# 1 Appendix

## 1.1 Intuitively Understanding Individual Rewards as Individual Costs

**Task Reward** represents the task objective. **Human Individual Reward** represents the human partner's collaborative preferences. And **Robot Individual Reward** represents the robot's soft constraints. The individual rewards are based on the actions each agent individually takes. We can interpret these as negative costs, where an action with lower rewards is straining on the agent performing it. In the dishwasher loading example, picking up an object, like a bowl or glass cup, are determined to be especially costly for the robot because of the difficulty in picking up objects with smooth surfaces with a two finger gripper, and the risk of the dropping and shattering the object. These high cost actions involved in picking up those objects would give low **Robot Individual Reward**.

## 1.2 Bayesian Inverse Reinforcement Learning: Implementation Details

We model the human as a noisily rational actor who selects actions based on the assumption that the robot will take the optimal complementary action to maximize the team's reward. The human implicitly sets expectations for the robot by assuming the robot will take the optimal action. For example, the predicted robot action is optimal under $R_{\theta^T,\theta^r,\theta^h}$, which the human has full knowledge of. By understanding these expectations, the robot is able to deduce which human individual reward function $R_{\theta^h}$ would lead to such expectations. We also adopt the Boltzmann rational model of human behavior [1, 2], which models that human decisions are exponentially likely with respect to reward. According to this model, the probability of choosing a particular option increases exponentially as its utility, or reward, increases compared to other available options.

Over $t$ timesteps, the robot observes the human perform some trajectory of actions, $\tau_t = \{(s_1, a_1^h, s_2), (s_2, a_2^h, s_3), ..., (s_t, a_t^h, s_{t+1})\}$. Since the human has full knowledge of the composite reward, we assume the human's policy is a stationary expert policy. Thus, we can make the following independence assumption for all candidate human reward functions $\hat{\theta}^h \in \Theta^h$ (Equation 1).

$$P(\hat{\theta}^h|\tau_t) \propto P((s_1, a_1^h)|\hat{\theta}^h)P((s_2, a_2^h)|\hat{\theta}^h)...P((s_t, a_t^h)|\hat{\theta}^h)P(\hat{\theta}^h) \tag{1}$$

We intuit that since the human has full knowledge of all three reward functions: task, human individual, and robot individual reward, and in absence of a model of the robot's behavior, the human will try to achieve the task objective while acting pedagogically by taking actions that seek to demonstrate a complementary action the human expected the robot to have performed. This best-case reward, $r_{bc}$, is the reward the team would receive from the composite function $R_{\theta^T,\theta^r,\hat{\theta}^h}$ if the robot had taken the action $a^r \in \mathcal{A}^r$ that would maximize the composite reward (Equation **??**). The composite function under candidate $\hat{\theta}^h$ uses $\hat{\theta}^h$ to parameterize the human individual reward term.

$$r_{bc}(s_t, a_t^h|\hat{\theta}^h) = \max_{a^r \in \mathcal{A}^r} \sum_{s_{t+1}} R_{\theta^T,\theta^r,\hat{\theta}^h}(s_t, a_t^h, s_{t+1})T(s_{t+1}|s, a_t^h, a_t^r)$$

$$P((s_t, a_t^h)|\hat{\theta}^h) \propto \exp(\beta \cdot r_{bc}(s_t, a_t^h|\hat{\theta}^h))$$

Consider a state in which the human has two actions to choose from $a$, and $b$. The robot's evaluates its best-case reward for two hypotheses: (1) best-case reward 1 to $a$, 3 to $b$, and (2) $r_{bc} = 2$ to $a$ and 3 to $b$. The human chooses action $b$; however, we do not want to necessarily update based on a distribution derived from the reward values themselves. Since our robot models the human is a reward maximizer, the human views no effective differences between the two strategies. Thus, we want to ensure the probabilities reflect equal probability for selecting $b$ by thresholding the Boltzmann potential if the action yields a best-case reward equal to the maximum.

