackwellized estimator [44] to estimate the EIG, obtaining the following formula for a given query q as:

$$\widehat{\text{EIG}}(q) = \sum_{k=1}^{N_q} \left[\frac{1}{S} \sum_{s=1}^{S} p(\xi_k \mid q, \theta, N_q, r^{(s)}) \, \log p(\xi_k \mid q, \theta, N_q, r^{(s)}) \right.$$
$$\left. - \, \hat{p}(\xi_k \mid q, \theta, N_q) \, \log \hat{p}(\xi_k \mid q, \theta, N_q) \right],$$

where

$$\hat{p}(\xi_k \mid q, \theta, N_q) = \frac{1}{S} \sum_{s=1}^{S} p(\xi_k \mid q, \theta, N_q, r^{(s)}), \quad r^{(s)} \sim p(r).$$

Among all candidate queries, the one with the highest estimated EIG is prioritized for selection.

4.6 Reward Learning

In our setup, the agent knows how to complete the task through a population of candidate policies, each of which has a different behavior trained on a population of rewards. We set the prior knowledge of the agent to the rewards' belief as a uniform Dirichlet distribution. In practice, we hand-craft a finite set of reward functions to represent this prior; the resulting population of policies reflects different styles of completing the task.

Following the procedure outlined in Algorithm 1, we maintain a posterior distribution belief over reward hypotheses, updated from the uniform Dirichlet prior. At each iteration, the agent selects the top-N queries ranked by chosen acquisition criterion (e.g., EIG, VR, or random). After receiving feedback, the posterior belief is updated with a particle-based approximation of Bayes' rule with a Boltzmann likelihood, whose temperature β is as defined in Section 4.2. The acquisition scores of the remaining queries are then recalculated with the updated belief, and the query pool is re-ranked for the next iteration.

5 EXPERIMENT

To capture the relation between query informativeness, modeled by the number of options in a query, and human rationality, we conduct a pilot user study for data collection and hypothesis fitting. Then, we show the effect of this relation to reward learning by performing simulation tests with MuJuCo environments [49], with different query selection strategies.

Algorithm 1 Posterior Sampling with EIG-driven Queries

```
Initialize reward function population $\{r_j\}_{j=1}^{J}$
Initialize prior belief $\text{Belief}^{(0)} \sim \text{Dirichlet}(\mathbf{1}_J)$
Initialize query pool $Q_{EIG} \leftarrow \{(q_i, EIG(q_i \mid \text{Belief}^{(0)}))\}$
Initialize log-likelihoods $\ell_m \leftarrow 0 \quad \forall m = 1, \dots, M$
Initialize samples $Prob = [Prob_m]_{m=1}^{M} \sim \text{Dirichlet}(\mathbf{1}_J)$
$t \leftarrow 0, N \leftarrow$ number of top queries per iteration
while $t < |Q|$ do
    for top $N$ $q_i \in Q_{EIG}$ ranked by $EIG(q_i)$ do
        Construct reward matrix $R(q_i) = \left[r_j(q_i)\right]_{j=1}^{J} \in \mathbb{R}^{J \times K}$
        Normalize $R(q_i)$ across $j$
        $\hat{r}_m \leftarrow R(q_i)^\top \cdot (Prob_m)^\top \quad \forall m$
        $\ell_m$ += $\log \text{Boltzmann}(\xi_{ki} \mid q_i, \theta, N_{q_i}, \hat{r}_m)$
        $t \leftarrow t + 1, \quad$ remove $q_i$ from $Q$
    end for
    Update weights $w_m \leftarrow \exp(\ell_m - \log \sum_s \exp(\ell_s))$
    Resample $update^{(t)}$ from $\{Prob_m\}$ according to $w_m / \sum_s w_s$
    Update belief $\text{Belief}^{(t)} \leftarrow \frac{1}{M} \sum_{m=1}^{M} update^{(t)}$
    Reforming $Q_{EIG}$ for remaining $q_i \in Q_{EIG}$ using $\text{Belief}^{(t)}$
end while
Output: Learned belief $\text{Belief}^{(|Q|)}$
```

5.1 Human Rationality User Study

To obtain empirical data for the rationality parameter β fitting, we conducted a pilot user study. The participants are presented with 8 preference queries, each containing between two and five all-lowercase word phrases. The distribution of queries was balanced with an equal number of queries for each option size.

