

# Leveraging Sub-optimal Data for Human-in-the-Loop Reinforcement Learning



Calarina  
Muslimani



Matthew E.  
Taylor

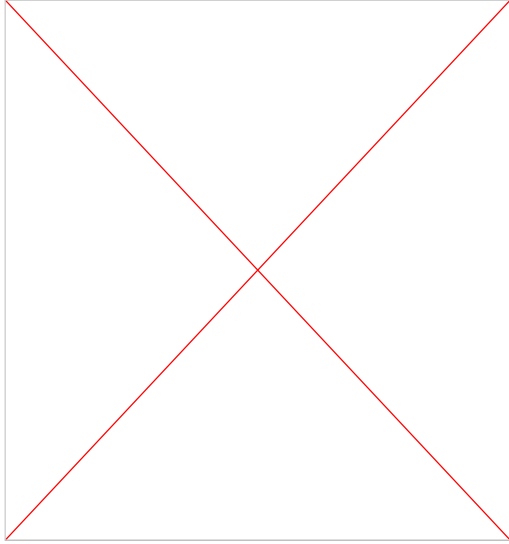
# Reinforcement learning



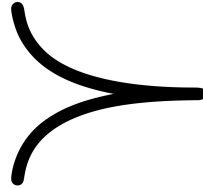
Where do rewards come from?  
Difficult to design in practice!



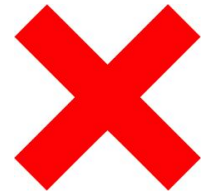
# We can learn reward functions from human feedback!



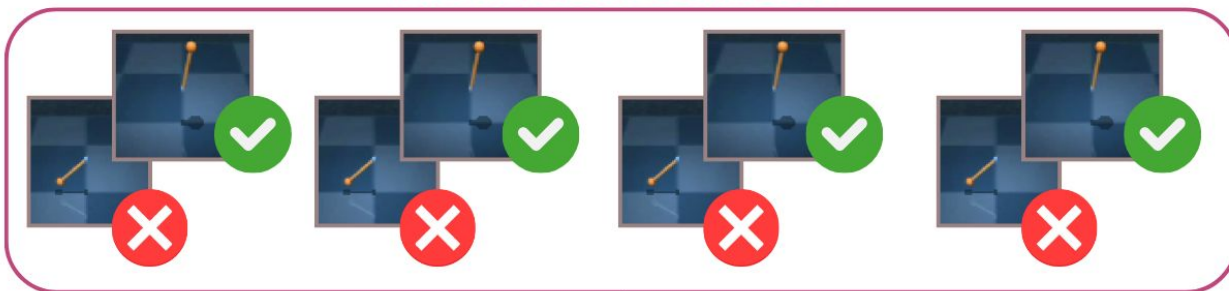
Trajectory A



Trajectory B



## Preference Data Set



Learn Reward Model

## Preference Data Set



Learn Reward Model

## Soft Max to get Probabilities

$$P_{\theta}(T^A > T^B) = \frac{\exp(\sum_t r_{\theta}(s_t^A, a_t^A))}{\exp(\sum_t r_{\theta}(s_t^A, a_t^A)) + \exp(\sum_t r_{\theta}(s_t^B, a_t^B))}$$

## Plug into Standard Binary Cross Entropy

$$L^{CE}(\theta, D) = -\mathbb{E}_{(T^A, T^B, y) \sim D} [y \log P_{\theta}(T^A > T^B) + (1 - y) \log P_{\theta}(T^B > T^A)]$$

## Preference Data Set




Requires a  
lot of human  
feedback!

$$P_{\theta}(T^A > T^B) = \frac{\exp(\sum_t r_{\theta}(s_t^A, a_t^A))}{\exp(\sum_t r_{\theta}(s_t^A, a_t^A)) + \exp(\sum_t r_{\theta}(s_t^B, a_t^B))}$$

Plug into Standard Binary Cross Entropy

$$L^{CE}(\theta, D) = -\mathbb{E}_{(T^A, T^B, y) \sim D} [y \log P_{\theta}(T^A > T^B) + (1 - y) \log P_{\theta}(T^B > T^A)]$$