$$r(s_t, a_t^h|\hat{\theta}^h) = \begin{cases} \lambda, & \text{if } r_{bc}(s_t, a_t^h|\hat{\theta}^h) = \max_{a^h \in \mathcal{A}^h} r_{bc}(s_t, a^h|\hat{\theta}^h) \\ 1-\lambda, & \text{otherwise} \end{cases}$$

**Algorithm 1** (BIRL) Bayesian Inverse Reinforcement Learning: Online Update

**Input**: State $s_t$, Human Action $a_t^h$, Belief prior $b_0(\hat{\theta}^h)\ \forall \hat{\theta}^h \in \Theta^h$
**Parameter**: Rationality threshold $\lambda$, Hypothesis space $\Theta^h$, Temperature $\beta$
**Output**: Updated beliefs $b$, Predicted human policy $\hat{\pi}^h$

1: **for** $\hat{\theta}^h \in \Theta^h$ **do**
2: $\quad r_{bc}(s_t, a_t^h|\hat{\theta}^h) = \max_{a^r \in \mathcal{A}^r} \sum_{s_{t+1}} R_{\theta^T,\theta^r,\hat{\theta}^h}(s_t, a_t^h, s_{t+1}) T(s_{t+1}|s, a_t^h, a_t^r)$
3: $\quad r(s_t, a_t^h|\hat{\theta}^h) = \begin{cases} \lambda, & \text{if } r_{bc}(s_t, a_t^h|\hat{\theta}^h) = \max_{a^h \in \mathcal{A}^h} r_{bc}(s_t, a^h|\hat{\theta}^h) \\ 1-\lambda, & \text{otherwise} \end{cases}$
4: $\quad Z_t = \sum_{\theta \in \Theta^h} e^{\beta \cdot r(s_t, a_t^h|\theta)}$
5: $\quad P((s_t, a_t^h)|\hat{\theta}^h) = \frac{1}{Z_t} e^{\beta \cdot r(s_t, a_t^h|\hat{\theta}^h)}$
6: $\quad b(\hat{\theta}^h) \leftarrow P(\hat{\theta}^h|s_t, a_t^h) = \frac{P((s_t, a_t^h)|\hat{\theta}^h)) b_0(\hat{\theta}^h)}{P(s_t, a_t^h)}$
7: **end for**
8: $\hat{\pi}^h(a^h|s; \hat{\theta}^h) \propto e^{\beta \cdot r(s_t, a_t^h|\hat{\theta}^h)}$
9: $\hat{\pi}^h(a^h|s) = \sum_{\hat{\theta}^h} b(\hat{\theta}^h) \hat{\pi}^h(a^h|s; \hat{\theta}^h)$

Algorithm 1 below describes the Bayesian inverse reinforcement learning approach. The predicted human policy is the expected policy under the updated beliefs of the robot. We use $\lambda = 0.9$ in our agents for all simulated and human study experiments.

### 1.2.1 Optimistic Information Gain: Implementation Details

In updating its beliefs, the robot assumes that the human will optimistically choose the action that would achieve maximum reward given that the robot takes the ideal action which would facilitate achieving the maximum reward. The robot updates its beliefs $b$ over the possible values of the human's individual reward, $\theta^h$ using a likelihood function $P((s_t, a_t^h)|\hat{\theta}^h)$ built on this assumption. [3] examines a solution to the CIRL problem in which the human teacher is expected to act pedaogically, while the robot learner, aware and expecting this pedagogy, acts practically under its learned beliefs. Research on human pedagogical reasoning demonstrates that when teaching, humans engage in actions aimed at influencing or altering the beliefs of learners [4]. Our assumption that humans will take actions expecting the optimal, complementary robot action interprets the expected human pedagogy as being through setting expectations of the robot.

While the robot uses the composite reward $R_{\theta^T,\theta^r,\hat{\theta}^h}$ to learn from the human's actions, the robot can seek out states that will give the human the opportunity to demonstrate its true composite reward $R_{\theta^T,\theta^r,\theta^h}$. Our key idea is that the robot's actions affect the state, and in turn affect the human's next actions, the robot can leverage this to actively take actions that will lead the team to states in which the human can provide more informative demonstrations. For example, consider the dishwasher unloading task, where the robot and human must collaboratively unload 3 bowls and 1 cup (see Figure 1). If the robot reaches for the cup, it leaves the human with no choice but to unload one of the three remaining bowls. Had the robot reached first for one of the bowls, the human would have the option of choosing between the cup or one of the three bowls. Opting for this more informative state would have given the robot more information about the human's preference between bowls and cups, since the human would have made a decision between the two objects. We will next formalize this desire to maneuver the team into informative state using an information gain metric.