The task was to select the option with the highest repeated-letter count under a 90-second time limit. For example, if a query is given as:

(1) bayesian experimental design
(2) active learning
(3) cognitive model
(4) expected information gain
(5) reinforcement learning

option 1's maximum repeated letter is ’e’ with 5 repeats; option 2 have three letters that have the same maximum count: ’e’, ’i’, and ’a’, with 2 repeats, since we are asking only for the maximum count of repeated letter, considering the count itself will be enough; option 3 have maximum count of 2 repeated letters (’e’ or ’i’); option 4 have 4 repeated letter ’e’. Therefore, the participant should report option 1. We collected data from 5 subjects, each with an independent set of queries, resulting in a total of 40 responses.

These responses form the dataset used to fit the rationality model parameters via MLE, with the reward function to be the maximum repeated-letter count for each option.

5.2 Reward Learning Simulation Test

We conduct the test on 3 MuJuCo environments, Walker2D, Hopper, and HalfCheetah, where the agent's goal is locomotion by moving forward.

Reward Prior and Policy Population. For each environment, we construct a population of $N = 50$ policies π_{ϕ^k}, each trained for 6 million steps with policy proximal optimization (PPO) [48] on a distinct reward which are hand-crafted by varying parameters $\phi^k, k = 1 \dots 50$ that encode specific behavioral styles. For Walker2D and Hopper, the parameter corresponds to the angle of the torso while moving forward, and the front thigh angle for Halfcheetah. We assume that the agent has no prior knowledge of which behavior is preferred. The prior over reward hypotheses is thus initialized as a uniform Dirichlet distribution:

$$p(r) = \text{Dirichlet}(\underbrace{1, \dots, 1}_{50}).$$

Query Generation. Each policy π_{ϕ^k}, for $k = 1, \dots, 50$, generates a set of $M = 10$ trajectory rollouts with 300 timestep,

$$\Xi^{(k)} = \{\xi_m^{(k)} \sim \pi_{\phi^k}\}_{m=1}^{10},$$

where $\xi_m^{(k)}$ denotes the m-th trajectory sampled from the k-th policy, resulting in 500 trajectories, serving as the basis of the queries. For each environment, we randomly generate 140 queries from their basis, each containing a uniform number of two to five trajectories, allowing us to study the effect of different query types. This constructs the pool of candidate queries for the learning phase.

Reward Learning. During the reward learning process, it is always based on our fitted human rationality model but with different query selection method: (i) using our fitted-rationality calculated EIG criterion, (ii) EIG criterion calculated assuming perfect rationality (constant $\beta = 5.74$, corresponding to the exponential model at 2-option queries, reflecting the assumption that humans act near-perfectly rational in pairwise choices), (iii) random pick, and (iv) VR under our fitted-rationality, which is calculated as:

$$VR(q) = 1 - \sum_{k=1}^{N_q} \left[\frac{1}{S} \sum_{s=1}^{S} p(\xi_k \mid q, \theta, N_q, r^{(s)}) \right], r^{(s)} \sim p(r).$$

Human feedback is simulated using a held-out target reward function, corresponding to the intended behavioral style (e.g., torso at −25°).

Posterior updates over reward hypotheses are performed using Bayes' rule with Boltzmann likelihoods. The agent's objective is to concentrate posterior mass on the reward function aligned with the target behavior.