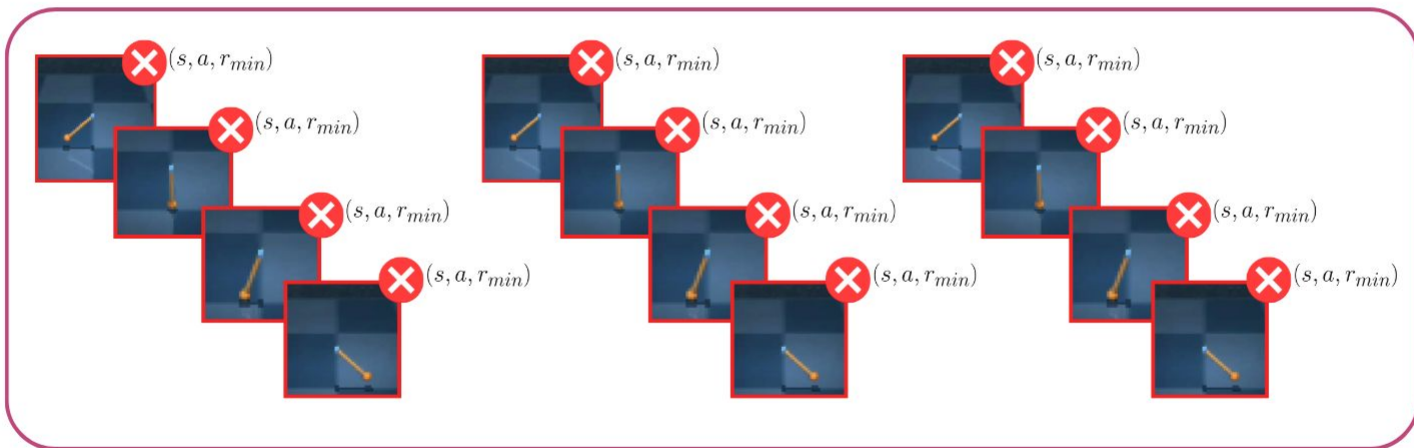
**Easy and  
cheap to  
obtain!**

**Can we leverage sub-optimal transitions  
to improve feedback efficiency of  
human-in-the-loop RL?**

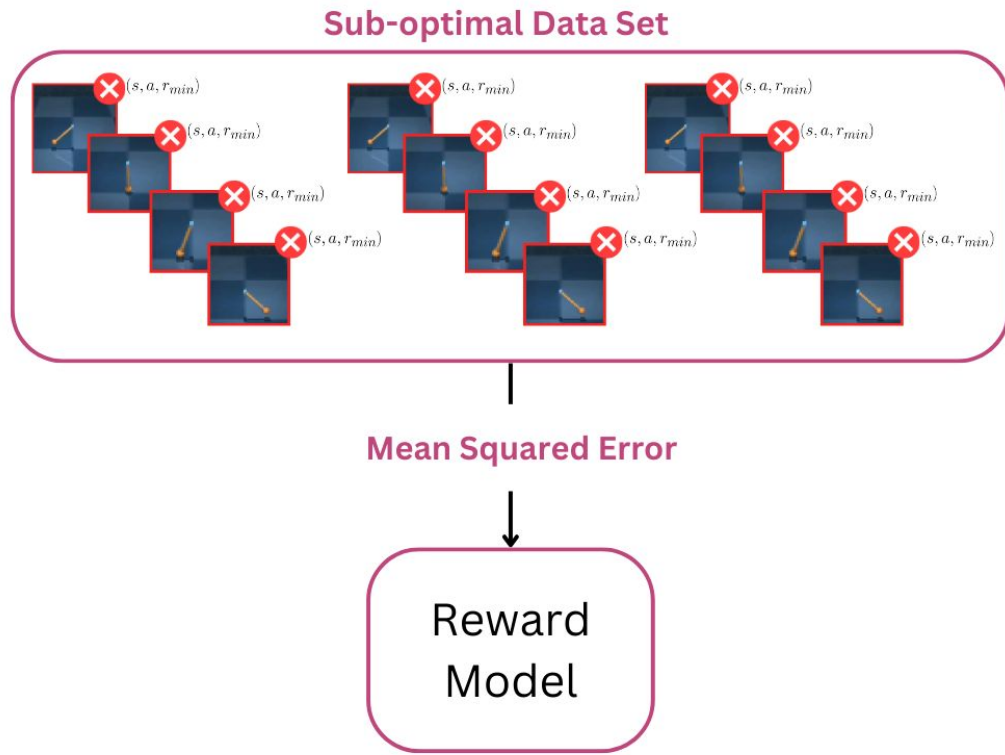


# Propose Sub-Optimal Data Pretraining–SDP

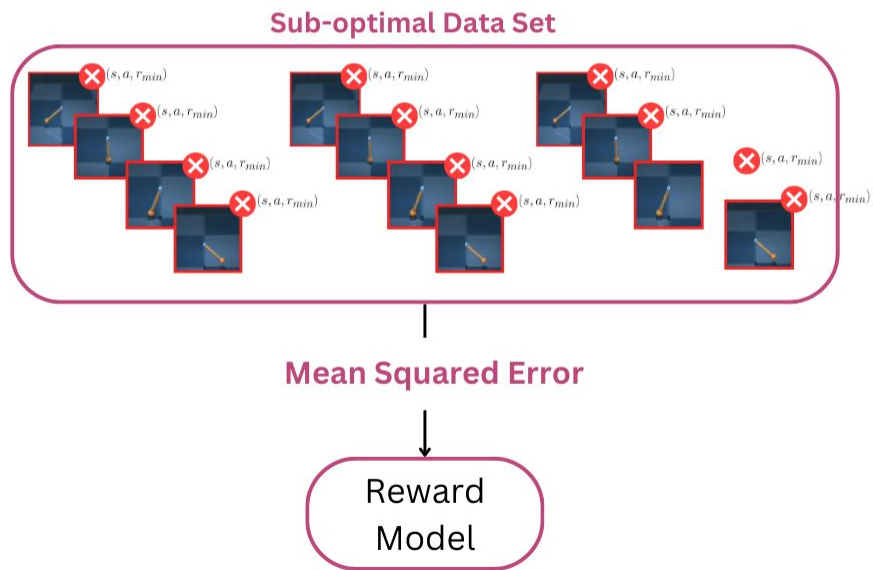
Leverages sub-optimal state, action transitions by *pseudolabeling all transitions* with *minimum* possible environment *reward*



# Simple technique of pre-training the reward model on the sub-optimal data set before applying off-the-shelf reward learning algorithms



Simple technique of pre-training the reward model on the sub-optimal data set before applying off-the-shelf reward learning algorithms