Algorithm 2 represents the planning algorithm which seeks out next states with high potential information gain. In line 5, we perform a Bellman backup to compute a task-based $Q$ given the current beliefs $b$ (Equation **??**). $Q(s, a^r, a^h, b)$ represents expected discounted future composite team rewards given $b$. The robot's information gain

$$I(b, s, a^r) = H(b) - \mathbb{E}_{\hat{\pi}^h, T}[H(b|s, a^r)]$$

with expected entropy $\mathbb{E}_{\hat{\pi}^h,T}[H(b|s,a^r)]$ under the predicted human policy $\hat{\pi}^h$ and the transition dynamics $T$, using Equation **??**. In line 11 of Algorithm 2, the information gain of $a^r$ in state $s$ is added to the Q-value as a boost to actions that would guide the team towards most informative states (Equation **??**). If the next state is not informative given any human action, then there will be no information gain boost given to any action. The $\alpha$ term (Equation **??**) is a switch that turns on and off the information gain boost. The information gain objective seeks to reduce uncertainty over the human individual reward functions in $\Theta^h$. However, multiple reward functions can lead to the same predicted human action. When the beliefs begin converging on the same predicted human action, the robot will deprioritize seeking information gain. $\alpha$ measures the intersection of the shared predicted human actions by the top probability human individual reward functions in the beliefs $b$.

$$\alpha = |\bigcap_{\hat{\theta}^h \in \Theta^h} \{a \mid \pi^h(a|s;\hat{\theta}^h) = \max_b \hat{\pi}^h(b|s;\hat{\theta}^h) \wedge b(\hat{\theta}^h) = \max_{\bar{\theta}^h} b(\bar{\theta}^h)\}| \tag{2}$$

The robot thus replans under updated beliefs, while seeking informative states in which the human has opportunity to provide informative decision. If the next state is not informative given any human action, then there will be no information gain boost.

Upon observing the action of the human, current state, and next state, the robot updates its beliefs ($b$) over the hypothesis space ($\Theta^h$) and produces a predicted policy for the human ($\hat{\pi}^h$). The robot then replans its actions seeking information gain. This online decision process comprises our full algorithm: Bayesian Information Seeking Learner (BaISL). See Algorithm 3. BaISL is an approach to the ICPL task. While the task is not complete, in each timestep, the robot updates its beliefs based on the previous human action and state using Algorithm 1. Then, the robot replans seeking potential information gain using Algorithm 2 (line 8). The robot samples an action from its computed policy (line 9), while the human chooses an action as well. The environment transitions to the next state. Once the task is complete, the taem receives the total composite reward obtained over the interaction (line 14).

## 1.3 Simulated Evaluation

### 1.3.1 Simulated Human Models

We experiment with three types of simulated humans. For each agent type, we experiment with temperatures $\beta \to \infty$ (rational), and the other with $\beta = 1$.

**Algorithm 2** (PSIG) Plan Seeking Information Gain

**Input**: Beliefs over human reward functions $b$, Predicted human policy $\hat{\pi}^h$
**Parameter**: $\theta^r, \theta^T$
**Output**: $\pi^r$

