Can You Rely on Synthetic Labellers in Preference-Based Reinforcement Learning? It's Complicated

Katherine Metcalf, Miguel Sarabia, Masha Fedzechkina, Barry-John Theobald
Apple
kmetcalf, miguelsdc, mfedzechkina, bjtheobald@apple.com

Abstract

Preference-based Reinforcement Learning (PbRL) enables non-experts to train Reinforcement Learning models using preference feedback. However, the effort required to collect preference labels from real humans means that PbRL research primarily relies on synthetic labellers. We validate the most common synthetic labelling strategy by comparing against labels collected from a crowd of humans on three Deep Mind Control (DMC) suite tasks: stand, walk, and run. We find that: (1) the synthetic labels are a good proxy for real humans under some circumstances, (2) strong preference label agreement between human and synthetic labels is not necessary for similar policy performance, (3) policy performance is higher at the start of training from human feedback and is higher at the end of training from synthetic feedback, and (4) training on only examples with high levels of inter-annotator agreement does not meaningfully improve policy performance. Our results justify the use of synthetic labellers to develop and ablate PbRL methods, and provide insight into how human labelling changes over the course of policy training.

Introduction

Preference-based Reinforcement Learning (PbRL) is a human-in-the-loop method to learn from teacher preference feedback instead of requiring either a hand-engineered reward function, or examples of task success. PbRL infers a reward function from preference feedback, then iterates between training the policy and seeking additional preference feedback, as shown in Figure 1. Given the efforts associated with collecting human feedback, the majority of PbRL development relies on synthetic labellers to compare methods and ablate method components. While prior work has evaluated proposed PbRL methods with real human feedback (Christiano et al. 2017; Lee, Smith, and Abbeel 2021; Stiennon et al. 2020; Wu et al. 2021), the evaluations are cursory, are run only on the newly proposed method, do not validate their synthetic labeller results, and do not provide insight into the human labels. Therefore, it is important to establish that synthetic labels are indicative of human labels to ensure experimental PbRL algorithm success translates to practical success (Amershi et al. 2014). We address this question in current work.

Specifically, we compare synthetic and human labels, and the policies resulting from training on these labels. We also analyze trends in human preference labels via inter-annotator agreement and label quality. The main findings of this paper are:

1. Policy performance from synthetic preference labels (Christiano et al. 2017; Lee et al. 2021) aligns with policy performance from human labels until the final round of preference feedback.
2. Policy performance from human feedback is higher at the start of training and higher from synthetic feedback at the end.
3. High human and synthetic preference label agreement is not required for similar policy performance.
4. As the policy produces behaviors more similar to the target behavior, the preference labelling task becomes challenging for humans.
5. Using human preference feedback with high inter-annotator agreement does not meaningfully improve policy performance.

Taken together these findings show that we can use synthetic labels as a proxy for human labels, but that practitioners need to measure the quality of their labels to obtain best results.

Related Work

The importance of studying the human in interactive machine learning settings was investigated by Amershi et al. (2014) as their behaviors directly impact the success of the developed algorithm. However, the majority of work in PbRL relies primarily on perfect, synthetic preference labelers for their evaluations and prior work has not studied their algorithms in the context of human behavior (Lee et al. 2021; Park et al. 2022; Liu et al. 2022; Bıyık, Talati, and Sadigh 2022; Hejna III and Sadigh 2022). Studies that include experiments with human feedback do not systematically evaluate with human labels, they do not compare across algorithms (Christiano et al. 2017; Lee, Smith, and Abbeel 2021), they do not investigate human labelling behaviors (Palan et al. 2019; Wang, Wang, and Min 2022), nor do they provide algorithm comparisons for both synthetic and human labellers (Palan et al. 2019; Lee, Smith,

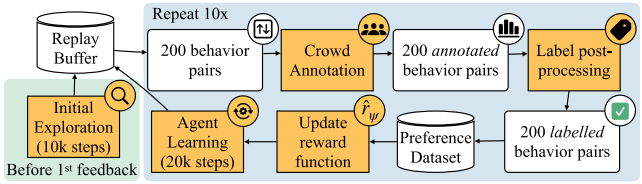


Figure 1: Schematic of the preference-based reinforcement learning loop. (1) the policy explores for some number of steps (either randomly (Christiano et al. 2017) or guided by an unsupervised exploration objective (Lee, Smith, and Abbeel 2021)). (2) trajectory pairs are selected from the buffer and presented to a teacher. The teacher uses their knowledge of the target behavior to assign preferences. (3) the labelled pairs are added to the preference dataset and the reward model is updated to be predictive of the preference labels. (4) the policy trains on the learned reward function before returning to the second step.

and Abbeel 2021). A systematic evaluation is important to understand whether performance gains hold in practice.

The prior work that has directly compared policy performance with human labels and synthetic labels suggests performance from human feedback is task dependent (Christiano et al. 2017). However, this work does not investigate what might contribute to the task-dependent performance, thus leaving unknown the conditions under which synthetic labels may be a good proxy for human labels.

