A On the practicality of the proposed interventions

We propose three interventions for influencing humans towards a specific preference model, and argue the benefits of doing so for alignment. Two of these interventions, tested in the Trained and Question experiments, hold promise of real world practicality. Our goal was to present—to the best of our knowledge—the first exploration into how to systematically influence human preference expression to better match RLHF algorithmic assumptions. This represents a fundamentally new and potentially impactful approach to improving model alignment. All experiments demonstrate that human preference expression can indeed be systematically influenced to better match specific preference models. While we used the ground-truth reward function for evaluation purposes, this establishes the core feasibility of our approach.

The Trained experiment lays the groundwork for extensions to more complex domains and scenarios where the ground-truth reward function is unknown or is known for the purposes of experimental evaluation—as in our experiments—but differs from any reward function(s) used during training. The significant effects observed in the Trained experiment suggest the potential efficacy of training humans to follow a chosen preference model. We plan to further address the practical application of this approach in future work, where we will elicit human preferences from a domain with a ground-truth reward function that is different from the one used for teaching subjects about a specific preference model.

The Question experiment already demonstrates a viable path forward, showing significant results for one condition in which the intervention does not rely upon any knowledge of the ground-truth reward function.

B The regret preference model

B.1 Intuition behind the regret preference model

Recall that, from Equation 3, the deterministic regret of a segment is given by $regret_d(\sigma|\tilde{r}) = V^*_{\tilde{r}}(s^{\sigma}_0) - (\Sigma_\sigma \tilde{r} + V^*_{\tilde{r}}(s^{\sigma}_{|\sigma|})$. Deterministic regret quantifies the extent to which a segment diminishes expected return from $V^*_{\tilde{r}}(s^{\sigma}_0)$. An optimal segment σ^* has 0 regret, while a suboptimal segment $\sigma^{\neg *}$ has positive regret. When two segments have deterministic transitions, end in terminal states, and share the same starting state, this regret preference model is equivalent to the partial return preference model: $P_{regret}(\cdot|\tilde{r}) = P_{\Sigma_r}(\cdot|\tilde{r})$. Conceptually, the partial return preference model assumes that preferences are determined solely by the reward-yielding outcomes *within* the segments, whereas the regret preference model bases preferences on how much the segments deviate from optimal behavior.

B.2 Comparing two segments with the same start state

When computing the difference in deterministic regret for two segments with the same start state, the start state value, $V^*_{\tilde{r}}(s^{\sigma}_0)$, cancels out:

$$regret_d(\sigma_1|\tilde{r}) - regret_d(\sigma_2|\tilde{r})$$

$$= V^*_{\tilde{r}}(s^{\sigma_{\sigma_1}}_0) - (\Sigma_{\sigma_{\sigma_1}} \tilde{r} + V^*_{\tilde{r}}(s^{\sigma_{\sigma_1}}_{|\sigma_{\sigma_1}|}) - V^*_{\tilde{r}}(s^{\sigma_{\sigma_2}}_0) + (\Sigma_{\sigma_{\sigma_2}} \tilde{r} + V^*_{\tilde{r}}(s^{\sigma_{\sigma_2}}_{|\sigma_{\sigma_2}|})$$

$$= -(\Sigma_{\sigma_{\sigma_1}} \tilde{r} + V^*_{\tilde{r}}(s^{\sigma_{\sigma_1}}_{|\sigma_{\sigma_1}|}) + (\Sigma_{\sigma_{\sigma_2}} \tilde{r} + V^*_{\tilde{r}}(s^{\sigma_{\sigma_2}}_{|\sigma_{\sigma_2}|})$$

C Additional information on the delivery domain and creating a human-labeled preference dataset

When teaching subjects about the delivery domain and constructing the preference datasets for the Privileged experiment, detailed in Section 5.1, we follow the same procedure as Knox et al. (2022). For the Trained and Question experiments, detailed in Sections 5.2 and 5.3 respectively, we modify the interface, preference elicitation, and human subject filtering procedure. These changes are detailed below where applicable.

The Privileged experiment was designed as a proof-of-concept; that human preferences could be influenced towards a specific preference model. When designing the Trained and Question experiments, which focus on interventions for the real world, we departed from the experimental protocol used by Knox et al. (2022) to better reflect the hypotheses we wished to test.

C.1 The delivery domain and task

The delivery domain is structured as a grid composed of cells, each containing a specific type of road surface. A task within the delivery domain is illustrated in Figure **??**. The state of the delivery agent is its location on the grid. The agent can move one cell in any of the four cardinal directions. Episodes conclude either successfully at the destination, earning a reward of +50, or in failure upon encountering a sheep, resulting in a reward of -50. Non-terminal transitions have a reward equal to the sum of their components: cells with a white road surface carry a -1 reward component, while cells with a brick surface carry a -2 component. Additionally, cells may contain a coin (+1) or a roadblock (-1). Coins remain in place and can, at best, cancel out the cost of the road surface.

Actions that would result in the agent moving into a house or beyond the grid's boundaries result in no movement. In such cases, the reward reflects the current cell's surface component but excludes any coin or roadblock components. The start state distribution, D_0, is uniformly random over non-terminal states.

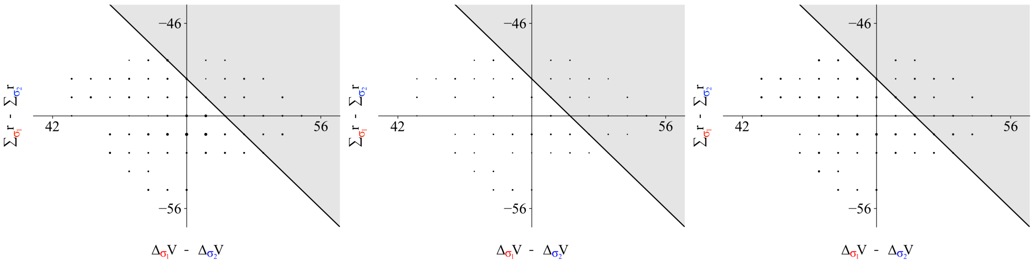
This domain was intentionally designed to make it easy for subjects to recognize poor behavior while making it challenging to discern optimal behavior from most states, mirroring many real-world tasks. This complexity means that some assumptions of the regret preference model, specifically that humans will always prefer optimal segments over suboptimal ones, are not always met, providing a robust test of the model's performance under realistic conditions.

C.2 Selecting segment pairs for preference elicitation

During the main preference elicitation portion of all experimental conditions, preferences are collected over trajectory segments sampled from the delivery task shown in Figure **??**. Below we outline our methodology for selecting segment pairs for labeling in the Privileged experiment, as well as separately for the Trained and Question experiments.

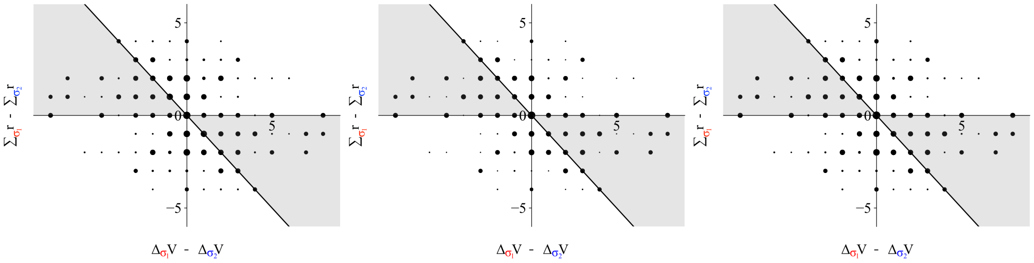
Privileged Experiment We followed the methodology of Knox et al. (2022) for collecting segment pairs, which involved two stages of data collection with differing goals. The first stage sought to characterize human preferences over a range of possible behaviors, including those that would highlight the differences between partial return and regret. The second stage sought to collect preferences over segment pairs that resolve the identifiablity issues of the partial return preference model related to a constant shift in the reward function. We refer readers to Knox et al. (2022) for a detailed description on how these segment pairs were constructed. Figure 11 plots the coordinates from which segment pairs where sampled for each condition in the first stage of data collection. Figures 12 and 13 plot these coordinates for the second of data collection.