1: $V(s) \leftarrow$ Initialize $V(s)$ randomly
2: **while** not converged **do**
3:  **for** $s \in \mathcal{S}$ **do**
4:  **for** $(a^r, a^h) \in \mathcal{A}^r \times \mathcal{A}^h$ **do**
5:  $Q(s,a^r,a^h,b) = \sum_{\hat{\theta}^h} b(\hat{\theta}^h) \sum_{s' \in S} T(s'|s,a^r,a^h)\Big(R_{\hat{\theta}^h,\theta^r,\theta^T}(s,a^r,a^h,s') + \gamma V(s')\Big)$
6:  **end for**
7:  $V(s) \leftarrow \max_{(a^r,a^h)\in\mathcal{A}^r\times\mathcal{A}^h} Q(s,a^r,a^h,b)$
8:  **end for**
9: **end while**
10: $\alpha = |\bigcap_{\hat{\theta}^h \in \Theta^h} \{a \mid \pi^h(a|s;\hat{\theta}^h) = \max_b \hat{\pi}^h(b|s;\hat{\theta}^h) \wedge b(\hat{\theta}^h) = \max_{\bar{\theta}^h} b(\bar{\theta}^h)\}|$
11: $\pi^r(s) \leftarrow \arg\max_{a^r\in\mathcal{A}^r} \sum_{a^h\in\mathcal{A}^h} \sum_{\hat{\theta}^h} b(\hat{\theta}^h)\hat{\pi}^h(a^h|s;\hat{\theta}^h)\Big(Q(s,a^r,a^h,b) + \alpha I(b,s,a^r)\Big)$

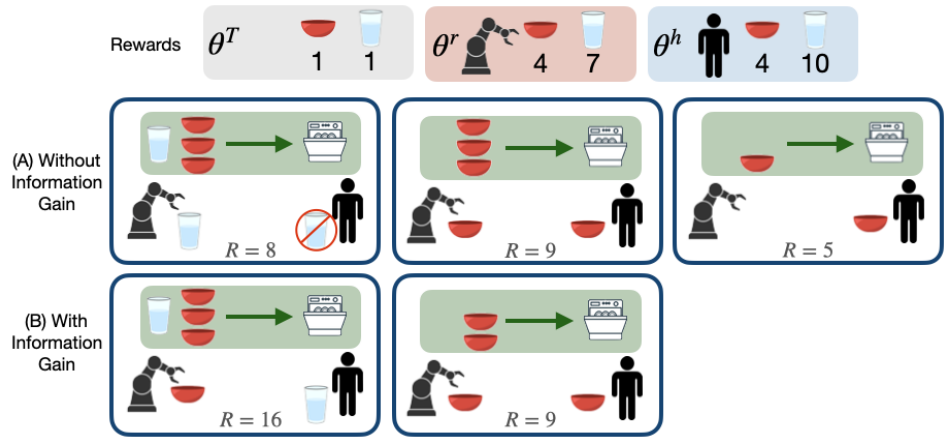


Figure 1: This toy example demonstrates an example of the effect of the information gain measure on the behavior of the robot. The robot and human must collaborate to load a dishwasher with 3 bowls and 1 cup; the team gets $+1$ reward for loading any object. The robot's individual reward is +7 for the cup, and +4 for the bowl. The human's individual reward is +10 for the cup, and +4 for the bowl. The robot's hypothesis space is two options: (1) +10 for the cup and +4 for the bowl, or (2) +4 for the cup and +10 for the bowl. Without information gain (A), the robot reaches for the cup, but it leaves the human with no choice but to unload one of the three remaining bowls, achieving a reward of $8 + 9 + 5 = 22$. With information gain (B), the robot opts for a more informative next state, where the robot is able to learn the human's preference of cups over bowls, achieving higher reward of $16 + 9 = 25$.

**Algorithm 3** (BaISL) Bayesian Information Seeking Learner: An approach to ICPL

**Input**: $s_0, \theta^T, \theta^r, \theta^h$
**Output**: Updated beliefs $b$, Predictive model of human action $f$

$b \leftarrow$ Initialize uniform prior over $\theta^h \in \Theta^h$
**for** task not over **do**
  Human observes $\theta^T, \theta^r, \theta^h$
  Robot observes $\theta^T, \theta^r$, and $\Theta^h$, but not $\theta^h$
  $s \leftarrow s_0$
  **while** game $n$ not over **do**
    $b, f \leftarrow \text{BIRL}(s, a^h, b)$ {update beliefs about human utility}
    $\pi^r \leftarrow \text{PSIG}(b, f)$ {plan seeking info gain using updated beliefs}
    $a^r \sim \pi^r(\cdot|s)$ {sample robot action $a^r$ from policy}
    $a^h \leftarrow$ Human decides $a^h$ based on $s$ {human takes $a^h$}
    $s \leftarrow s' \sim T(s'|s, a^h, a^r)$ {environment transitions to state $s'$}
  **end while**
**end for**
Return $R_F = \sum_t R_{\theta^T, \theta^r, \theta^h}(s_t, a^r_t, a^h_t, s_{t+1})$ {team observes final reward once task complete}