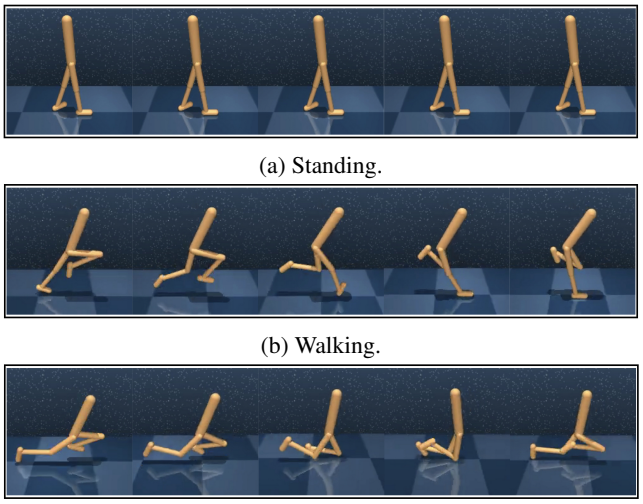
Preference-based Reinforcement Learning

An overview of PbRL is shown in Figure 1. PbRL learns a policy, π_ϕ, to maximize the expected discounted return $\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k})$, where the reward function $r(s_t, a_t)$ is inferred from preference labels over behavior pairs. The learning process iterates between training π_ϕ with the current estimate, $\hat{r}_\psi$, of the target reward function, r_ψ, and updating $\hat{r}_\psi$ using the preference labels from a teacher. The teacher is queried every K steps of policy training with m_p behavior pair queries, and assigns a preference label to each pair forming a dataset of preference triplets $(\sigma^1, \sigma^2, y_p)$, where σ_1 and σ_2 is the behavior pair and y_p is the preference label.

The behaviors selected for labelling are those most informative to the reward model according to model uncertainty or reward ensemble disagreement (Metcalf et al. 2023; Christiano et al. 2017; Lee, Smith, and Abbeel 2021; Park et al. 2022). We follow Lee et al. (Lee et al. 2021) learning an ensemble of reward models and selecting the m_p behaviors with the highest preference-label ensemble disagreement from m_b randomly selected pairs.

To learn from preference triplets, PbRL follows the Bradley-Terry model (Bradley and Terry 1952) and assumes the preferred behavior has the larger cumulative reward according to r_ψ, and the probability of σ_1 being preferred over σ_2 is given by:

$$P_\psi[\sigma^1 \succ \sigma^2] = \frac{\exp \sum_t \hat{r}_\psi(s_t^1, a_t^1)}{\sum_{i\in\{1,2\}} \exp \sum_t \hat{r}_\psi(s_t^i, a_t^i)}, \quad (1)$$



(a) Standing.

(b) Walking.

(c) Running.

Figure 2: Examples of `stand` (top), `walk` (middle), and `run` (bottom) behaviors learned from human feedback after $1,000$ pieces of preference feedback and $500,000$ policy train steps.

where s_t^i and a_t^i is the state and action from behavior i at time-step t, and $\hat{r}_\psi(s_t^i, a_t^i)$ is the estimated reward for s_t^i and a_t^i. The parameters ψ of $\hat{r}_\psi$ are optimized to minimize:

$$\mathcal{L}^\psi = -\mathbb{E}_{(\sigma^1,\sigma^2,y_p)\sim\mathcal{D}_{\text{pref}}} y_p(0) \log P_\psi[\sigma^2 \succ \sigma^1] + y_p(1) \log P_\psi[\sigma^1 \succ \sigma^2], \quad (2)$$

where $\mathcal{D}_{\text{pref}}$ is the accumulated dataset of preference feedback.

We additionally trained the reward function on an auxiliary objective using the reward learning technique introduced in REED (Metcalf et al. 2023): before selecting behavior pairs for teacher feedback and updating the reward function, environment dynamics are encoded in the reward function's state-action representation via a self-supervised objective and the transitions accumulated in the policy's replay buffer. We use the SimSiam (Chen et al. 2020) self-supervised objective and use the architecture and hyper-parameter values from REED (Metcalf et al. 2023). The hyper-parameters are provided in Table **??** in the Appendix.

Methods

We train a total of nine policies across three different task conditions (`walker-stand`, `walker-walk`, and `walker-run` DMC tasks (Tassa et al. 2018)) and different reward function conditions (from human feedback, from synthetic feedback, and hand-engineered). Two of the reward function conditions rely on learning the policies with PbRL and the hand-engineered reward function condition uses standard RL. Each policy is trained for a total of $500,000$ steps with $1,000$ labelled behavior pairs (100 pairs per feedback round) for the PbRL policies. All policies are identical for the first $10,000$ steps (1000 seeding steps and

9000 unsupervised exploration) after which the policies are updated according to task and labeller-specific reward models.

Collecting Preference Feedback and Creating the Preference Dataset

Preference feedback was collected for each task every $20,000$ policy steps over 10 rounds of feedback following prior schedules (Metcalf et al. 2023; Lee et al. 2021; Lee, Smith, and Abbeel 2021). For each feedback round we selected ($m_p = 200$) queries for labelling by the crowd rather than ($m_p = 100$) (Metcalf et al. 2023; Christiano et al. 2017; Lee, Smith, and Abbeel 2021; Lee et al. 2021) to ensure at least 100 usable crowd-labelled pairs. Only 100 behavior pairs are used to update the reward model.