The first stage of data collection resulted in $1,359$ segment pairs from 39 subjects for the P_{Σ_r}-Privileged condition, $1,418$ segment pairs from 42 subjects for the Privileged-Control condition, and $1,501$ segment pairs from 43 subjects for the P_{regret}-Privileged condition. All trajectory segments consisted of 3 actions, and the start state for each segment in a pair was different. The second stage of data collection resulted in $1,173$ segment pairs from 25 subjects for the P_{Σ_r}-Privileged condition, 375 segment pairs from 8 subjects for the Privileged-Control condition, and $1,030$ segment pairs from 22 subjects for the P_{regret}-Privileged condition. For each segment pair in the second stage, the agent in one segment takes 3 actions while in the other segment it reaches a terminal state in fewer than 3 actions. Each subject is asked to label preferences for between 35 and 50 segment pairs. No two subjects see the same segment pairs.



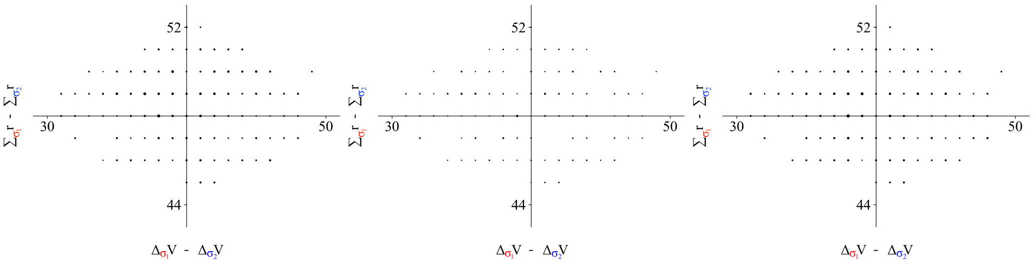
P_{Σ_r}-Privileged Privileged-Control P_{regret}-Privileged

Figure 13: The coordinates of the segment pairs shown to subjects for preference labeling in the second stage of data collection for the Privileged experiment. Each segment pair belonging to these graphs contain one segment where the agent terminates at a positive terminal state and one where it does not. The proportionality of the circles are consistent across this plot and the 3 subplots of Figure 12 and 13.



P_{Σ_r}-Privileged Privileged-Control P_{regret}-Privileged

Figure 11: The coordinates of the segment pairs shown to subjects for preference labeling in the first stage of data collection for the Privileged experiment. The x-axis is the difference in the change in state value between the two segments and the y-axis is partial return differences between the two segments. The areas of the circles are proportional to the number of segment pairs at that point. The proportionality is consistent across this plot and the 3 subplots of Figures 12 and 13.



P_{Σ_r}-Privileged Privileged-Control P_{regret}-Privileged

Figure 12: The coordinates of the segment pairs shown to subjects for preference labeling in the second stage of data collection for the Privileged experiment. Each segment pair belonging to these graphs contain one segment where the agent terminates and one where it does not. The proportionality of the circles are consistent across this plot and the 3 subplots of Figure 11 and 13.

Trained and Question Experiments Subjects label the same dataset of 500 segment pairs for each of the 3 conditions in the Trained and Question experiment. We chose 500 segment pairs for preference labeling by splitting the P_{Σ_r}-Privileged, Privileged-Control, and P_{regret}-Privileged datasets into different numbers of same sized partitions. We then identified the smallest partition size where the likelihood of the influenced preference dataset was always significantly higher than that of the control condition, defined as being 100 times more likely.

In 72 of the 500 pairs used for labeling, the agent in one trajectory segment takes 3 actions, while in the other segment, it reaches a terminal state in fewer than 3 actions. Knox et al. (2022) found that these trajectory segments are essential for learning a reward function with the partial return preference model in this domain. The remaining trajectory segments consist of 3 actions sampled uniformly randomly from all possible actions. The trajectory segments in a segment pair have the same start state. The order of trajectory segments within a pair and the order of segment pairs shown to subjects is uniformly random. Each subject is asked to label preferences for 50 segment pairs. No two subjects within a condition see the same segment pairs, except for the last segment pair used to test for task comprehension and attention. Figure 19 illustrates the segment pairs sampled for labeling for all experimental conditions within these two experiments.

In the Trained experiment we collected data from 10 subjects per condition. In the Question experiment, we sought to do the same but ran into the following issue: We randomly sample subjects to complete our study from a standard sample of available subjects who meet certain criteria (see Appendix C.3), utilize random assignment to assign each subject to a condition, and recollect data removed from subjects who failed the comprehension tests (detailed in Appendix C.4). For the P_{regret}-Question condition, we were unable to find a 10th subject who passed the comprehension test before the subject sample population potentially changed significantly over time. As such, we were no longer confident that we could claim the subjects from all conditions in the Question experiment were drawn from the same population. Therefore, we collected data from only 9 subjects per condition in the Question experiment, resulting in a maximum dataset size of 450 preferences per condition.

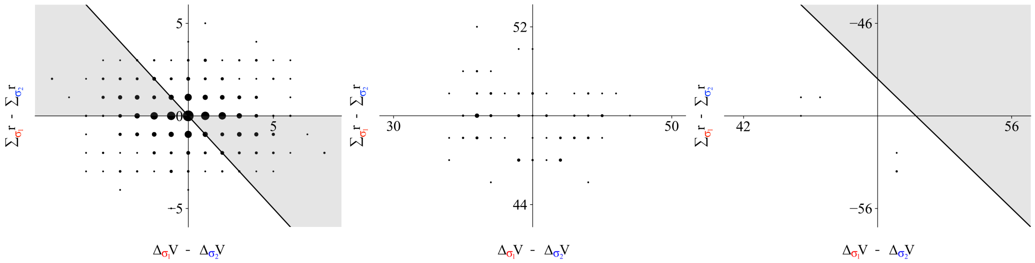


Figure 14: The coordinates of the segment pairs shown to subjects for preference labeling for all conditions for the Trained and Question experiments. The areas of the circles are proportional to the number of segment pairs at that point. The proportionality is consistent across this plot.

C.3 Recruiting human subjects

All subject compensation amounts were chosen using the median time subjects took during a pilot study and then calculating the payment to result in \$15 USD per hour. This hourly rate of \$15 was chosen because it is commonly recommended as an improved US federal minimum wage.

Privileged Experiment We recruited subjects with IRB approval via Amazon Mechanical Turk and paid subjects \$5 per experiment. Subjects had to be located in the United States, have an approval rating of at least 99%, and have completed at least 100 other studies on Mechanical Turk to join our study. Due to an experimental error, we did not show the IRB-approved consent form to participants after they accepted our study on Mechanical Turk. We reported this issue to our IRB and received approval to use the collected data.

Trained and Question Experiments We recruited subjects with IRB approval via Prolific. We paid subjects in the Trained experiment (see Section 5.2) \$7.50 for completing the study. We observed that subjects took about half the time to complete the study when they were not taught about a specific preference model, such as in the control condition, and so for the Question experiment (see Section 5.3) we paid subjects \$3.75 for completing the study. Subjects were recruited via Prolific from a standard sample, and were required to be both fluent in English and located in the United States.

C.4 FILTERING SUBJECT'S DATA

We evaluated each subject's understanding of the delivery domain and excluded those who lacked sufficient understanding. Participants were required to complete a task-comprehension survey, from which we derived a task-comprehension score. The questions and corresponding answer choices are detailed in Table 1. Participants received 1 point for fully correct answers and 0.5 points for partially correct answers. For the Privileged experiment—and separately for the Trained and Question experiments—the threshold score for removing worker data was determined by visually inspecting a histogram of the scores, aiming to strike a balance between upholding high comprehension standards and retaining a sufficient dataset for analysis.

In addition to filtering based on the task-comprehension survey, we also removed data from any participant who ever preferred a segment where the agent ends in a negative terminal state—the worst possible outcome—over a segment where the agent does not.

Privileged Experiment Subjects could achieve a score on the task-comprehension survey ranging from 0 to 7. Data from participants scoring below 4.5 was discarded. All subjects were shown at least one segment pair containing a segment where the agent ends in a negative terminal state and a segment where it does not, and their data is removed if they prefer the former. These segment pairs are used to test for task comprehension and attentiveness. Because of a data management error, the filtered-out data was lost and we don't otherwise know how many subjects were filtered out for this experiment.

Trained and Question Experiment Subjects could achieve a score on the task-comprehension survey ranging from 0 to 6. Data from participants scoring below 3.5 was discarded. The last segment pair shown to participants during preference elicitation always contained a segment where the agent ends in a negative terminal state and a segment where it does not. Other segment pairs shown to subjects may also illustrate this scenario.

Across all conditions in the Trained experiment, the data from $19/49$ subjects were removed: $9/19$ from the Trained-Control condition, $9/19$ from the P_{Σ_r}-Trained condition, and $1/11$ from the P_{regret}-Trained condition. The data from $33/60$ subjects in the Question experiment were removed: $11/20$ from the Question-Control condition, $9/18$ from the P_{Σ_r}-Question condition, and $13/22$ from the P_{regret}-Question condition.