Evaluation. To compare the reward learning under different query selection methods, we track the posterior distribution of the reward belief for each query update. For the ground truth reward, we use the mean of the ten trajectory rollouts $\Xi^{(k^*)}$ from the policy trained on the target parameter ϕ^{k^*}. This ground truth trajectory is passed into the reward population, yielding a distribution of rewards on the ground truth trajectory. Under each update's posterior, we calculate the expected reward as:

$$\text{Expected Reward} = \sum_{i=1}^{50} p_i r_i$$

where $p_i \in \text{Belief}^{(t)}, \quad r_i \in \left[r_j(\bar{\xi}^{(k^*)})\right]_{j=1}^{J}, \quad \bar{\xi}^{(k')} = \frac{1}{M} \sum_{m=1}^{M} \xi_m^{(k^*)}$, and finally normalize the values across query methods.

6 RESULT

In this section, we first present the results of our human rationality study, where we model the relationship between query informativeness and decision reliability. We then evaluate how incorporating this model into an active learning setup affects reward learning dynamics. Finally, we analyze the role of different query types and their contribution to efficient learning under bounded rationality.

6.1 Human Rationality Study

The result of the human rationality study showed that among 40 queries, the wrong choices made by the subjects concentrated at four and five option queries, with two and three wrong choices, respectively, and with all other types of queries responded correctly. This result matches previous studies in psychology [21, 24] where the number of options in a decision task increases cognitive load and reduces the decision reliability, implying some implicit relation between query informativeness and human rationality.

Model	**Average Likelihood**
Uniform	0.3021
Exponential ($\beta = \alpha N_{qi}^{-t}$)	**0.6468**
Quadratic ($\beta = \alpha(N_{qi} - \gamma)^2 + c$)	0.6391
Constant β	0.4350
Fixed $\beta = 1$ (softmax)	0.5582

Table 1: Average likelihood of different models fitted to user choice data.

Then, we capture this relation by fitting our human rationality models in Section 4.2 with the result data, and evaluate with the average likelihood of the data with respect to the model as shown in Table 1. From which we can see that both the quadratic and exponential models capture the data better compared to pure softmax or fixed constant, indicating that human rationality is not constant but decreases as the number of options increases in a query.

The result confirms that when query informativeness is modeled by the number of options, higher informativeness comes at the cost of reduced human rationality, indicating a trade-off shaping the learning dynamics. To show how this trade-off influences reward learning in practice, we next show the evaluation of reward learning dynamics in simulation under different query selection strategies.

6.2 Simulation Test

In the simulation test, we evaluate how different query selection strategies perform when reward learning is carried out under our fitted human rationality model. Specifically, we compare EIG with