Selecting behaviors for feedback The policies take random actions to populate their replay buffers for $1,000$ steps and then learn on an unsupervised exploration objective for $9,000$ steps prior to selecting behavior pairs for feedback. As no information about the tasks has yet been provided, the policies are identical and the same behaviors are labelled for all three tasks. Using the same behaviors allows us to understand the impact of task difficulty on the level of agreement in the labels provided by a crowd. Subsequent feedback rounds use different sets of behaviors as the policies are exposed to task-specific preference feedback.

Collecting Labels from the Crowd Each behavior pair was labelled by 50 participants who were provided with a description of successful task completion. For the task descriptions, we used the walk instructions from Christiano et al. (Christiano et al. 2017) and created similar instructions for stand and run. The specific instructions given to the participants can be found in the Appendix (Figure 9). Participants were asked to identify which behavior is a better example of the specified task. The participants selected from: (1) σ_1 is preferred ([1, 0]), (2) σ_2 is preferred ([0, 1]), (3) both are equally preferred ([0.5, 0.5]), and (4) "cannot tell", indicating unable to label. The label assigned to a behavior pair was the majority label, and samples for which option (4) was the majority were discarded. For cohort details and recruitment please see Appendix **Cohort Information and Recruitment Process**.

Generating the Synthetic Labels Following Lee et al. (Lee et al. 2021), the synthetic preference labels were created by assigning preference to the behavior with the larger cumulative reward, or a tie if the two behaviors had the same cumulative return, according to the task's hand engineered reward function. The synthetic labels can be thought of as *oracle* labels that are always optimal given the target reward function. Our experiments use the *oracle* labeller as it is the one most commonly used in PbRL research (Christiano et al. 2017; Lee, Smith, and Abbeel 2021; Liang et al. 2022; Park et al. 2022; Liu et al. 2022).

The Policies

To compare the effects of training with synthetic and human labels, we trained three policies for each of the three tasks using a different reward function:

SAC on the DMC task reward standard soft actor-critic (SAC) (Haarnoja et al. 2018) trained using the DMC rewards. This policy acts as a performance upper-bound.

SAC on reward function from synthetic feedback SAC trained using the reward function inferred from the synthetic preference labels. The type of policy that is usually reported in PbRL works.

SAC on reward function from crowd feedback SAC trained using the reward function inferred from the human preference labels.

Note that it is not possible to directly compare policy performance across tasks, as the stand, walk, and run tasks have different difficulty levels for SAC (e.g., stand is easy, and run is difficult). Therefore we evaluate policy performance using normalized returns, detailed in Equation 3.

All policies for stand, walk, and run are trained for the same number of steps $(500,000)$ and share the same network architecture. While the SAC hyper-parameters differ by DMC task, they are kept constant within a task. All original hyper-parameters are used and specifics are given in the Appendix Tables **??**, **??**, and **??**.

Experiments and Results

We investigate four specific questions regarding the crowd annotations and policy performance:

Q1. How does the performance of policies trained with the synthetic feedback compare to those trained with crowd feedback?

Q2. How well do human and synthetic preference labels agree?

Q3. Is human and synthetic labeller agreement necessary for aligned policy performance?

Q4. What is the relationship between inter-annotator agreement and policy performance?

The performance of each policy is evaluated according to the DMC task reward function, which is used to generate the synthetic labels and is distinct from the reward function used to train the policy. For each policy we show its learning curve, which plots the mean, and standard error from the mean return per episode $(1,000$ steps). Additionally, for policies trained with synthetic or human feedback, performance is evaluated relative to SAC performance trained with the DMC ground-truth reward function. A single score per policy is computed as the mean normalized returns over the course of policy training:

$$NR(\hat{r}_\psi) = \frac{1}{T}\sum_t \frac{r_\psi(s_t, a_t, \pi_\phi^{\hat{r}_\psi}(o_t))}{r_\psi(s_t, a_t, \pi_\phi^{r_\psi}(o_t))}, \tag{3}$$

where NR is the mean normalized returns, r_ψ is the DMC task reward for the given state s_t, and action a_t, at time t, $\pi_\phi^{r_\psi}$ is the policy trained with the DMC task reward, $\pi_\phi^{\hat{r}_\psi}$ is the policy trained with the learned reward, and T is the total number of policy training steps. The closer the normalized