D PRIVILEGED EXPERIMENT INTERFACE DETAILS

For all conditions in the Privileged experiment, when subjects interact with episodes from the grid-world domain we display four segment statistics: the "score" or partial return, the "best possible score from start" or $V_r^*(s_0^\sigma)$, the "best possible score given your moves" or $\tilde{r} + V_r^*(s_{|\sigma|}^\sigma)$, and the "opportunity cost" or regret, which is the difference between the previous two components. We explain $V_r^*(s_0^\sigma)$ as "the most money the vehicle could have made from the start", $\tilde{r} + V_r^*(s_{|\sigma|}^\sigma)$ as "the most money the vehicle can make from the start, including the route you've taken so far", and regret as "the difference between the two" and the "minimum amount of money lost by taking your route instead of the best route". See Figure 15 for an example of what humans see when interacting with an episode from the domain.

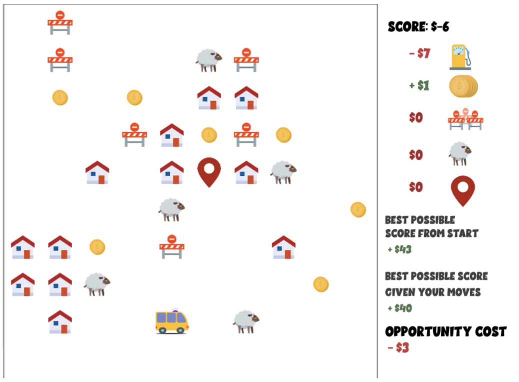
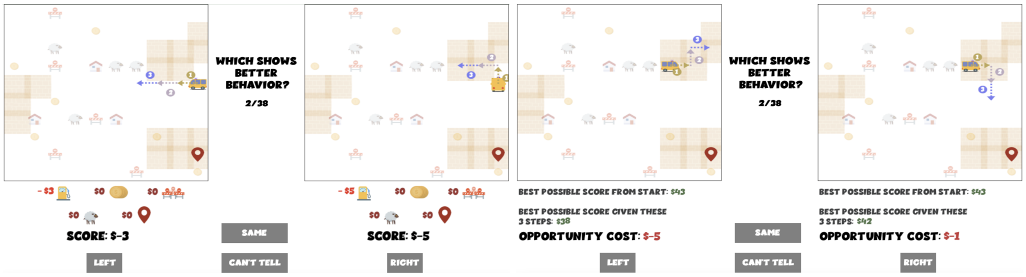

SCORE: $-6
- $7
+ $1
$0
$0
$0
BEST POSSIBLE SCORE FROM START
+ $43
BEST POSSIBLE SCORE GIVEN YOUR MOVES
+ $40
OPPORTUNITY COST
- $3

Figure 15: An episode from the grid-world domain used for teaching subjects about the domain transition and reward function, as well as to aid in their understanding of partial return (i.e., "score") and regret (i.e., "opportunity cost").

For all conditions in the Privileged experiment, we ask subjects to label preferences for $35-50$ segment pairs using the question "Which shows better behavior?". Only the information shown during preference elicitation differs between conditions, as detailed below.

Table 1: The task comprehension survey, designed to test participant's comprehension of the domain for the purpose of filtering data. Each full credit answer earned 1 point; each partial credit answer earned 0.5 points. The left most column indicates the experiment where the given question was used to filter subjects. We discarded the data of participants who scored less than 4.5 points overall for the Privileged experiment, and less than 3.5 points overall for the Trained and Question experiments.

Experiment	Question	Full credit answer	Partial credit answer	Other answer choices
Privileged, Trained, Question	What is the goal of this world? (Check all that apply.)	• To maximize profit	• To get to a specific location. • To maximize profit Partial credit was given if both answers were selected.	• To drive as far as possible to explore the world. • To collect as many coins as possible. • To collect as many sheep as possible. • To drive sheep to a specific location.
Trained, Question	What happens when you run into a house?	• You incur a gas cost and don't go anywhere.	• You incur a gas cost and a cost for hitting the house, and you don't go anywhere. • You incur a gas cost and a cost for hitting the house, and you drive over the house. • Nothing happens.	• The episode ends. • You get stuck. • To collect as many sheep as possible.
Privileged	What happens when you run into a house? (Check all that apply.)	• You pay a gas penalty. • You can't run into a house; the world doesn't let you move into it. Full credit was given if both answers were selected.	• You pay a gas penalty. • You can't run into a house; the world doesn't let you move into it. Partial credit was given if only one answer was selected.	• The episode ends. • You get stuck. • To collect as many sheep as possible.
Privileged, Trained, Question	What happens when you run into a sheep? (Check all that apply.)	• The episode ends. • You are penalized for running into a sheep. Full credit was given if both answers were selected.	• The episode ends. • You are penalized for running into a sheep. Partial credit was given if only one answer was selected.	• You are rewarded for collecting a sheep.
Privileged, Trained, Question	What happens when you run into a roadblock? (Check all that apply.)	• You pay a penalty.		• The episode ends. • You get stuck. • You can't run into a roadblock; the world doesn't let you move into it.
Privileged, Trained, Question	Is running into a roadblock ever a good choice in any town?	• Yes, in certain circumstances.		• No.
Privileged	What happens when you go into the brick area? (Check all that apply.)	• You pay extra for gas.		• The episode ends. • You get stuck in the brick area. • You can't go into the brick area; the world doesn't let you move into it.
Privileged, Trained, Question	Is entering the brick area ever a good choice?	• Yes, in certain circumstances		• No



BEST POSSIBLE SCORE FROM START: $43

BEST POSSIBLE SCORE GIVEN THESE 3 STEPS: $38

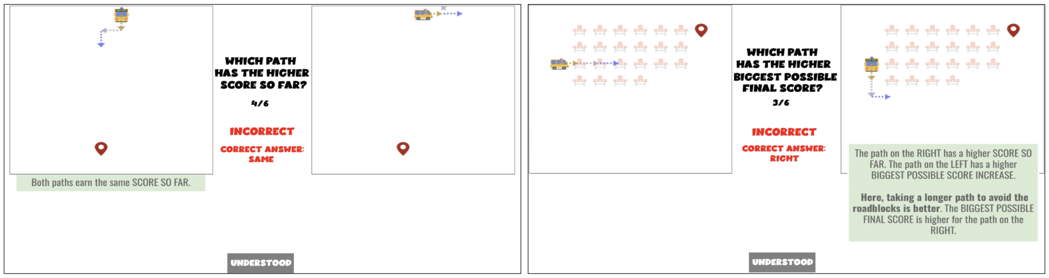
OPPORTUNITY COST: $-5

BEST POSSIBLE SCORE FROM START: $43

BEST POSSIBLE SCORE GIVEN THESE 3 STEPS: $42

OPPORTUNITY COST: $-1

Figure 16: An example of the preference elicitation interface shown to subjects in the P_{Σ_r}-Privileged condition (Left) and the P_{regret}-Privileged condition (Right).


WHICH PATH HAS THE HIGHER SCORE SO FAR?
4/6
INCORRECT
CORRECT ANSWER: SAME
UNDERSTOOD
WHICH PATH HAS THE HIGHER BIGGEST POSSIBLE FINAL SCORE?
3/6
INCORRECT
CORRECT ANSWER: RIGHT
UNDERSTOOD

Both paths earn the same SCORE SO FAR.

The path on the RIGHT has a higher SCORE SO FAR. The path on the LEFT has a higher BIGGEST POSSIBLE SCORE INCREASE.

Here, taking a longer path to avoid the roadblocks is better. The BIGGEST POSSIBLE FINAL SCORE is higher for the path on the RIGHT.

Figure 18: In the Trained experiment, subjects are shown two sets of six segment pairs where, after labeling their preference, they are given feedback on its correctness. For these practice segment pairs, subjects are given an explanation of why one segment is preferable to another regardless of whether their preference is correct. The left interface above displays an example of preference feedback for the partial return preference model and the right interface for the regret preference model. Preference feedback is only given for domain tasks separate from the delivery task used during the main preference elicitation session. To avoid technical jargon, we refer to partial return as "score so far", the end state value as "biggest possible score increase", and regret as "biggest possible final score". Subjects are taught these concepts during training.

During preference elicitation, we display the "score", or the partial return, for the vehicle's path and each corresponding reward component. See Figure 16 for the preference elicitation interface for the P_{Σ_r}-Privileged condition.