1. The *optimistic reward human* selects actions that maximize composite reward assuming the human takes the optimal complementary action. This model is the one used in our BIRL likelihood function, making it easier for *Ours* and *Ours wo IG* to perform well with. The simulated human model selects action $a^h$ with probability proportional to the reward, with a best-case prediction for $a^r$.

$$a^r \sim \pi^h(a^h|s_t) \propto exp\left(\beta \max_{a^r \in \mathcal{A}^r} \sum_{s_{t+1}} R_{\theta^T, \theta^r, \hat{\theta}^h}(s_t, a^h_t, s_{t+1}) T(s_{t+1}|s, a^h_t, a^r_t)\right)$$

2. The *optimistic reward human* selects actions that maximize *expected* composite reward under a uniform probability over all robot actions. This human decision making function does not reflect Prag-Ped nor Ours, making this human model out of distribution for both. The human

model computes its reward by marginalizing out $a^r$.

$$\pi^h(a^h|s_t) \propto exp\Bigg(\beta \sum_{a^r \in \mathcal{A}^r} \sum_{s_{t+1}} R_{\theta^T,\theta^r,\hat{\theta}^h}(s_t, a_t^h, s_{t+1}) T(s_{t+1}|s, a_t^h, a_t^r)\Bigg)$$

3. The *pedagogic human* uses the Pragmatic-Pedagogic Value Alignment [3] Q-value corresponding to the true human individual reward. Since the human computes the expected Q-value for its own actions by marginalizing over robot actions. This model is the one used in the *Prag-Ped* robot's likelihood function, making it easier for *Prag-Ped* to work well with. The Pragmatic-Pedagogic Value Alignment solution to CIRL leverages the assumption that the human can observe the robot's action at the current timestep before selecting its own action. The human policy maximizes the best expected outcome for each available action:

$$\pi^h(a^h|s, b, a^r, \hat{\theta}^h) \propto exp(\beta Q(s, b, a^h, a^r; \hat{\theta}^h)$$

In order to compute $Q$, the human considers the belief update of the robot, where the update of the robot's belief is determine given by the Bayesian update:

$$b'(\hat{\theta}^h|s, b, a^r, a^h) \propto \pi^h(a^h|s, b, a^r, \hat{\theta}^h) b(\hat{\theta}^h)$$

The robot's policy under the new beliefs maximizes the expected $Q$ under the new beliefs.

$$\pi^{r*}(s', b') = \arg\max_{a^r} \sum_{a^h, \hat{\theta}^h} Q(s, b, a^h, a^r; \hat{\theta}^h) b(\hat{\theta}^h)$$

The Bellman equation for the human is as follows:

$$Q(s, b, a^h, a^r; \hat{\theta}^h) = R_{\theta^T,\theta^r,\hat{\theta}^h}(s, a^r, a^h) + \mathbb{E}_{s',a^{h'}}\Big[\gamma Q'(s', b', a^{h'}, \pi^{r*}(s', b'); \hat{\theta}^h)\Big]$$

The human is pedagogic because the Bellman equation takes into account how the robot's beliefs will change based on the actions of the human. Since the human cannot actually see the robot's action ahead of time during this collaborative decluttering task, the human marginalizes out $a^r$:

$$\pi^h(a^h|s, b, a^r, \hat{\theta}^h) \propto exp(\sum_{a^r \in \mathcal{A}^r} \beta Q(s, b, a^h, a^r; \theta^h))$$

and acts according this policy.