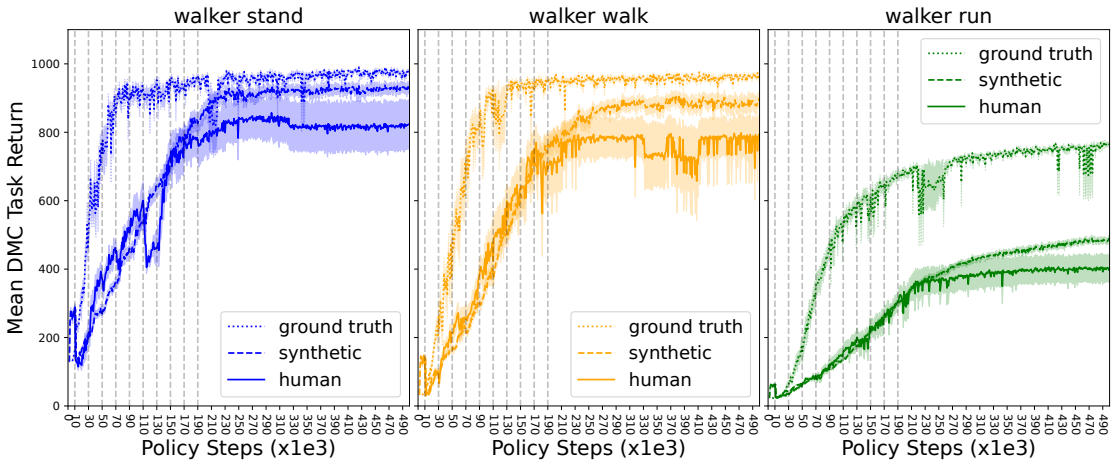


Figure 3: Policy learning curves for `stand`, `walk` and `run`. For each task, three lines are plotted: SAC from the DMC task reward, SAC from synthetic feedback, and SAC from crowd feedback. SAC on the DMC task reward shows the expected upper bound on policy performance. The mean and standard error DMC task return plotted over 10 random seeds per episode. The dashed vertical lines mark policy steps at which feedback is given.

return is to 1.0 the better a policy was able to recover ideal performance.

To better understand the crowd annotations, we use two metrics to analyze inter-annotator agreement. First, Fleiss' κ measures inter-annotator agreement (Fleiss 1971) using the interpretation of κ proposed by Landis and Koch (Landis and Koch 1977). Second, to establish the level of agreement for a single behavior pair, we use the normalized L2 similarity between the distribution of annotations and the uniform distribution representing a tie ([0.5, 0.5]):

$$\text{agreement} = \alpha \left(\frac{\|\mathbf{x}\|_2}{\|\mathbf{u}\|_2} - 1 \right), \tag{4}$$

where $\mathbf{x}$ is a *normalized* histogram of d votes: $\mathbf{x} \in [0.0, 1.0]^d$ and $\sum_{i=0}^{d} x_i = 1.0$, $\mathbf{u}$ is a uniform histogram of d dimensions $[1/d, 1/d]$), and $\alpha = \sqrt{2} - 1$ is a normalizing constant to keep the agreement between 0.0 and 1.0. The agreement is zero when the votes are uniformly distributed, and 1.0 when $\mathbf{x}$ is a one-hot histogram of votes.

After each feedback round, we assess the sensitivity of the reward model and the policy to the preference triplets in the preference dataset. We create 10 versions of the preference dataset by randomly sub-sampling 100 out of the 200 preference triplets from the human feedback. Across feedback rounds the maximum number of rejected behavior pairs (those with a majority "cannot tell" vote) was 27 resulting in at least 173 labelled trajectory pairs from which to subsample. The choice of preference triplets impacts $\hat{r}_\psi$, which in turn impacts $\pi_\phi^{\hat{r}_\psi}$.

Synthetic and Crowd End-to-End Policy Performance

We compare training a policy end-to-end on the `stand`, `walk`, and `run` tasks with crowd labels against training end-to-end with synthetic labels. Since behaviors that get selected for labelling depend on the state-actions the agent visits, which in turn depend on the current reward function, human and synthetic trained policies each will have been exposed to different datasets after the first round of feedback The policy learning curves are shown in Figure 3, the normalized returns in Table 1, and the agreement between the crowd and synthetic labels per feedback round are shown in Figure 4. **Take-away**: Policy performance given the crowd versus the synthetic labels is nearly identical until the last round of feedback (at step $190,000$), at which point performance diverges and the policies trained from the crowd under-perform those trained with the synthetic labellers.

Steps	Label	Stand	Walk	Run
0 - 190k	H	0.59 (0.12)	**0.56 (0.14)**	0.40 (0.18)
	S	**0.60 (0.19)**	0.54 (0.28)	**0.42 (0.27)**
190 - 500k	H	0.61 (0.03)	0.80 (0.03)	0.53 (0.03)
	S	**0.96** (0.03)	**0.91 (0.03)**	**0.60 (0.05)**
0 - 500k	H	0.60 (0.08)	0.71 (0.14)	0.48 (0.13)
	S	**0.82 (0.21)**	**0.77 (0.25)**	**0.53 (0.19)**

Table 1: PbRL policy normalized returns (across 10 random seeds) for the `stand`, `walk`, and `run` tasks learned for $500,000$ steps with human labels versus synthetic labels. H refers to human and S to synthetic preference feedback.

Take-away: The drop in performance after the last round of feedback suggests the crowd is not able to guide the policy as well as the synthetic labellers once the examples behaviors become similar to the target behavior, consequentially the reward functions learned from the crowd did not generalize as well to the unseen behaviors with high similarity to the target behavior, and conclusions about policy perfor-

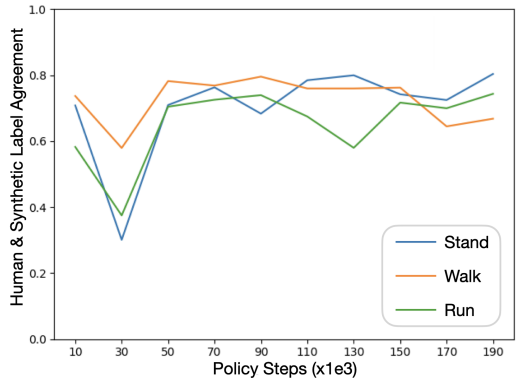


Figure 4: The degree of *label matching* between the human and synthetic labellers per feedback round for each task.

mance drawn from experiments using the proxy synthetic labels do not directly generalize to human labels.