D.1 P_{regret}-Privileged condition Interface details

During preference elicitation, we display the "best possible score from start" or $V^*_r(s^\sigma_0)$, the "best possible score given your moves" or $\tilde{r} + V^*_r(s^\sigma_{|\sigma|})$, and the "opportunity cost" or regret, which is the difference between the previous two components. See Figure 16 for the preference elicitation interface for the P_{regret}-Privileged condition.

D.2 Privileged-Control condition Interface details

During preference elicitation, we do not display any segment statistics. See Figure 17 for the preference elicitation interface for the Privileged-Control condition.

E Trained Experiment Interface details

The interfaces employed for each condition in the Trained experiment differ in what preference model—if any—human subjects' preferences are influenced towards. Therefore, the concepts taught throughout the study and the preference elicitation instruction differ between conditions, as outlined below.

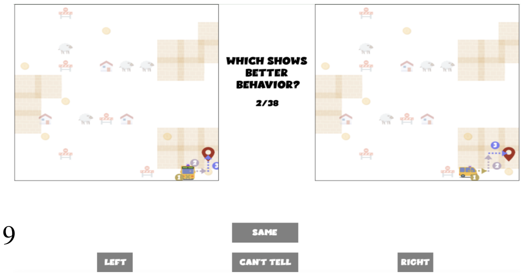


Figure 17: An example of the preference elicitation in-

E.1 P_{Σ_r}-TRAINED CONDITION INTERFACE DETAILS

When subjects interact with episodes from the grid-world domain, we display the "score so far", or the partial return, for the vehicle's path. We explain how the "score so far" is computed, have them compute it for three trajectory segments while providing feedback, and then let them interact with additional episodes to observe the "score so far".

After understanding partial return, subjects were instructed to use it when generating preferences. Initially, they labeled preferences for six segment pairs without specific guidance. Subsequently, they were told to generate preference labels based on the "score so far" and walked through a detailed example. They then labeled six more segment pairs, receiving feedback on the correctness of their preferences. An example of this interface is shown in Figure 18. Finally, subjects were instructed on how *not* to generate preferences (i.e., "do not select the path where the van looks like it might achieve a higher score in the future"), and given another six segments pairs to label with feedback on their preferences.

The trajectory segments used to teach subjects the partial return preference model were collected from various delivery tasks, excluding the one depicted in Figure **??**, which is reserved for the main preference elicitation portion. These pedagogical segment pairs were selected to illustrate scenarios where the partial return of both segments were equal, where one segment had a higher partial return but higher regret than the other (i.e., the two competing preference models would disagree on the preference label), and where one segment had a higher partial return and lower regret than the other. These preferences were not used for reward learning.

After learning to use the partial return preference model, subjects interacted with the delivery task shown in Figure **??** and generated preferences for 50 segment pairs from this task. When labeling these segment pairs, subjects were asked "which path has the highest score so far?". This is the main preference elicitation phase, where no feedback or information about the ground-truth reward function is provided. The preference elicitation interface is shown in Figure 4.

E.2 P_{regret}-TRAINED CONDITION INTERFACE DETAILS

We progressively teach subjects how to compute the regret of a trajectory segment by sequentially introducing its components outlined in Equation 3. When subjects interact with episodes from the delivery domain, we display the components of regret as we introduce them. First, we introduce the "score so far," $\Sigma_\sigma r$. Next, we explain the "biggest possible score increase," $V_r^*(s^\sigma_{|\sigma|})$. Finally, we present the "biggest possible final score," which combines both components as $\Sigma_\sigma r + V_r^*(s^\sigma_{|\sigma|})$. For each component, we explain how it is computed and allow subjects to interact with various tasks to observe the corresponding values. Additionally, we ask subjects to compute $V_r^*(s^\sigma_{|\sigma|})$ and $\Sigma_\sigma r + V_r^*(s^\sigma_{|\sigma|})$ for three different trajectory segments each, providing feedback on their answers. Since we always present segment pairs that share the same start state, we do not introduce $V_r^*(s^\sigma_0)$ to subjects because this component cancels out when computing preference distributions using the regret preference model (See Appendix B.2).

After understanding regret, subjects are taught to use it when generating preferences following a procedure similar to that in Section E.1. They are first asked to label preferences for six trajectory segments. Then, they are instructed to generate preference labels based on the "biggest possible final score" and shown a detailed example. Subjects label six more trajectory segments, receiving feedback on the correctness of their preferences as illustrated in Figure 18. Following this, they are instructed on how *not* to generate preferences (i.e., "do not select the path that merely has the higher score so far"). Finally, they are given another six segment pairs to label, with feedback provided on their preferences. Note that these pedagogical segment pairs are the same as those used in the P_{Σ_r}-Trained condition, detailed in Appendix E.1.

The subsequent preference elicitation procedure used to build the dataset of preferences for reward learning is identical to the procedure outlined in Section E.1, except subjects are asked “which path has the highest biggest possible final score” when generating preferences.

E.3 Trained-Control condition Interface details

When subjects interact with episodes from the grid-world domain, we display the “score so far”, or the partial return, for the vehicle’s path and explain this statistic in the same way as for the P_{Σ_r}-Trained condition. We do not train subjects to compute any segment statistic, nor do we instruct them on how to generate preferences like in the other conditions. During preference elicitation, we ask subjects “Which path do you prefer?”, a question that does not seek to influence subjects towards any preference model. This preference elicitation procedure aims to be generally representative of standard approaches for collecting feedback for RLHF.

E.4 Testing subjects comprehension of the taught preference model

When teaching subjects to follow a specific preference model, we present them with two sets of six practice segment pairs and provide feedback on their preference labels for these pairs. We test the correlation between the subject’s adherence to the taught preference model during the last six practice pairs and in the main preference elicitation portion of the study. We compute the Spearman correlation coefficient between the fraction of human preferences that agree with the noiseless version of the taught model in the last six practice segment pairs and in the fifty segment pairs shown during the main preference elicitation portion.

We are not able to perform this analysis for the P_{Σ_r}-Trained condition; the fraction of human preferences that agree with the noiseless version of the partial return preference model in the last six practice segment pairs remains constant for all subjects and therefore the Spearman correlation coefficient is undefined. For the P_{regret}-Trained condition, we compute a Spearman correlation coefficient of -0.137 with $p = 0.706$. We suspect that the high p-value is a result of the small sample size of only 10 subjects.

E.5 Surveying subject agreement of the taught preference model

During the post-study task-comprehension survey, we assess subjects’ personal agreement with the taught preference model. For the P_{Σ_r}-Trained condition, we ask, “We told you that the better path is always the one with the higher SCORE SO FAR. How often did you agree with this?” For the P_{regret}-Trained condition, we ask, “We told you that the better path is always the one with the higher BIGGEST POSSIBLE FINAL SCORE. How often did you agree with this?” Responses are given on a 7-point Likert scale, where 1 indicates “always disagreed” and 7 indicates “always agreed.” The mean response for the P_{Σ_r}-Trained condition was 4.2 with a variance of 1.56, while the P_{regret}-Trained condition had a mean response of 6.3 with a variance of 1.61. These results suggest that subjects personally aligned more with the regret-based labeling of segment pairs than with the partial return-based approach.

For both the P_{Σ_r}-Trained and P_{regret}-Trained conditions, we also asked subjects “How helpful were our explanations on why one path was better than another path for your own decision making?” The mean response for the P_{Σ_r}-Trained condition was 5.8 with a variance of 2.36, while the P_{regret}-Trained condition had a mean response of 6.4 with a variance of 0.84. We interpret these results as general satisfaction with our protocol for teaching subjects about a specific preference model.

E.6 Number of preferences per collected dataset

We collected 500 preferences for each condition outlined in the Trained experiment. Because we discard samples where a subject chose “Can’t Tell” instead of a preference, each dataset contains a different number of preferences indicated in Table 2.

F Question Experiment Interface details

The conditions in the Question experiment only differ in what question is asked during preference elicitation. The control condition used for the Question experiment is identical to that used for the Trained experiment, detailed in Appendix E.3.

Table 2: The number of preferences in each preference dataset resulting from the Trained experiment.