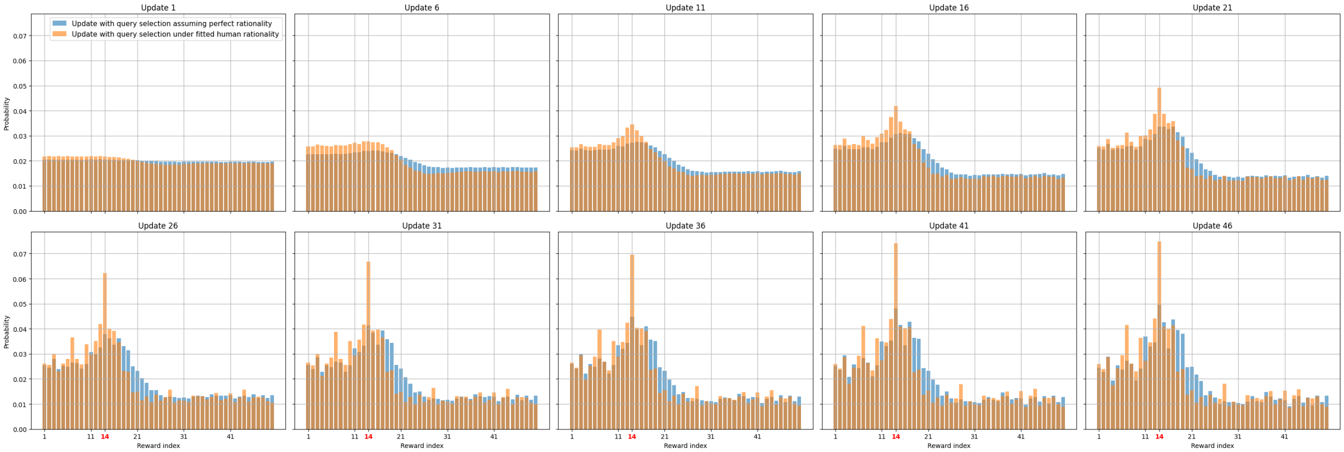


Figure 2: Agent's belief distribution among rewards during learning with Walker2d, where the highlighted reward index is the ground truth.

	Iteration							
Method	1	5	10	20	30	40	80	140
EIG with fitted rationality	**0.0651**	**0.3508**	**0.4707**	**0.6889**	**0.8086**	**0.8533**	0.9267	0.9993
EIG assuming perfect rationality	0.0000	0.1776	0.3465	0.5793	0.7356	0.8167	0.9620	0.9993
VR with fitted rationality	0.0095	0.1649	0.2934	0.5212	0.7058	0.7920	0.9747	0.9993
Random	0.0001	0.3121	0.4180	0.6481	0.7374	0.8027	0.9304	0.9993

(a) Walker2d

	Iteration							
Method	1	5	10	20	30	40	80	140
EIG with fitted rationality	**0.0507**	**0.1948**	**0.3374**	**0.4998**	**0.6087**	0.6727	0.8643	0.9997
EIG assuming perfect rationality	0.0000	0.1079	0.2166	0.3891	0.5374	0.6837	0.9418	0.9997
VR with fitted rationality	0.0032	0.1066	0.2059	0.3660	0.5030	0.6489	0.9412	0.9997
Random	0.0165	0.1926	0.2438	0.3574	0.5170	0.5954	0.8981	0.9997

(b) HalfCheetah

	Iteration							
Method	1	5	10	20	30	40	80	140
EIG with fitted rationality	**0.0494**	**0.3012**	**0.4334**	**0.6546**	**0.7484**	0.7748	0.9147	1.0000
EIG assuming perfect rationality	0.0084	0.1288	0.2721	0.5028	0.6858	0.8151	0.9820	1.0000
VR with fitted rationality	0.0000	0.1478	0.2814	0.4464	0.5999	0.7406	0.9660	1.0000
Random	0.0215	0.2033	0.3726	0.5635	0.6827	0.7954	0.9504	1.0000

(c) Hopper

Table 2: Normalized cumulative rewards at selected update steps for different strategies across environments.

fitted rationality, EIG assuming perfect rationality, VR with fitted rationality, and random selection.

As shown in Table 2, we observe that accounting for human rationality via the fitted Boltzmann model yields faster learning in a limited feedback horizon, where selecting the most informative queries from the pool under human rationality early on is crucial. As informative queries are exhausted in the pool, the rate of learning slows down and its advantage diminishes, while all method converges to the same reward levels due to the fixed query pool, this can also be visualized through the belief posterior (Figure 2) and the reward dynamic plot (Figure 3) for the Walker2d environment. Results for the other environments, which exhibit the same trend, are provided in Appendix A.

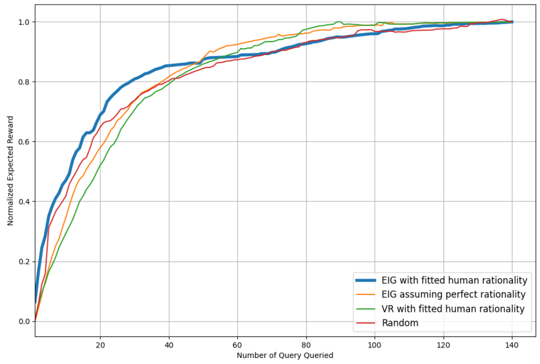


Figure 3: Normalized cumulative reward under the agent's belief upon updating accounting human rationality with different query selection methods.