An example for each of the `stand`, `walk`, and `run` behaviors learned from human feedback are given in Figure 2. Interestingly, the behavior for `walk` is actually a `run` as in the second frame both feet are off the ground. For the `run` behavior, it is difficult to tell if both of the agent's feet leave the ground at the same time. If both feet do not come off the ground at the same time, technically, the `walk` and `run` behaviors may be flipped.

Crowd and Synthetic Label Matching The degree of *label matching* between synthetic and crowd feedback is assessed by assigning a synthetic label to each crowd-labeled behavior pair and evaluating how frequently the human and synthetic labels are the same (e.g., both the human and synthetic label express a preference for behavior one). Figure 4 shows the changes in *label matching* over the course of policy training. In general, there is strong *label matching* between the synthetic and human preference labels indicating that the synthetic labels are a decent proxy. However, the degree of *label matching* depends on the task and the feedback round. The `walk` task has the highest *label matching* between the crowd and synthetic labels, with `run` having the second highest agreement, and `stand` the least. For both `stand` and `run` there is a large drop in human-synthetic *label matching* in the second round of feedback with a smaller drop for `walk`. There is a drop in `walk` human-synthetic *label matching* for the last two rounds of feedback, which is roughly when the policies trained from crowd versus synthetic feedback begin to diverge. In contrast, both the `run` and `stand` tasks have a slight increase in human-synthetic *label matching* and have lower standard deviation in policy performance. **Take-away**: The results further support that as policy performance approaches the target behavior, people are less able to guide the policy closer to the target behavior through preference feedback than the synthetic labeller.

Comparing Synthetic and Crowd Labels: First Feedback Round

Given the differences in policy performance with human versus synthetic labellers, we directly compare the effects

Task	Human NR	Synthetic NR	NR Difference
stand	0.66	0.54	0.12
walk	0.87	0.69	0.18
run	1.01	0.96	0.05

Table 2: Policy normalized returns (NR, cf. Equation 3) for the `stand`, `walk`, and `run` with human versus synthetic feedback. After updating each reward function, a policy was trained for $20,000$ steps, and then the normalized returns are calculated. Mean are reported for each task over 10 random seeds.

of the synthetic and crowd labels on policy performance. To that end, we evaluate the policy performance after the first round of feedback when *the behavior pairs are identical* both across tasks and labeller conditions. After the first $10,000$ policy steps, 200 behavior pairs were sent to the human crowd. For each behavior pair, the crowd labelled which behavior was a better example of `stand`, of `walk`, and of `run`. In total, 27 behavior pairs were marked as *cannot label* and were removed from each task's preference dataset. Each remaining behavior pair was then labelled by the synthetic labeller. The synthetic and crowd preference datasets were used to learn separate reward functions and subsequent policies (following Section **Preference-based Reinforcement Learning**). We repeated this process for each task, and trained for $20,000$ steps. The resulting normalized returns are shown in Table 2.

At early stages of policy training when the behaviors are mostly unstructured, the crowd feedback results in better policy performance than the synthetic feedback, as shown in columns *Human NR* and *Synthetic NR* of Table 2. Referring back to the results in Section **Synthetic and Crowd End-to-End Policy Performance**, human feedback quality decreases during final feedback rounds. **Take-away**: Human feedback outperforms synthetic feedback earlier in training when the behaviors are furthest from the target behavior and appear mostly random as opposed to later in training when the example behaviors are more similar to the target behavior.

As shown in column *NR Difference* in Table 2, the `walk` task has the highest relative policy performance gains from human feedback followed by `stand` and then `run`, an ordering which aligns with how effectively SAC trained on the DMC reward is able to learn each task. Therefore, we conclude that the "easier" the task, the higher the difference between human and synthetic feedback. **Take-away**: Tasks that are "easier" for the policy are also "easier" for people to assess behavior quality.

Next, we examine the frequency with which the crowd majority vote and synthetic labels match per behavior pair and observe more *label matching* than not between the crowd and synthetic labels with 0.71 *label matching* for `stand`, 0.75 for `walk`, and 0.58 for `run`. The greatest amount of *label matching* occurs for the `walk` followed by `stand` and then `run`. This ordering aligns the task difficulty according to SAC performance with the DMC reward

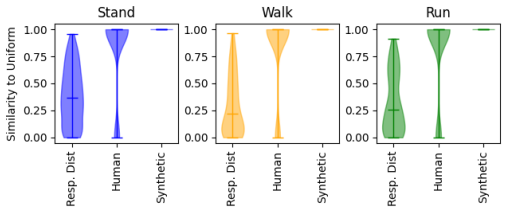


Figure 5: Violin plots displaying the distribution of human and synthetic labels from the first round of feedback. Labels are evaluated for the strength with which they indicate a strong behavior preference (1.0) versus a tie (0.0) via the normalized L2 agreement (Equation 4) between the label and the uniform distribution ([0.5, 0.5]). "Resp. Dist." indicates the degree of inter-annotator agreement and how strongly one behavior was preferred by the crowd over another.

and the difference between human and synthetic label normalized returns (see Table 2 – the lower the *label matching* the smaller the performance gap. **Take-away**: These results lead us to conclude that easier tasks have more *label matching* between crowd and synthetic labels and strong *label matching* is not necessary for policy performance to be aligned between the two label sources (e.g., `run`).