Condition	Number of Preferences in Dataset
P_{Σ_r}-Trained	497
Trained-Control	473
P_{regret}-Trained	491

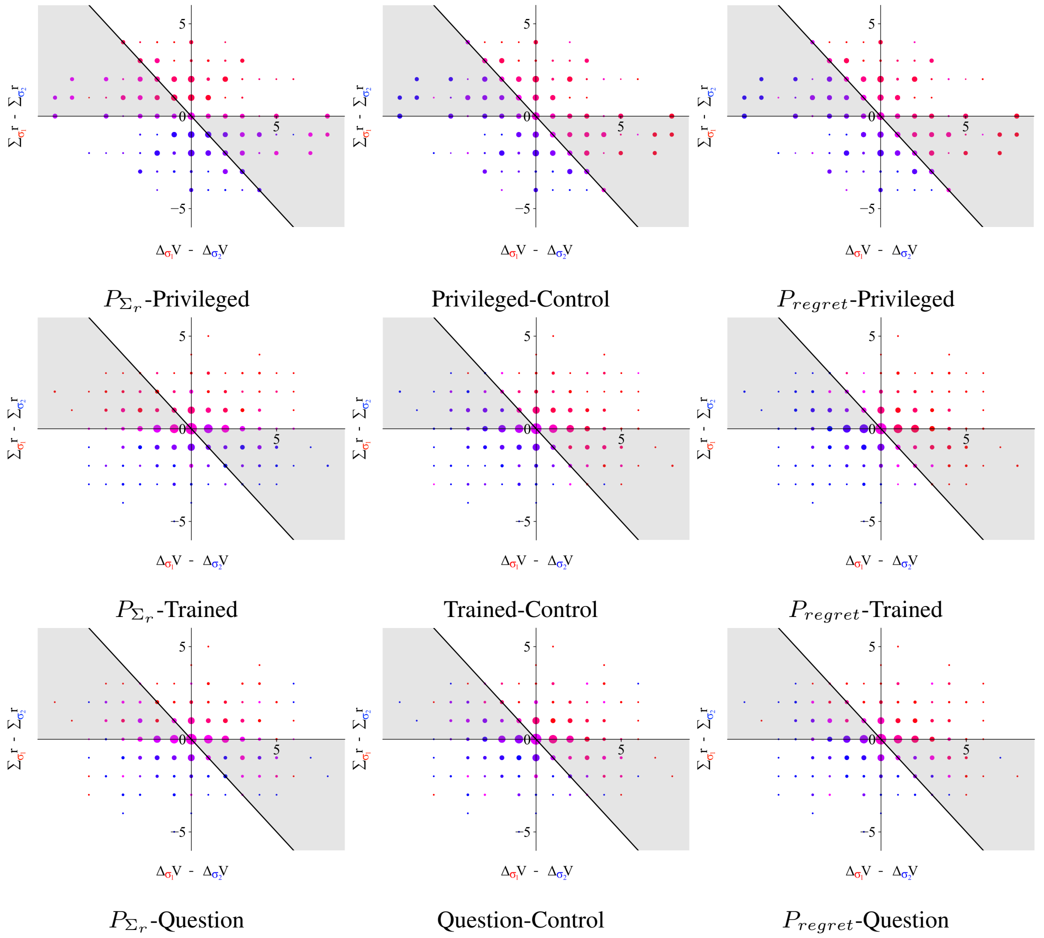
F.1 Choosing the preference elicitation question

The authors of this paper were asked to propose possible questions to ask subjects during preference elicitation that might influence their preferences toward either preference model. The first author selected the three most appealing options for each preference model, each author ranked these options in order of desirability, and then ranked-choice voting was employed to select the winner. For guiding human preferences towards the regret preference model the three options ranked by each author were “Which path shows better decision-making?”, “Which path reflects better decision-making?, and “Which path is more likely to be taken by an expert?”. For guiding human preferences towards the partial return preference model, the three options were “Which path would be better if the task ended after the path?”, “Which path has better immediate outcomes?”, and “Which path looks better, considering only exactly what happened during the path?”. After ranking these questions by their likely ability to guide human preferences towards the regret or partial return preference models, the second question won 80% and 60% of the time respectively.

F.2 Number of preferences per collected dataset

Table 3 displays the number of preferences collected for each condition in the Question experiment after discarding samples where a subject chose “Can’t Tell” instead of a preference.

G Correlations between preferences and segment statistics



P_{Σ_r}-Privileged

Privileged-Control

P_{regret}-Privileged

P_{Σ_r}-Trained

Trained-Control

P_{regret}-Trained

P_{Σ_r}-Question

Question-Control

P_{regret}-Question

Figure 19: Proportions of subjects preferring each segment in a pair, plotted by the difference in segments' change in state values (x-axis) and partial returns (y-axis). The areas of the circles are proportional to the number of segment pairs at that point. The proportionality is consistent across across all plots for the Trained and Question experiments, and separately for the Privileged experiments. The diagonal line indicates points of indifference for P_{regret}, while indifference points for P_{Σ_r} are on the x-axis. The shaded gray area highlights where the partial return and regret preference models disagree, each preferring a different segment. To visually assess which preference model better fits the data: if subjects used the partial return preference model to generate preferences, the color gradient would be orthogonal to the x-axis. Conversely, if they followed the regret preference model, the gradient would be orthogonal to the diagonal line, as regret here is $x + y$.

Recall that we compute the regret of a trajectory segment with deterministic transitions as follows: $regret_d(\sigma|\tilde{r}) = V^*_{\tilde{r}}(s^\sigma_0) - (\Sigma_\sigma \tilde{r} + V^*_{\tilde{r}}(s^\sigma_{|\sigma|}))$, where one of the 3 components of regret is partial return, $\Sigma_\sigma \tilde{r}$. We combine two components of $regret_{\text{d}}(\sigma|r)$ to simplify analysis, introducing the following shorthand: $\Delta_\sigma V_{\tilde{r}} \triangleq V^*_{\tilde{r}}(s^\sigma_{|\sigma|}) - V^*_{\tilde{r}}(s^\sigma_0)$.

The change in state value, $\Delta_\sigma V_{\tilde{r}}$, should have a greater effect on human preferences that are more aligned with the regret preference model, and partial return should have a greater effect on human preferences more aligned with the partial return preference model. The datasets of preferences are visualized in Figure 19. Note that on the

Table 3: The number of preferences in each preference dataset resulting from the Question experiment.

Condition	Number of Preferences in Dataset
P_{Σ_r}-Question	437
Question-Control	434
P_{regret}-Question	442

diagonal line in Figure 19, $regret_{\mathtt{d}}(\sigma_2|r) = regret_{\mathtt{d}}(\sigma_1|r)$, making the P_{regret} preference model indifferent.

Figure 19 shows that, when influencing human preferences towards the regret preference model in the Privileged and Trained experiments, $\Delta_\sigma V_{\tilde{r}}$ has influence on the resulting preference dataset independent of partial return. This is evident for the P_{regret}-Privileged and P_{regret}-Trained condition's dataset plots when focusing only on points at a chosen y-axis value; if the colors along the corresponding horizontal line reddens as the x-axis value increases, then $\Delta_\sigma V_{\tilde{r}}$ appears to have independent influence. Visually, $\Delta_\sigma V_{\tilde{r}}$ also exhibits independent influence on the preference datasets from the control condition for all experiments. When influencing human preferences towards the partial return preference model in the Privileged and Trained experiments, $\Delta_\sigma V_{\tilde{r}}$, has significantly less influence on the resulting preference dataset as evident by the x-axis–rather than the diagonal line–partitioning most red and blue points in the P_{Σ_r}-Privileged and P_{Σ_r}-Trained condition's dataset plots.

Visual inspects leads us to conclude that $\Delta_\sigma V_{\tilde{r}}$ independently influences human preferences for all conditions in the Question experiment, indicating that regardless of the preference elicitation question, the change in state value is still correlated with how subjects label preferences.

H Likelihood of preference dataset

When we evaluate the likelihood of a preference dataset given a preference model (under the assumption that it follows a Boltzmann distribution), we seek to evaluate which class of preference model can better express the human data, given equivalent versions of the ground-truth reward function that was taught to human subjects prior to preference elicitation. Specifically, we note that reward functions that differ only by a constant scaling factor are equivalent under most definitions—including how they order policies given a start state distribution and, by consequence, their sets of optimal policies—and different scalings of the same ground-truth reward function are considered an equivalence class. Concretely, to evaluate the likelihood of a dataset given a preference model class and this equivalence class that includes the ground-truth reward function, we use the highest likelihood across a predefined list of positive scaling parameters, each of which multiplies the output of the ground-truth reward function. This scaling parameter can also be seen as scaling the difference in the two segment statistics and it therefore affects entropy of the probabilities given to preferring each segment, pushing them closer to 0.5 or to 0 and 1. Mathematically, it has the same effect as using a Boltzmann temperature parameter, making such a temperature parameter redundant in most settings and therefore not part of our standard description of the preference models. The predefined list of scaling parameters was chosen to cover the space in which these preference models have relatively high likelihoods. Alternatively, this scaling parameter could be learned via gradient descent and tested on a heldout set, like in k-fold cross validation, but we decided against this approach out of concerns that learning a scaling parameter is not representative of actually learning a reward function, where the reward function has unknown parameters beyond its scale.