The results indicate that, despite the high informativeness of the query, if human rationality is not considered during query selection, the irrationality of human decisions will affect the learning efficiency within a limited human feedback horizon.

6.3 Impact of Query Informativeness on Learning Dynamics

To further understand how query informativeness under bounded rationality affects the efficiency of reward learning, we also evaluate the learning with only one type of query (2-5 options) in the same query pool. As shown in Figure 4 for the Walker2d environment, pairwise queries yield the fastest initial progress but plateau early due to limited information, while higher-option queries are slower initially but contribute more once sufficient feedback accumulates. The mixed strategy benefits by leveraging these shifting strengths across query types, highlighting the impact of query informativeness on learning dynamics, from which the efficiency arises from accounting for bounded rationality and adaptively balancing query informativeness during query selection.

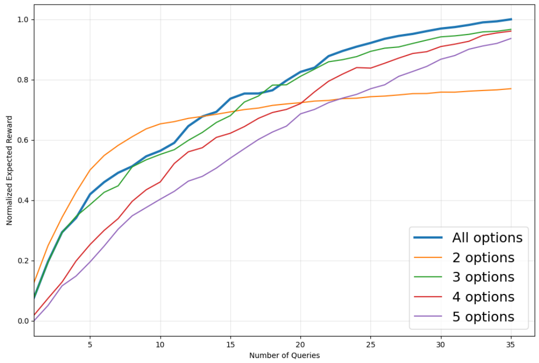


Figure 4: Comparison of different types of query updates independently in the query pool, along with the whole query pool updates with Walker2d.

7 CONCLUSION

In this study, we studied the relation between query informativeness and human rationality by modeling query informativeness with the number of options in a query. We showed how this relation affects learning dynamics in a Bayesian active learning setup, with efficiency gains arising when taking bounded human rationality into account under a limited human feedback horizon.

These findings provide insight into query design for RLHF, suggesting that, beyond pairwise comparisons or other canonical formats, combining queries with different levels of informativeness can be effective. When human rationality is correctly modeled and incorporated into learning, it boosts efficiency under limited feedback while ensuring the reliability of human responses.

However, this study models informativeness only through the number of options in a query, while in practice, various factor that affects a query's informativeness. Moreover, the efficiency observed in the simulation tests depends strongly on the assumed human rationality model, making more representative modeling critical. These limitations point to future work directions, including a more comprehensive modeling of query informativeness, a more detailed human rationality study and modeling, and a real human in the loop test for validation.

ACKNOWLEDGMENTS

This work was conducted during an internship in Intelligent Robotic Labs at Aalto University. I would like to express my sincere gratitude to the lab for hosting and supporting my research. I am especially thankful to Baimukashev Daulet and Oskar Keurulainen for their invaluable supervision, guidance, and encouragement, which made this work possible.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[2] Chandrayee Basu, Erdem Bıyık, Zhixun He, Mukesh Singhal, and Dorsa Sadigh. 2019. Active learning of reward dynamics from hierarchical queries. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 120–127.

[3] Chandrayee Basu, Mukesh Singhal, and Anca D Dragan. 2018. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 132–140.

[4] Syrine Belakaria, Joshua Kazdan, Charles Marx, Chris Cundy, Willie Neiswanger, Sanmi Koyejo, Barbara E Engelhardt, and Stefano Ermon. 2025. Sharpe Ratio-Guided Active Learning for Preference Optimization in RLHF. *arXiv preprint arXiv:2503.22137* (2025).

[5] José M Bernardo. 1979. Expected information as expected utility. *the Annals of Statistics* (1979), 686–690.