Crowd and Synthetic Label Agreement We were surprised to find a gap in policy performance between the crowd and synthetic labels. Consequently, we examine the types of labels the crowd and synthetic labellers each assign. Label types are considered to exist on a range from expressing a strong preference to expressing no preference. The votes and the labels are, therefore, scored based on how strong of a preference they convey by computing similarity to the uniform distribution using the normalized L2 agreement (Equation 4). The stronger the preference, the more dissimilar to the uniform distribution and the closer the score is to 1.0. We compare the crowd response (votes per response), the crowd majority label, and the synthetic label distribution across behavior pairs for each task (see Figure 5).

Across tasks there are large differences between the types of labels assigned by the crowd ("Human") and synthetic ("Synthetic") labeller. Unlike the crowd, the synthetic labeller never assigns a single tie and therefore provides only strong preferences. For all tasks, the crowd provides label types indicating no preference (a tie) with most labels indicating a preference. We find that the proportion of tied behavior pairs and strong preference behavior pairs is similar for all tasks. The crowd expresses a weak preference (less than 0.4) for the majority of behavior pairs. For `walk` and `run` the majority of responses do not indicate a strong preference behavior pairs and for `stand` the majority of responses have a slight preference over behavior pairs. **Take-away**: These results suggest that, for the first round of feedback, the human labels outperform the synthetic labels due to having weaker preferences, indicated by providing tie labels.

Assessing Inter-annotator Agreement and Preference Strength

Given the prevalence of weak preferences (low inter-annotator agreement) in the first round of feedback, we explore how inter-annotator agreement changes over the course of policy training. To analyze inter-annotator agreement, we examine the distribution of votes over the four possible responses the crowd participants can select from. We find that at the beginning of policy training, inter-annotator agreement is poor (Fleiss' $\kappa \approx 0.1$) and at the last round of feedback inter-annotator agreement is moderate (Fleiss' $\kappa = 0.45$) when computed across all tasks (see Figure 6). Conversely, policies trained from human feedback outperform those trained from synthetic feedback at the start of training, but not at the end.

To better understand the spread of inter-annotator agreement over the 200 behavior pairs, we make use of the L2 normalized agreement (Equation 4). Figure 7 shows that as policy training progresses, inter-annotator agreement increases. How inter-annotator agreement changes depends on the task. For example, by the eighth round of human feedback `walk` and `run` have approximately the same number of samples with low and high scores, whereas for `stand` the majority of samples have high inter-annotator agreement. The final rounds of feedback have the highest degree of inter-annotator agreement and are where human and synthetic policy performance diverge. (**Take away**) Strong inter-annotator agreement does not guarantee strong policy performance.

Inspecting Figure 8 we can see that in addition to an overall increase in inter-annotator agreement over the course of policy training, the number of ties (i.e., the density around 0.0) decreases slightly. The shift in inter-annotator agreement and prevalence of tie labels suggests that the preference selection task becomes easier over the course of policy training, possibly due to the behavior pairs more frequently containing one behavior that is a better example of the target behavior. Moreover, Figures 6 and 8 suggest the increase in agreement is driven by increases in a preference for one behavior over another (rather than an increase in ties). However, the shift in crowd labels indicates behavior preference differs between tasks. For `stand`, the crowd responses steadily shift to preferring one behavior over another, whereas for `walk` and `run`, ties and behavior preferences appear equally common in the last feedback round. Interestingly, for `walk` and `run`, there is a decrease in prevalence for tie labels until policy step $90,000$ at which point there is an increase, and then a steady decrease until the last feedback round at policy step $190,000$. This difference in ties for `walk` and `run` may be due to the task difficulty: with the DMC task reward, SAC takes longer to learn to `walk` and `run` than to `stand`. Further, we see that the number of labels indicating a tie at first decrease for all three tasks and continue to decrease for `stand`, but start to increase again after policy step $110,000$ for `walk` and `run` (that is, the bottoms of Figure 8 grow thicker).

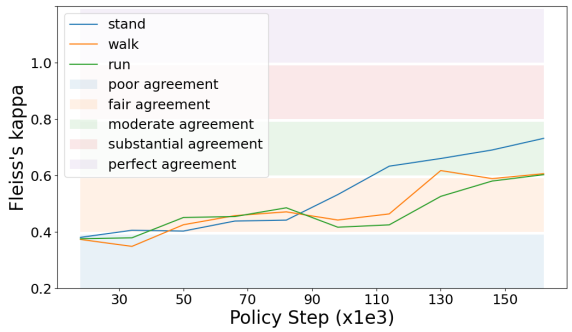


Figure 6: Fleiss' kappa for each round of feedback at 10k, 30k, 50k, 70k, 90k, 110k, 130k, 170k, and 190k policy steps for each task. Each point in the graph is the Fleiss' κ score for responses by 50 participants to 200 preference-selection samples.