For each preference dataset resulting from each experimental condition, we evaluate how well P_{regret} and P_{Σ_r} predict the dataset. We explore a range of possible reward scaling parameters for P_{regret} and P_{Σ_r}, computing the mean cross-entropy loss for each parameter and model over the dataset. The reward scaling parameters were selected to be exponentially spaced between approximately 1 and -1. The n-th reward scaling parameter is given by $p_n = ar^{n-1}$. We used 25 reward scaling parameters: the first 12 were generated with $a = 0.01$ and $r = 1.236$, the next 12 with $a = -0.01$ and $r = 1.236$, and the final parameter was set to 0.

The plotted losses for each reward scaling parameter are illustrated in Figure 20 for the preference datasets obtained from the Privileged experiment, Figure 21 for the Trained experiment, and Figure 22 for the Question experiment. Note that when fitting the regret preference model to a preference dataset for these plots, we apply the scaling parameter to the negated regret of a segment for easier visual comparison. These plots extend the results shown in Figures 5, 7, and 9. One incorrect conclusion to draw from those figures is that the proposed interventions simply train humans to better understand the ground-truth reward function, rather than to follow a specific preference model. We acknowledge the possibility of entangled effects relating to learning more about the ground-truth reward function rather than a specific preference model. However, we expect that in our experiments such effects are relatively minimal. Firstly, all human subjects already had a good understanding of

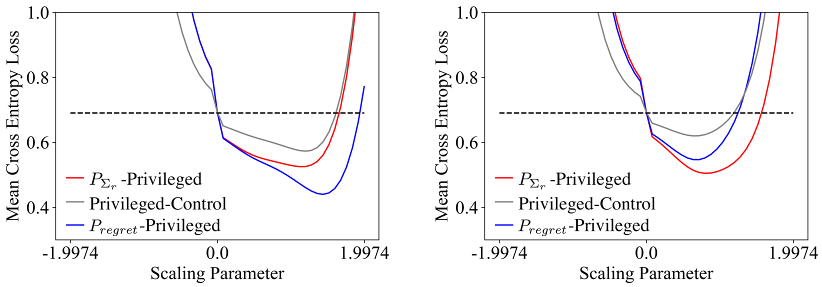


Figure 20: The mean cross entropy loss over each preference dataset resulting from the Privileged experiment (see Section 5.1) for all scaling parameters given the regret preference model (Left) and the partial return preference model (Right). If the loss is **lower** for the dataset of preferences influenced towards the target preference model than the control condition's dataset at a specific scaling parameter, it means the former is better predicted by—and more likely under—the target preference model given that scaling parameter. The regret preference model achieves the lowest mean cross-entropy loss over—and is therefore most predictive of—the P_{regret}-Trained dataset for all scaling parameters greater than 0. Similarly, the partial return preference model best predicts the P_{Σ_r}-Trained dataset for all scaling parameters greater than 0. This supports our hypothesis that showing subjects privileged information about each segment's regret or partial return during preference elicitation does influence their preferences towards a specific preference model.

the ground-truth reward function; we employed a comprehension test to filter out subjects who did not (see Appendix C.4). Further, in Figures 5, 7, and 9, we see that for all experiments the loss over a condition's dataset is lower under the target preference model than all other conditions datasets under the same preference model. Had one condition trivially resulted in subjects better understanding the ground-truth reward function rather than the target preference model, we would expect to see that condition's dataset induce the lowest loss under either preference model. Figures 5, 7, and 9 illustrate that this is not the case.

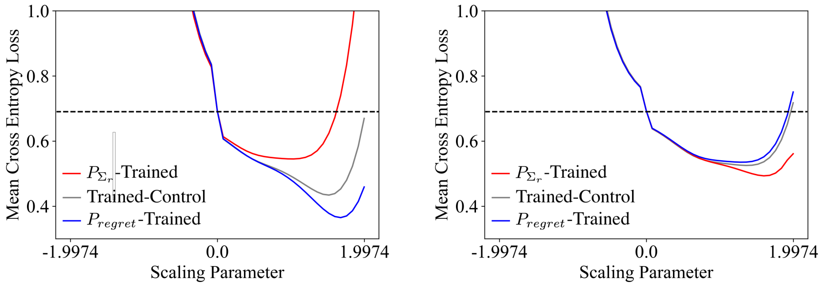


Figure 21: The mean cross entropy loss over each preference dataset resulting from the Trained experiment (see Section 5.2) for all scaling parameters given the regret preference model (Left) and the partial return preference model (Right). See Figure 20 for details on how to interpret this graph. For scaling parameters greater than 0, the regret preference model achieves the lowest mean cross-entropy loss over—and is therefore most predictive of—the P_{regret}-Trained dataset, followed by the Trained-Control dataset, and finally the P_{Σ_r}-Trained dataset. Similarly, the partial return preference model best predicts the P_{Σ_r}-Trained dataset followed by the P_{regret}-Trained and Trained-Control condition's datasets. This supports our hypothesis that teaching subjects about a specific preference model does influence their preferences towards that model.

We also seek to test whether there is a difference in the likelihood of the control condition's dataset and the likelihood of the dataset that arises from influencing humans towards a specific preference model. The dataset of preferences in the Privileged experiment is unpaired so we perform a Mann-Whitney U test, while the dataset of preferences in the Trained and Question experiments are paired so we perform a Wilcoxon paired signed-rank test. See Appendix C.2 for more details on how these datasets were constructed. All statistical tests are applied between the likelihoods over each dataset

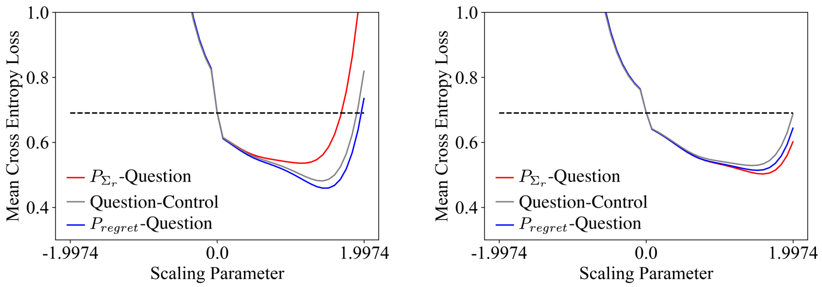


Figure 22: The mean cross entropy loss over each preference dataset resulting from the Question experiment (see Section 5.3) for all scaling parameters given the regret preference model (Left) and the partial return preference model (Right). See Figure 20 for details on how to interpret this graph. The regret preference model achieves the lowest mean cross-entropy loss over the P_{regret}-Question dataset, closely followed by the Question-Control dataset, for all scaling parameters greater than 0. The partial return preference model best predicts the P_{Σ_r}-Question condition's dataset, closely followed by the P_{regret}-Question dataset and then the Question-Control dataset for all scaling parameters greater than 0. These results suggest that changing the preference elicitation question to influence preferences towards the regret preference model may not be effective, though the loss over the Question-Control dataset given the regret preference model is already relatively low. Modifying the question to guide preferences towards the partial return preference model has a moderate effect on the datasets conformity to the preference model.

that result from the best scaling parameter—meaning the scaling parameter that induces the lowest mean cross-entropy loss.

Performing the Mann-Whitney U test between P_{Σ_r}-Privileged and Privileged-Control datasets results in $U = 1418607.0$, $p < 0.01$, and between the P_{regret}-Privileged and Privileged-Control datasets results in $U = 1643392.5$, $p < 0.01$. In the Trained and Question experiments, to ensure that each condition contains the same segment pairs, we removed a segment pair from all conditions if any subject selected "Can't Tell" instead of indicating a preference for that pair. Performing the Wilcoxon test between the P_{Σ_r}-Trained and Trained-Control datasets results in $W = 6366.0$, $p < 0.01$, and between the P_{regret}-Trained and Trained-Control datasets results in $W = 12083.0$, $p < 0.01$. Additionally, performing the Wilcoxon test between the P_{Σ_r}-Question and Question-Control datasets results in $W = 7321.0$, $p < 0.05$—which we consider statistically significant—and between the P_{regret}-Question and Question-Control datasets results in $W = 2217.5$, $p = 0.685$—which we do not consider statistically significant.