[6] Erdem Biyik, Nima Anari, and Dorsa Sadigh. 2024. Batch active learning of reward functions from human preferences. *ACM Transactions on Human-Robot Interaction* 13, 2 (2024), 1–27.

[7] Erdem Bıyık, Dylan P Losey, Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. 2022. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research* 41, 1 (2022), 45–67.

[8] Erdem Bıyık, Malayandi Palan, Nicholas C Landolfi, Dylan P Losey, and Dorsa Sadigh. 2019. Asking easy questions: A user-friendly approach to active reward learning. *arXiv preprint arXiv:1910.04365* (2019).

[9] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[10] Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. 2020. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*. PMLR, 1165–1177.

[11] Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*. PMLR, 783–792.

[12] Lawrence Chan, Andrew Critch, and Anca Dragan. 2021. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956* (2021).

[13] Heewoong Choi, Sangwon Jung, Hongjoon Ahn, and Taesup Moon. 2024. Listwise reward estimation for offline preference-based reinforcement learning. *arXiv preprint arXiv:2408.04190* (2024).

[14] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[15] Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

[16] Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. 2024. Active preference optimization for sample efficient rlhf. *arXiv preprint arXiv:2402.10500* (2024).

[17] Brett Day, Ian J Bateman, Richard T Carson, Diane Dupont, Jordan J Louviere, Sanae Morimoto, Riccardo Scarpa, and Paul Wang. 2012. Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of environmental economics and management* 63, 1 (2012), 73–91.

[18] Gaurav R Ghosal, Matthew Zurek, Daniel S Brown, and Anca D Dragan. 2023. The effect of modeling human rationality level on learning rewards from multiple feedback types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5983–5992.

[19] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

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

[21] William E Hick. 1952. On the rate of gain of information. *Quarterly Journal of experimental psychology* 4, 1 (1952), 11–26.

[22] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745* (2011).

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

[24] Sheena S Iyengar and Mark R Lepper. 2000. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology* 79, 6 (2000), 995.

[25] Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems* 33 (2020), 4415–4426.

[26] Kaixuan Ji, Jiafan He, and Quanquan Gu. 2024. Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:2402.09401* (2024).

[27] SM Katz, A Maleki, E Bıyık, and MJ Kochenderfer. [n. d.]. Preference-based learning of reward function features. arXiv 2021. *arXiv preprint arXiv:2103.02727* ([n. d.]).

[28] Vijay Keswani, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, and Walter Sinnott-Armstrong. 2024. On the Pros and Cons of Active Learning for Moral Preference Elicitation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 711–723.

[29] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] https://arxiv.org/abs/1412.6980

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

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

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

[33] David Lindner, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, and Andreas Krause. 2021. Information directed reward learning for reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 3850–3862.

[34] Pangpang Liu, Chengchun Shi, and Will Wei Sun. 2024. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504* (2024).

[35] Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior. (1972).

[36] Sören Mindermann, Rohin Shah, Adam Gleave, and Dylan Hadfield-Menell. 2018. Active inverse reward design. *arXiv preprint arXiv:1809.03060* (2018).

[37] Ithan Moreira, Javier Rivas, Francisco Cruz, Richard Dazeley, Angel Ayala, and Bruno Fernandes. 2020. Deep reinforcement learning with interactive feedback in a human–robot environment. *Applied Sciences* 10, 16 (2020), 5574.

[38] Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. 2022. Learning multimodal rewards from rankings. In *Conference on robot learning*. PMLR, 342–352.

[39] Vivek Myers, Erdem Bıyık, and Dorsa Sadigh. 2023. Active reward learning from online preferences. *arXiv preprint arXiv:2302.13507* (2023).

[40] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[41] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[42] Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. 2019. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928* (2019).

[43] Mattia Racca, Antti Oulasvirta, and Ville Kyrki. 2019. Teacher-aware active robot learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 335–343.

[44] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. 2024. Modern Bayesian experimental design. *Statist. Sci.* 39, 1 (2024), 100–114.