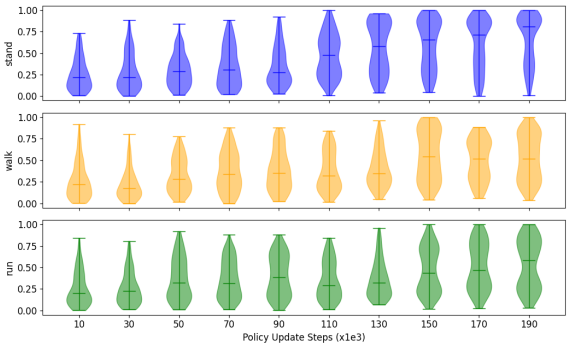


Figure 7: The distribution of inter-annotator agreement scores per round of human feedback for the stand (top), walk (middle), and run tasks (bottom). Agreement is computed for each behavior pair using the normalized L2 distance (Equation 4) between the distribution of responses and the uniform distribution. The violin plot at each round of feedback covers responses from 50 participants on 200 behavior pairs, where each response selects one of four classes.

Using Inter-Annotator Agreement to Select Reward Function Training Samples

Given the co-occurring drop in policy performance and increase in inter-annotator agreement at the end of training, we investigate whether selecting the reward function's training samples (preference triplets) using inter-annotator agreement leads to better policy performance. To this end, we train a policy for each task on the top and bottom 100 behavior pairs according to inter-annotator agreement (Equation 4).

We remark that at the first round of feedback, roughly 30 behavior pairs have moderate inter-annotator agreement (measured using Fleiss' κ) of 0.49 for stand, 0.54 for walk, and 0.49 for run. Note that as 27 behavior pairs were rejected by participants, 27 behavior pairs overlap between the 100 top and bottom behavior pairs (almost one third).

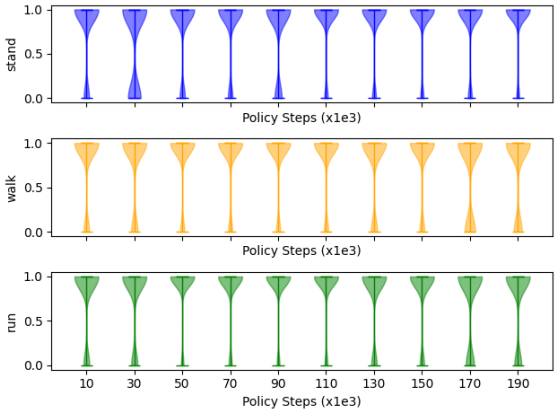


Figure 8: The spread of the crowd labels ranging from tie to a strong behavior preference over the source of policy training and feedback rounds for stand, walk, and run.

Task	Highest Agreement	Random	Lowest Agreement
stand	0.55	0.66	0.44
walk	0.56	0.87	0.49
run	1.08	1.01	0.87

Table 3: Normalized returns on policies trained after 30k with one round of reward learning where the reward was trained with 100 behavior pairs selected with the highest inter-annotator agreement (Equation 4), the lowest, or selected randomly.

Take-away: The results in Table 3 demonstrate that selecting annotations by highest inter-annotator agreement is not an optimal strategy. Random sub-sampling outperforms highest agreement sub-sampling in stand and walk, and is within a standard deviation for run. On the other hand, as expected, selecting annotations by lowest inter-annotator agreement results in clearly worse policies for all three tasks.

Conclusion and Future Work

In this work, we set out to validate the use of "perfect" synthetic preference labellers in place of human labellers during the development and evaluation of PbRL algorithms. We present evidence that human and synthetic labellers produce similar, but not identical, policy performance. Specifically, human feedback results in policies that perform better at the start of training and synthetic feedback policies that perform better at the end of training. The gap in policy performance at the end of policy training is especially concerning as it means the human guided reward functions do not capture the target behavior as well. Therefore conclusions about the performance of PbRL algorithms that rely on synthetic labellers may not generalize to human labellers. However, until the final feedback round conclusions drawn from experiments using synthetic labellers are likely to generalize.

Appendices

Human Preference Feedback Instructions

The annotator instructions are given in Figure 9.

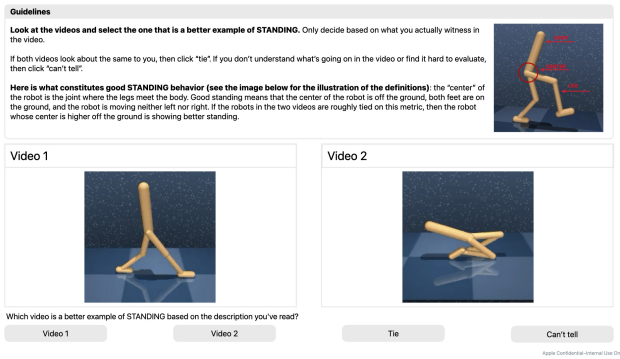
Crowd Information and Recruitment Process

The participants were recruited through an internal crowd sourcing annotation platform. Tasks were submitted to the platform specifying that the ability to read English was a requirement, the number of evaluations needed per data point, the geographical regions the tasks should be posted to, and the expected completion time for each evaluation. Participants were then able to decide whether to provide evaluations for the task, they were allowed to stop providing evaluations at any point without any penalty, and were compensated per evaluated preference pair based on the expected completion time. The platform continued to post the task until all preference pairs were evaluated the requested number of times. No individual was able to provide multiple evaluations for the same data point. Participants were not guaranteed to annotate all preference pairs in the task.