I Accuracy over preference dataset

Computing the likelihood of a preference dataset given a preference model is an informative measure of how well that preference model describes the dataset. But, computing that likelihood also requires preference model P—defined in Equation 2 for partial return and 4 for regret—which rests on the assumption that humans are Boltzmann-rational as instantiated via the logistic function. Therefore, to circumvent this assumption, we also compute the accuracy of the *noiseless* version of a given preference model over each dataset. These results are detailed below for all experiments.

We test the significance of these results using the Fisher's exact test (Upton, 1992). When executing the Fisher's exact test, for each condition in each experiment we construct a 2x2 contingency table where the first row is the number of preferences the noiseless target preference model classified correctly, the second row is the number of preferences the noiseless target preference model classified incorrectly, the first column is the dataset where subjects are influenced towards the target preference model, and the second column is the control condition's dataset.

Privileged Experiment Table 4 shows that, consistent with the results in Section 5.1, the noiseless version of the target preference preference model achieves higher accuracy on the preference dataset influenced toward the target model compared to the control condition dataset.

Table 6: The accuracy of each condition's preference dataset from the Question experiment with respect to the *noiseless* version of the target preference model. Higher is better. See Table 4 for more details on how to interpret this table.

Condition	Noiseless-P_{Σ_r} Accuracy	Noiseless-P_{regret} Accuracy
Control Condition	54.8%	69.9%
Influenced Towards Target Model	65.2%	68.7%

We conduct a Fisher exact test (Upton, 1992) to determine whether there is a significant difference in the proportion of preferences that the noiseless target preference model correctly classifies between the influenced preference dataset and the control condition dataset. We find a p-value of less than 0.001 when comparing the P_{Σ_r}-Privileged dataset to the Privileged-Control dataset, as well as when comparing the P_{regret}-Privileged dataset to the Privileged-Control dataset.

Trained Experiment Table 5 presents the accuracy of the noiseless target preference model when predicting the P_{Σ_r}-Trained and P_{regret}-Trained datasets compared to the Trained-Control dataset. The accuracy over both the P_{Σ_r}-Trained and P_{regret}-Trained datasets is notably higher than the accuracy over the Trained-Control dataset given the respective preference model, supporting the results in Section 5.2.

Table 4: The accuracy of each condition's preference dataset from the Privileged experiment with respect to the *noiseless* version of the target preference model. If the accuracy is **higher** for the dataset of preferences influenced towards the target preference model than the control condition's dataset, it means the former is better predicted by the noiseless version of the target preference model.

Condition	Noiseless-P_{Σ_r} Accuracy	Noiseless-P_{regret} Accuracy
Control Condition	48.6%	55.9%
Influenced Towards Target Model	75.3%	75.8%

We conduct the Fisher exact test over the proportion of preferences predicted correctly by the noiseless target preference model; we find a p-value of 0.0012 when comparing the P_{Σ_r}-Trained dataset to the Trained-Control dataset, and of 0.0025 when comparing the P_{regret}-Trained dataset to the Trained-Control dataset.

Question Experiment As shown in Table 6, the accuracy of the noiseless partial return preference model is higher over the P_{Σ_r}-Question dataset than the Question-Control dataset. This indicates that changing the preference elicitation instruction to influence preferences towards partial return results in a preference dataset that is better predicted by partial return. The P_{regret}-Question dataset, on the other hand, is not better predicted by regret than the control condition.

Table 5: The accuracy of each condition's preference dataset from the Trained experiment with respect to the *noiseless* version of the target preference model. Higher is better. See Table 4 for more details on how to interpret this table.

Condition	Noiseless-P_{Σ_r} Accuracy	Noiseless-P_{regret} Accuracy
Control Condition	53.7%	68.8%
Influenced Towards Target Model	75.5%	75.4%

We conduct Fisher's exact test and find a p-value of 0.0157 when comparing the proportion of preferences predicted correctly by the noiseless partial return preference model for the P_{Σ_r}-Trained dataset versus the Trained-Control dataset, indicating statistical significance. We do not find a statistically-significant p-value when comparing the P_{regret}-Trained dataset to the Trained-Control dataset ($p = 0.7144$).

J LEARNING REWARD FUNCTIONS FROM PREFERENCES

J.1 DESIGN PATTERN FOR LEARNING A REWARD FUNCTION FROM PREFERENCES

We follow the general procedure for learning a reward function from a dataset of preferences depicted in Figure 23. This procedure is executed for all preference datasets in each experiment, which all share the same ground-truth reward function r.

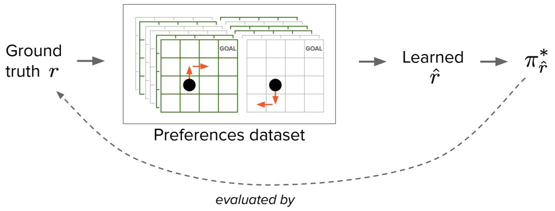


Figure 23: An outline of the general procedure for learning a reward function from a preference dataset and then evaluating that reward function. The generic gridworld shown is for illustrative purposes only. Figure provided by Knox et al. (2024).

J.2 ADDITIONAL DETAILS FOR LEARNING A REWARD FUNCTION FROM PREFERENCES

Doubling the preference dataset by reversing preference samples When learning a reward function from a preference dataset, we double the amount of data by duplicating each preference sample and then flipping the preference label and segment pair ordering. This provides more training data and avoids learning any segment ordering effects.

The reward representation The ground-truth reward function r is assumed to be a linear combination of weights and features. Any reward function learned from a preference dataset $\hat{r}$ takes the same form. This linearity assumption enables us to use the tractable algorithm for learning a reward function with P_{regret} proposed by Knox et al. (2022).

Discounting during value iteration The delivery domain is an episodic environment but a policy derived from a poorly learned reward function can endlessly avoid terminal states, resulting in a return of negative infinity. Therefore during value iteration and when computing a policy's mean return with respect to r, we apply a discount factor of $\gamma = 0.999$. We chose this high discount factor to avoid returns of negative infinity while having a negligible effect on the returns of high-performing policies and still allowing value iteration to converge within a reasonable time.

Early stopping when learning with P_{regret} Knox et al. (2022) found that when learning a reward function using P_{regret}, the training loss tended to fluctuate cyclically. To handle this, they use the $\hat{r}$ that achieved the lowest loss during training instead of the final $\hat{r}$. We follow the same procedure.

J.3 HYPERPARAMETERS FOR LEARNING A REWARD FUNCTION FROM PREFERENCES

The following hyperparameters were used by Knox et al. (2024) and across all our experiments. See Knox et al. (2024) for more details on how they were chosen.

Reward learning with the partial return preference model
learning rate: 2; number of training epochs: $30,000$; and optimizer: Adam (with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and eps= $1e-08$).

Reward learning with the regret preference model
learning rate: 0.5; number of training epochs: $5,000$; optimizer: Adam (with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and eps=$1e-08$); and softmax temperature: 0.001.

Additionally, learning with P_{regret} following the algorithm proposed by Knox et al. (2024) requires a set of successor features from candidate policies which are used to approximate $V^*_{\hat{r}}(.)$, a component of the regret preference model. Because we use the same delivery ask as Knox et al. (2024), we use the set of successor features that they generate.

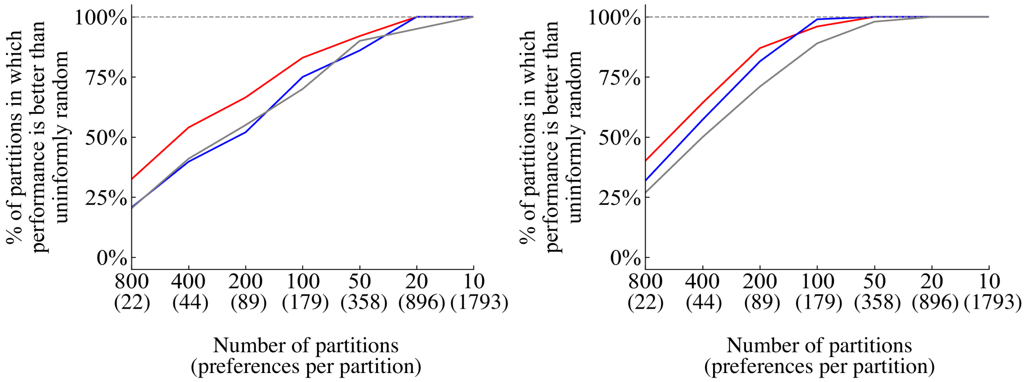


Figure 24: Learning a reward function with the partial return preference model (Left) and regret preference model (Right) from the preferences collected in the Privileged experiment. This figure complements Figure 6.