[45] Deepak Ramachandran and Eyal Amir. 2007. Bayesian Inverse Reinforcement Learning.. In *IJCAI*, Vol. 7. 2586–2591.

[46] Stuart Russell. 1998. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*. 101–103.

[47] Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. 2017. Active preference-based learning of reward functions. (2017).

[48] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[49] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 5026–5033.

[50] David Zhang, Micah Carroll, Andreea Bobu, and Anca Dragan. 2022. Time-efficient reward learning via visually assisted cluster ranking. *arXiv preprint arXiv:2212.00169* (2022).

A REWARD LEARNING DYNAMIC RESULT FOR OTHER ENVIRONMENT

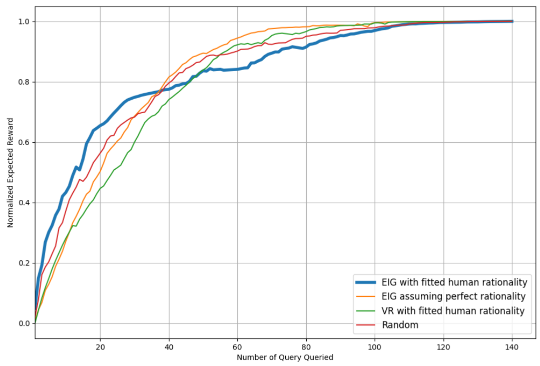


Figure 5: Reward learning dynamics of Hopper environment.

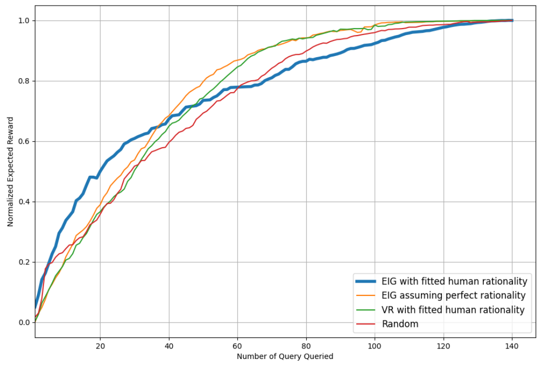


Figure 6: Reward learning dynamics of HalfCheeta environment.