**We obtain large amounts of labeled data for “free”!**

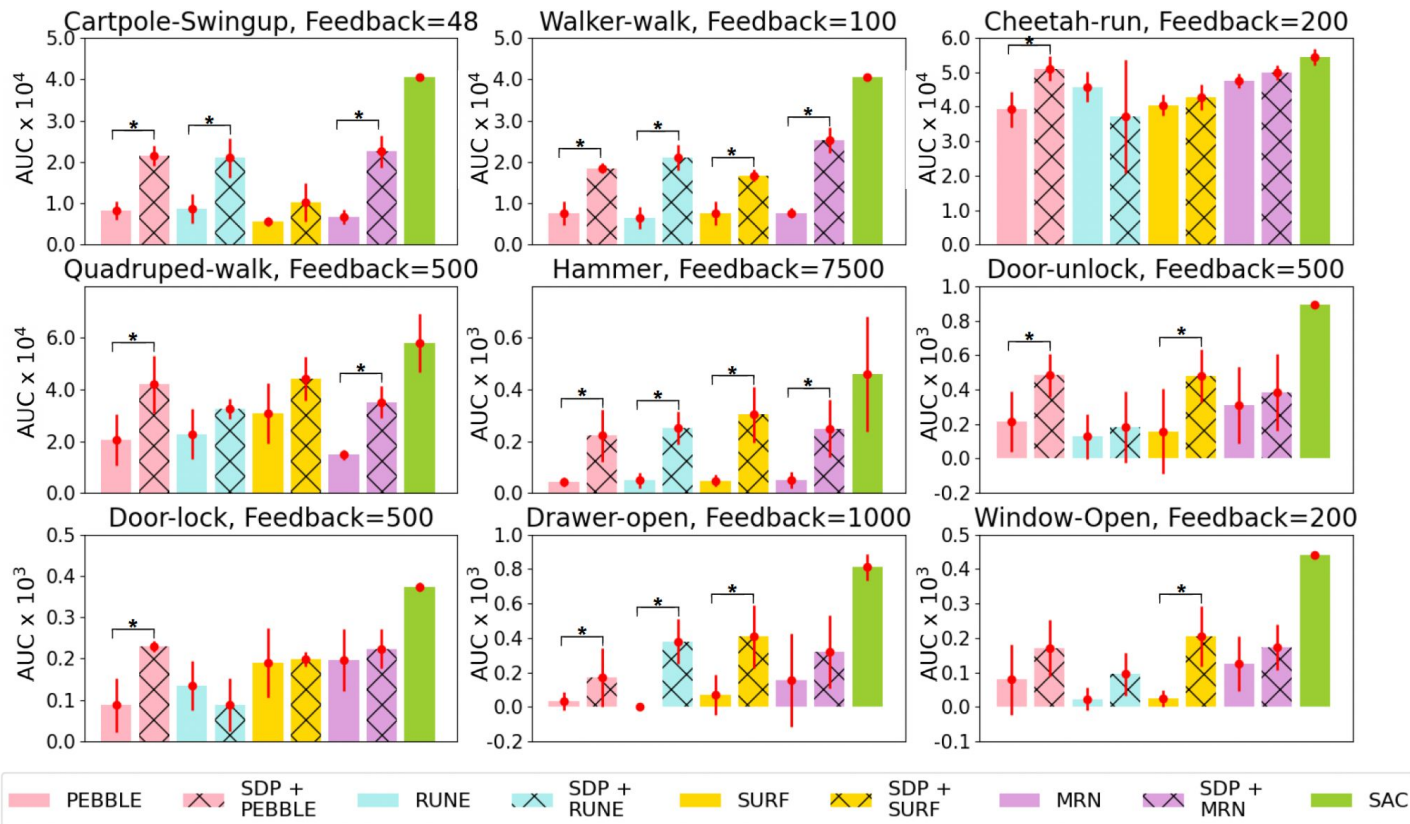


**The reward model learns to associate low-quality transitions with a low reward!**



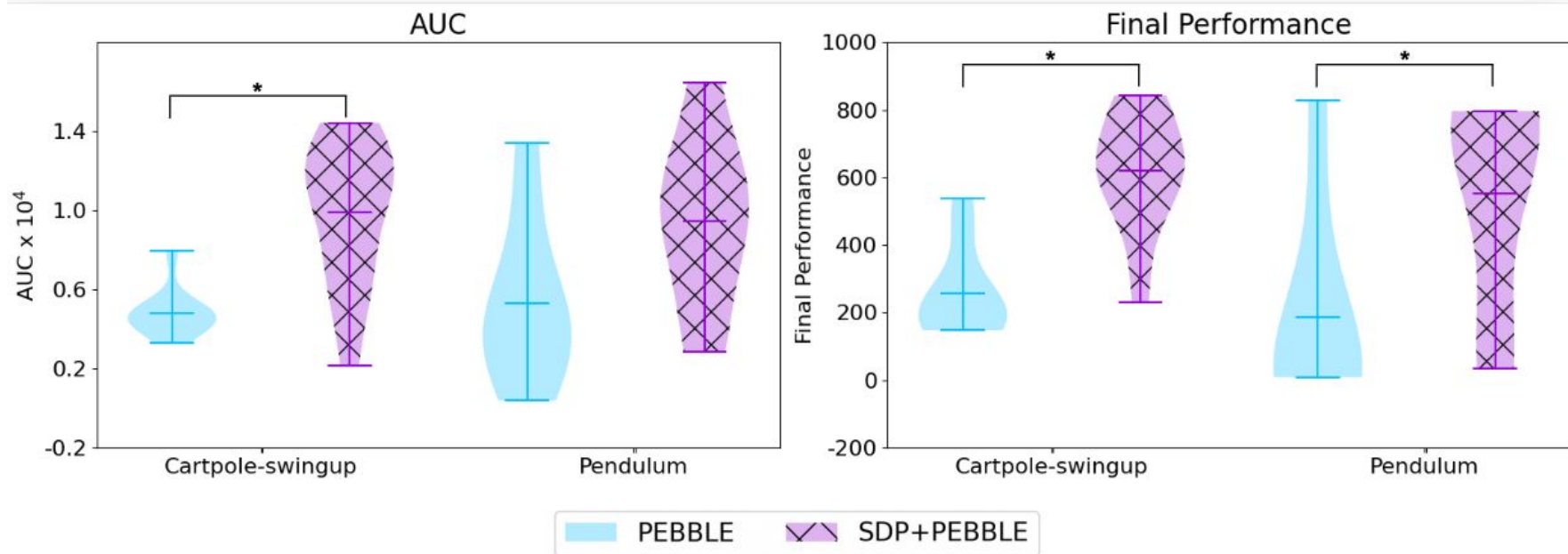
**Human will not need to provide preferences for low-quality behaviors!**

# SDP significantly improved performance in over 63% of experiments



# SDP can work with real humans!

*Ethics-approved user study of 16 participants (CS and non-CS background)*



## Key Takeaways

1. SDP is a simple approach that makes use of sub-optimal data to reduce the amount of human feedback needed in human-in-the-loop RL
2. Effective in simulated robotic environments with both real and simulated humans!