### 1.3.2 Baselines

1. **BaISL: Bayesian Information Seeking Learner** The robot acts according to $\pi^r$ (Eq. **??**), modeling the human as an optimistic reward teacher and planning towards informative states. The robot selects the action $a^r$, according to Algorithm 3, maximizing the expected-reward-based Q-values and information gain boost.

$$\pi^r(s) \leftarrow \max_{a^r \in \mathcal{A}^r} \sum_{a^h \in \mathcal{A}^h} \sum_{\hat{\theta}^h} b(\hat{\theta}^h) \hat{\pi}^h(a^h|s; \hat{\theta}^h) \Big( Q(s, a^r, a^h, b) + \alpha I(b, s, a^r) \Big)$$

2. **BaL: Bayesian Learner** In this baseline, we ablate the information gain term. The robot policy $\pi^r$ is defined by Equation **??**, but without the information gain term. The robot models an *optimistic reward human* and plans using expected reward under current beliefs. The robot selects the action $a^r$ maximizing the expected-reward-based Q-values only.

$$\pi^r(s) \leftarrow \max_{a^r \in \mathcal{A}^r} \sum_{a^h \in \mathcal{A}^h} \sum_{\hat{\theta}^h} b(\hat{\theta}^h) \hat{\pi}^h(a^h|s; \hat{\theta}^h) \Big( Q(s, a^r, a^h, b) \Big)$$

3. **Prag-Ped: Pragmatic-Pedagogic Value Alignment [3]** As a baseline, we compare the performance of our agent against a solution to the CIRL [5] problem in which the human acts pedagogically while the robot reasons practically. The robot policy assumes that the human will act pedagogically with a Q value function that accounts for the robot's beliefs.

The Pragmatic-Pedagogic value alignment solution further assumes the human observes the robot's action before selecting their own action. The human policy maximizes the best expected outcome for each available action:

$$\pi^h(a^h|s,b,a^r,\hat{\theta}^h) \propto exp(\beta Q(s,b,a^h,a^r;\hat{\theta}^h)$$

Consider a state in which the human has only one action. an incorrect hypothesize reward achieved will be higher for , since under Thus, we want to ensure the probabilities reflect equal probability for selecting $b$ by thresholding the Boltzmann potential if the action yields a best-case reward equal to the maximum.

$$r(s_t,a_t^h|\hat{\theta}^h) = \begin{cases} \lambda, & \text{if } r_{bc}(s_t,a_t^h|\hat{\theta}^h) = \max\limits_{a^h \in \mathcal{A}^h} r_{bc}(s_t,a^h|\hat{\theta}^h) \\ 1-\lambda, & \text{otherwise} \end{cases}$$

In order to compute $Q$, the human considers the belief update of the robot, where the update of the robot's belief is determine given by the Bayesian update:

$$b'(\hat{\theta}^h|s,b,a^r,a^h) \propto \pi^h(a^h|s,b,a^r,\hat{\theta}^h)b(\hat{\theta}^h)$$

The robot's policy under the new beliefs maximizes the expected $Q$ under the new beliefs.

$$\pi^{r*}(s',b') = \arg\max_{a^r} \sum_{a^h,\hat{\theta}^h} Q(s,b,a^h,a^r;\hat{\theta}^h)b(\hat{\theta}^h)$$

The Bellman equation for the human is as follows:

$$Q(s,b,a^h,a^r;\hat{\theta}^h) = R_{\theta^T,\theta^r,\hat{\theta}^h}(s,a^r,a^h) + \mathbb{E}_{s',a^{h'}}\left[\gamma Q'(s',b',a^{h'},\pi^{r*}(s',b');\hat{\theta}^h)\right]$$

While the Prag-Ped solution computes both a policy for the human and robot, we take and execute the policy of the robot for this robot baseline. The robot takes the action maximizing expected reward under the beliefs, using $\pi^{r*}(s',b')$. The *pedagogic human* simulated model acts according to the human policy part of the Prag-Ped solution.