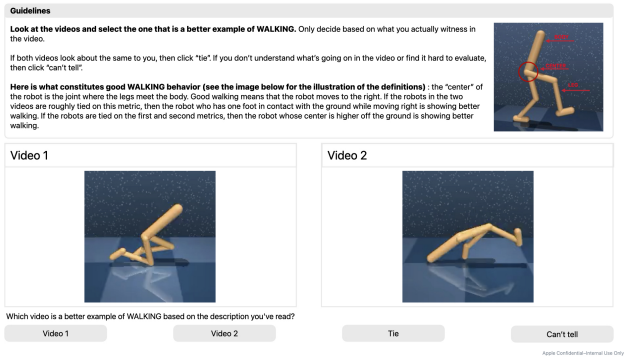
The crowd sourcing platform was designed to maintain participant anonymity. Therefore, we do not have access to any biographical data about the participants. Participants were recruited from countries spanning North American, Europe, and Asia. Participants spanned a total of 46 unique countries with a mean of 6090 (stdev=8685; min=299; max=39748) annotations per region. Evaluations were provided by 2033 unique participants with a mean of 138 (stdev=166; min=1; max=1897) annotations per participant.

Preference-based Reinforcement Learning Hyper-parameters

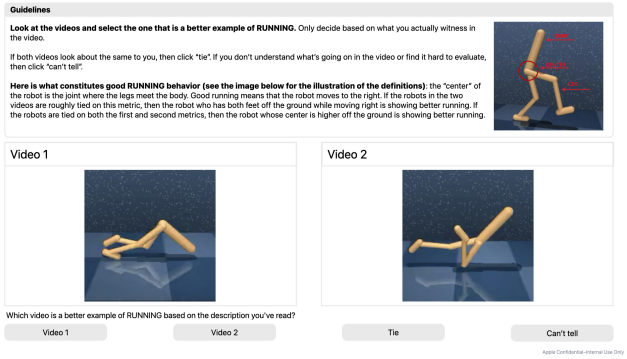
The hyper-parameters used to train the policies and the reward model follow those used in (Haarnoja et al. 2018) for SAC and (Metcalf et al. 2023) for PEBBLE.



(a) Instructions for stand.



(b) Instructions for walk.



(c) Instructions for run.

Figure 9: The instructions given to participants for the preferred behavior selection task. At the top of information how to complete the task is given (e.g., only consider behaviors you see in the video) followed by a description of what it means to complete the target behavior. Below the instructions videos of the two videos are presented side-by-side. Below the videos is the question posed to participants and the response options for participants to choose between. participants are only able to select one of the four possible responses and make their selection by clicking the response button.

References

Amershi, S.; Cakmak, M.; Knox, W. B.; and Kulesza, T. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4): 105–120.

Bıyık, E.; Talati, A.; and Sadigh, D. 2022. Aprel: A library for active preference-based reward learning algorithms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 613–617. IEEE.

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

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. volume 30.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Hejna III, D. J.; and Sadigh, D. 2022. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, 2014–2025. PMLR.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Lee, K.; Smith, L.; Dragan, A.; and Abbeel, P. 2021. B-Pref: Benchmarking Preference-Based Reinforcement Learning. *Neural Information Processing Systems (NeurIPS)*.

Lee, K.; Smith, L. M.; and Abbeel, P. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. In *International Conference on Machine Learning*, 6152–6163. PMLR.

Liang, X.; Shu, K.; Lee, K.; and Abbeel, P. 2022. Reward uncertainty for exploration in preference-based reinforcement learning.

Liu, R.; Bai, F.; Du, Y.; and Yang, Y. 2022. Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35.

Metcalf, K.; Sarabia, M.; Mackraz, N.; and Theobald, B.-J. 2023. Sample-Efficient Preference-based Reinforcement Learning with Dynamics Aware Rewards. In *Conference on Robot Learning*.

Palan, M.; Landolfi, N. C.; Shevchuk, G.; and Sadigh, D. 2019. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*.

Park, J.; Seo, Y.; Shin, J.; Lee, H.; Abbeel, P.; and Lee, K. 2022. SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning. *arXiv preprint arXiv:2203.10050*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; Casas, D. d. L.; Budden, D.; Abdolmaleki, A.; Merel, J.; Lefrancq, A.; et al. 2018. Deepmind control suite. *arXiv preprint arXiv:1801.00690*.

Wang, R.; Wang, W.; and Min, B.-C. 2022. Feedback-efficient Active Preference Learning for Socially Aware Robot Navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11336–11343. IEEE.

Wu, J.; Ouyang, L.; Ziegler, D. M.; Stiennon, N.; Lowe, R.; Leike, J.; and Christiano, P. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.