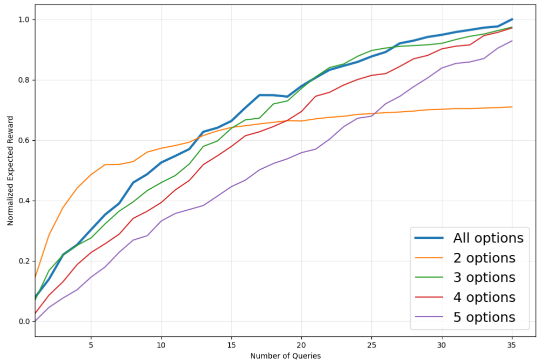
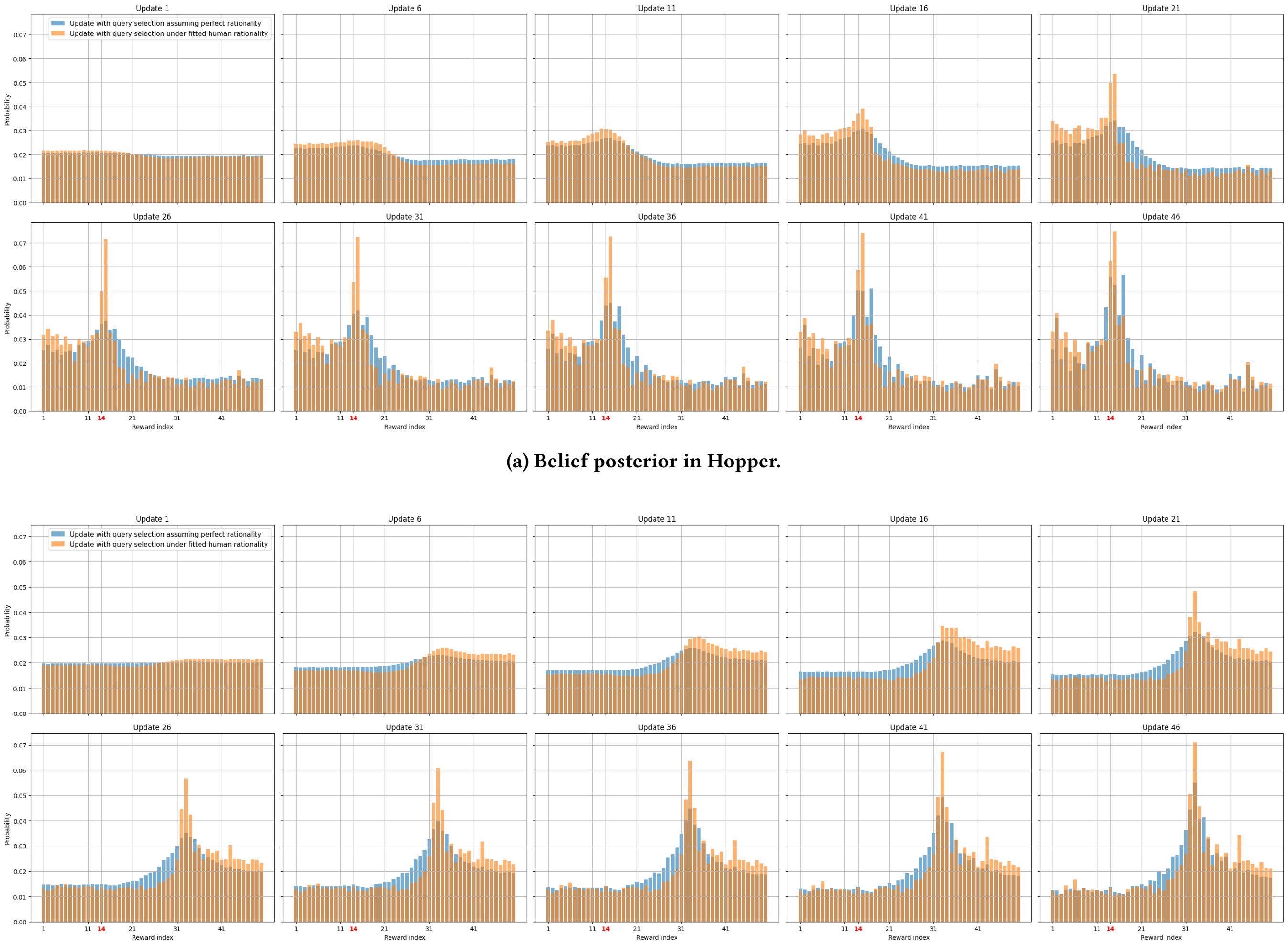


Figure 7: Reward learning dynamics of HalfCheeta environment.



(a) Belief posterior in Hopper.

(b) Belief posterior in HalfCheeta.

Figure 9: Belief posterior updates across environments

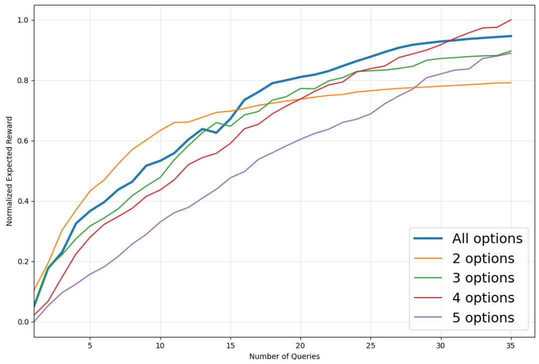


Figure 8: Reward learning dynamics of Hopper environment.