4. **MaxEnt** Our second baseline is Maximum Entropy Invese Reinforcement Learning [6], followed by replanning using the learned reward. The robot learns a reward function $f(s,a^r,a^h)$ representing $R_{\theta^T,\theta^r,\hat{\theta}^h}(s,a^r,a^h)$ using the demonstrated states and joint actions previously seen. The robot evaluates Q-values based on the learned reward function $f$:

$$Q(s,a^r,a^h;f) = f(s,a^r,a^h,s') + \gamma \sum_{s' \in S} T(s'|s,a^r,a^h) \max_{a^{r'},a^{h'}} Q(s',a^{r'},a^{h'};f)$$

$$\pi^{r*}(s) = \arg\max_{a^r} \sum_{a^h} Q(s,a^r,a^h;f)$$

### 1.3.3 Simulated Human Evaluation Results

*Ours* performs comparably to *Ours wo IG* when paired with the fully rational human partners. Both models utilize the *optimistic reward human* in the BIRL likelihood function, and thus reach optimal performance quickly by the second round. The information gain objective causes *Ours* to perform more exploratory actions, which do not contribute to task reward in the first round. When paired with the $\beta = 1$ human, *Ours* outperforms *Ours wo IG* on all three human models. With the rational *pedagogic human*, *Prag-Ped* outperforms the other three methods in the first round. However, *Ours* outperforms the other three baselines when paired with the $\beta = 1$ *pedagogic human*.

Table 1: Simulated ICPL: $\beta = 1$ Human. Entries in the table represent (mean, standard deviation) of the percent of optimal composite reward achieved by the team.

| | Optimistic Reward Human | | | Expected Reward Human | | | Pedagogic Human | | |
|---|---|---|---|---|---|---|---|---|---|
| | n=1 | n=2 | n=3 | n=1 | n=2 | n=3 | n=1 | n=2 | n=3 |
| BaISL | 0.92,0.01 | 0.96,0.1 | 0.97,0.1 | 0.95,0.1 | 0.96,0.1 | 0.97,0.1 | 0.83,0.2 | 0.88,0.1 | 0.91,0.1 |
| BaL | 0.95,0.1 | 0.97,0.1 | 0.96,0.1 | 0.95,0.1 | 0.97,0.1 | 0.96,0.1 | 0.85,0.2 | 0.87,0.1 | 0.89,0.2 |
| Prag-Ped | 0.91,0.1 | 0.92,0.1 | 0.92,0.1 | 0.89,0.2 | 0.91,0.2 | 0.92,0.1 | 0.94,0.1 | 0.90,0.2 | 0.90,0.2 |
| MaxEnt IRL | 0.92,0.1 | 0.91,0.1 | 0.91,0.1 | 0.85,0.1 | 0.91,0.1 | 0.91,0.1 | 0.77,0.2 | 0.85,0.2 | 0.85,0.2 |

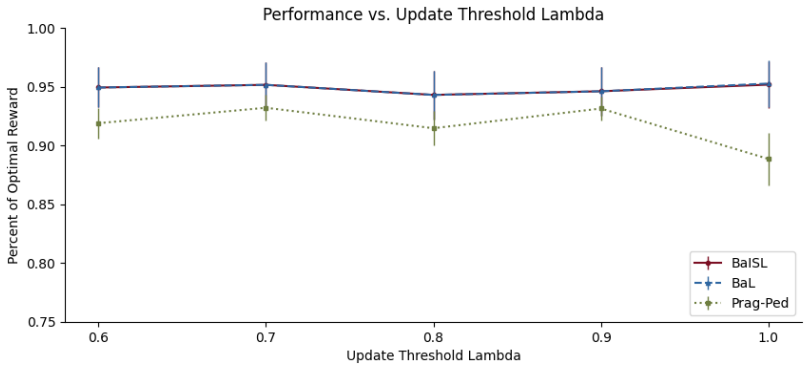


Figure 2: Performance evaluation of our models across different values of our $\lambda$ hyperparameter used in BIRL (line 3, Algorithm 1) shows that the **BaISL, BaL** methods are unaffected by $\lambda$ when partnering with the $\beta = 1$ *optimistic reward human*.