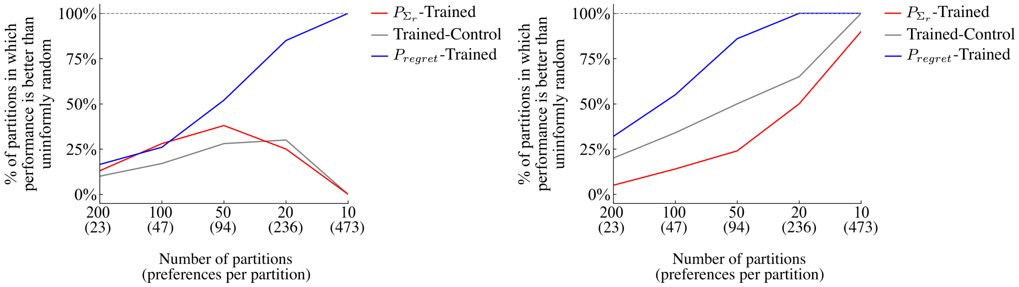


Figure 25: Learning a reward function with the partial return preference model (Left) and regret preference model (Right) from the preferences collected in the Trained experiment. This figure complements Figure 8.

When generating Figures 6, 8, 10, 25, and 26, we use random seeds $1 - 10$. The human preference datasets contain varying numbers of preferences (see Table 2). For all datasets within an experiment, we randomly subsample preferences to match the size of the smallest dataset—1793 preferences for the Privileged experiment datasets, 473 preferences for the Trained experiment datasets and 434 preferences for the Question experiment datasets. This allows for easier comparison when partitioning the resulting datasets.

J.4 Computer specifications and software libraries used

The computer used to run all experiments had the following specification. Processor: 2.8 GHz Quad-Core Intel Core i7; Memory: 16 GB. Pytorch 2.0.1 (Paszke et al., 2019) was used to implement all reward learning models, and statistical analyses were performed using Scikit-learn 1.3.0 (Pedregosa et al., 2011).

J.5 Additional results for learning reward functions

Figure 24 complements Figure 6, Figure 25 complements Figure 8, and Figure 26 complements Figure 10, showing the percentage of partitions where the learned reward functions results in better performance than a policy that selects actions uniformly. In general, for each partition size, ranking each preference dataset by the percentage of better-than-random performance induced by the learned reward functions produces the same order as when using near-optimal performance.

J.6 Addressing the partial return preference model's identifiablity issues

Learning with the partial return preference model from the P_{Σ_r}-Trained and Trained-Control datasets often fails to recover a reward function that induces near-optimal performance (see Figure 8). Knox et al. (2022) demonstrated that, in this grid-world domain, learning with the partial

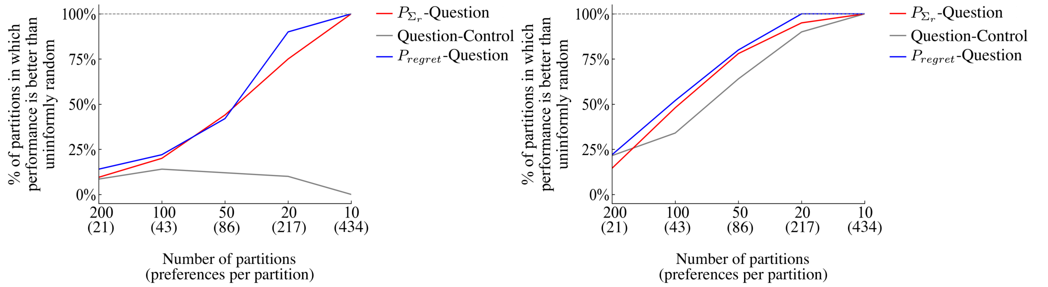


Figure 26: Learning a reward function with the partial return preference model (Left) and regret preference model (Right) from the preferences collected in the Question experiment. This figure complements Figure 10.

return preference model requires preference labels over pairs of trajectory segments in which one segment terminates earlier than the other at both the positive and negative terminal states (i.e., the inverted teardrop and sheep in Figure **??**). Such segment pairs help mitigate the identifiability issue of the partial return preference model related to a constant shift in the reward function. To address this issue, Knox et al. (2022) manually selected types of segment pairs for preference labeling. We follow their segment pair selection methodology for the Privileged experiment, but for the Trained and Question experiments, trajectory segments were constructed by randomly sampling actions to emulate a more realistic preference elicitation procedure. Consequently the partial return preference model may not recover the ground-truth reward function from the resulting preference datasets, which we hypothesize as explaining the partial return preference model's poor performance when learning from the P_{Σ_r}-Trained and Trained-Control datasets in Figure 8. Knox et al. (2024) showed that using preferences generated by P_{regret} to learn a reward function with the partial return preference model results in a reward function that is equivalent to an optimal advantage function, which may explain why the partial return preference model recovers performant reward functions from the P_{regret}-Trained dataset. We leave an investigation into this hypothesis to future work.

To empirically test whether the absence of specific segment pairs contributed to the poor performance of the partial return preference model when learning reward functions from the P_{Σ_r}-Trained and Trained-Control datasets, we added 50 additional segment pairs to each dataset. In these additional segment pairs, one segment terminates at the positive terminal state in fewer than three time-steps while the other segment does not terminate after three time-steps. These segment pairs would appear in the right-most graph in Figure 14, and were assigned preference labels by the partial return preference model with the ground-truth reward function. Figure 27 present the results when learning reward functions using these additional *synthetic* preferences.

Learning with the regret preference model using these additional preferences (bottom row of Figure 27) induces comparable results to those in Figure 8 and Figure 25 which matches our expectations; the regret preference model does not suffer from the same identifiability issues as the partial return preference model. Learning with the partial return preference model when including the additional preferences (top row of Figure 27) results in reward functions that induce near-optimal behavior more often for all datasets across all partition sizes. Including these preferences also results in better-than-uniformly-random behavior significantly more often across all partition sizes and datasets. These results therefore support our hypothesis that the partial return preference model's poor performance when learning from the P_{Σ_r}-Trained and Trained-Control datasets is, at least in part, due to the datasets missing specific segment pairs that account for the partial return preference model's identifiability issues.

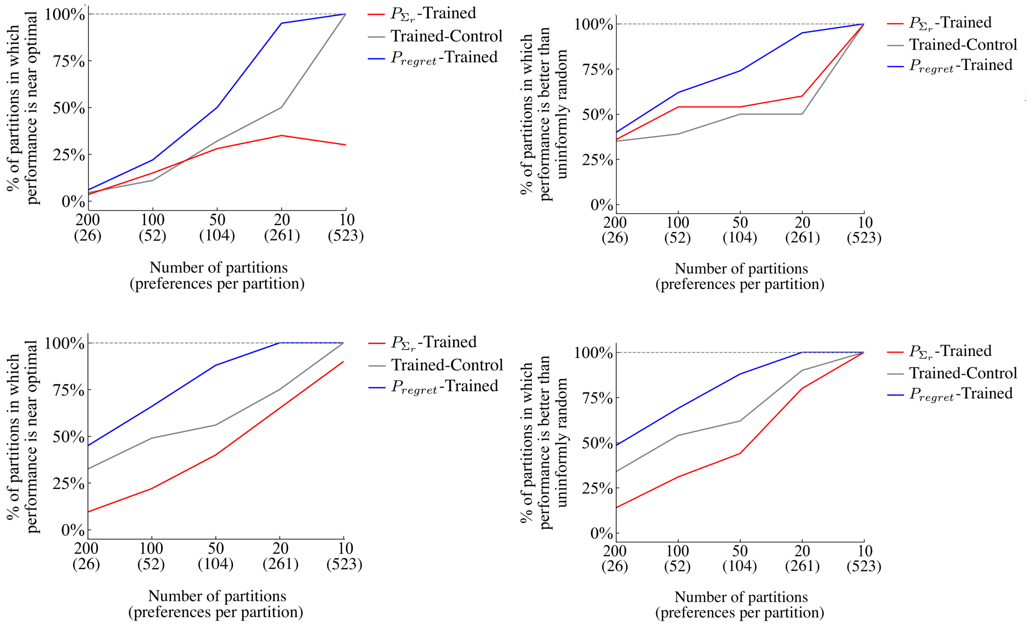


Figure 27: Learning a reward function with the partial return preference model (Top Row) and regret preference model (Bottom Row). The Training experiment's preference datasets are partitioned following the same methodology as when generating Figure 8. Each preference dataset contains 50 additional segment pairs that aim to compensate for the identifaibility issues of the partial return preference model.