Table 2: Simulated ICPL: $\beta = \inf$ Human. Entries in the table represent (mean, standard deviation) of the percent of optimal composite reward achieved by the team.

| | Optimistic Reward Human | | | Expected Reward Human | | | Pedagogic Human | | |
|---|---|---|---|---|---|---|---|---|---|
| | n=1 | n=2 | n=3 | n=1 | n=2 | n=3 | n=1 | n=2 | n=3 |
| BaISL | 0.97,0.1 | 1.0,0.0 | 1.0,0.0 | 0.91,0.3 | 1.0,0.0 | 1.0,0.0 | 0.86,0.1 | 0.91,0.1 | 0.93,0.1 |
| BaL | 0.99,0.0 | 1.0,0.0 | 1.0,0.0 | 0.91,0.3 | 1.0,0.0 | 1.0,0.0 | 0.90,0.1 | 0.91,0.1 | 0.92,0.1 |
| Prag-Ped | 0.92,0.2 | 0.91,0.2 | 0.91,0.2 | 0.88,0.2 | 0.88,0.2 | 0.89,0.2 | 0.93,0.2 | 0.93,0.2 | 0.95,0.2 |
| MaxEnt IRL | 0.81,0.2 | 0.92,0.1 | 0.93,0.1 | 0.83,0.1 | 0.91,0.1 | 0.92,0.1 | 0.78,0.2 | 0.88,0.1 | 0.85,0.2 |

### 1.3.4 Hyperparameter Analysis

We analyze the performance of our models across the $\lambda$ hyperparameter used in BIRL (line 3, Algorithm 1). We test $\lambda$ from 0.6 to 1.0 in 0.1 intervals, using the $\beta = 1$ *optimistic reward human* with robots **BaISL, BaL, Prag-Ped**, using 50 random game configurations. We find that performance of our algorithm BaISL is not affected by the choice of $\lambda$ value. The **Prag-Ped** robot is mostly unaffected as well, but performance drops slightly with $\lambda = 1.0$. We use $\lambda = 0.9$ in our agents for all other simulated experiments. Additionally, we analyze the performance of our models across the $\beta$ hyperparameter used in the simulated human model for the *optimistic reward human* with robots **BaISL, BaL, Prag-Ped**, using 50 random game configurations. The robots update their beliefs using a beta of 1, while the TRUE $\beta$ values for the simulated human are plotted on the x-axis in Figure 3. We find that performances of **BaISL, BaL**, and **Prag-Ped** are not affected by true $\beta$ value for the simulated human.

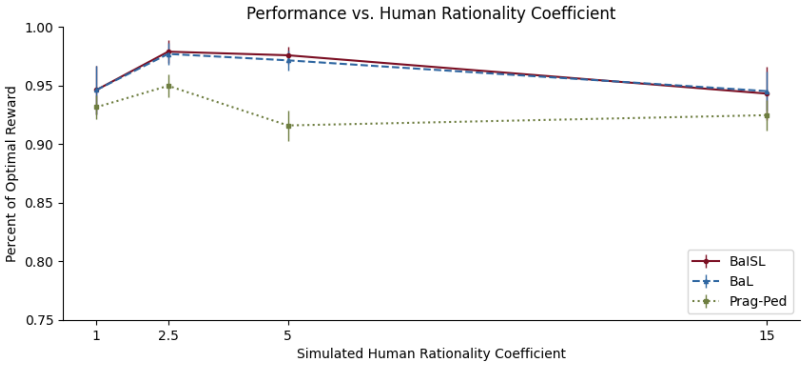


Figure 3: We find that performances of **BaISL, BaL**, and **Prag-Ped** are not affected by true $\beta$ value for the simulated human.

# References


[1] H. J. Jeon, S. Milli, and A. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

[2] C. Laidlaw and A. Dragan. The boltzmann policy distribution: Accounting for systematic suboptimality in human models. *arXiv preprint arXiv:2204.10759*, 2022.

[3] J. F. Fisac, M. A. Gates, J. B. Hamrick, C. Liu, D. Hadfield-Menell, M. Palaniappan, D. Malik, S. S. Sastry, T. L. Griffiths, and A. D. Dragan. Pragmatic-pedagogic value alignment. In *Robotics Research*, pages 49–57. Springer, 2020.

[4] P. Shafto, N. D. Goodman, and T. L. Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89, 2014.

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

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