Deep Reinforcement Learning from Human Preferences

AUTHORS

Paul F. Christiano[1], Jan Leike[2], Tom B. Brown[3], Miljan Martic[2], Shane Legg[2], Dario Amodei[1]

AFFILIATIONS

[1] OpenAI,
[2] DeepMind
[3] Google Brain

Group: Matilde Carvalho, Lourenço Carvalho; January 2026

1. *MOTIVATION/PROBLEM*

Recent advances in deep reinforcement learning (RL) rely on well-specified reward functions, yet many real-world tasks involve goals that are complex, subjective or poorly defined. The authors highlight that approximate or poorly designed rewards often lead to unintended behaviour, motivating the need for alternative ways to specify goals that do not depend on explicit reward engineering.

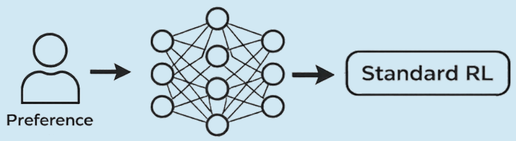
2. *RESEARCH QUESTIONS*

The authors evaluate whether RL from human preferences can achieve performance comparable to standard RL trained with access to the true reward. This leads to two central research questions:

- Can deep reinforcement learning agents learn complex tasks using only human preference feedback, without access to an explicit reward function?
- Can this approach scale to modern deep RL tasks while requiring minimal human supervision?

3. *CORE IDEA*

To address the difficulty of specifying reliable reward functions for complex tasks, the paper treats reward as an unknown signal inferred from human preferences rather than engineered by hand. Instead of a fixed reward r(s,a), the agent receives pairwise judgments over short trajectory segments, learns a neural reward model from these comparisons, and then optimizes this learned signal with standard deep RL.



4. *METHOD OVERVIEW*

The proposed method replaces an explicit reward function with a **learned reward model trained from human preference comparisons**. At each time step, the system maintains two components:

a. a policy that controls the agent's actions, and
b. a reward model that estimates how well the agent's behaviour aligns with human preferences.

1. The agent interacts with the environment and generates trajectories;
2. Humans compare short trajectory segments and indicate which behaviour is preferred;
3. A neural network reward model is learned from preferences:

$$\hat{P}\left[\sigma^1 \succ \sigma^2\right] = \frac{\exp \sum \hat{r}(o_t^1, a_t^1)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}$$

   a. The reward model assigns a score to each trajectory segment;
   b. Scores are converted into preference probabilities;
   c. Model parameters are updated to match human choices;
   d. The reward model is optimized via supervised learning by minimizing cross-entropy loss between predicted and actual human preferences;
4. Standard reinforcement learning optimizes the agent using the learned reward.

5. *KEY TECHNICAL CHOICES*

The following design choices improve stability, sample efficiency, and scalability of the preference-based reward learning approach:

- **Learned reward model**: The reward function is parameterized by a neural network and used as a surrogate reward for RL;
- **Ensemble of reward models**: Multiple reward predictors are trained and averaged to improve stability and estimate uncertainty;
- **Active query selection**: Human queries are selected based on disagreement across the reward model ensemble, prioritizing informative comparisons;
- **Online reward learning**: The reward model is continuously updated during training to adapt to the agent's evolving behavior.

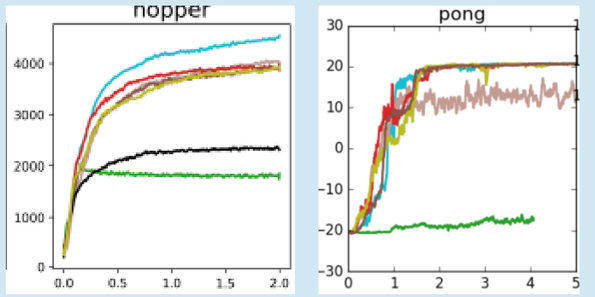
6. *EXPERIMENTS*

The experiments validate our method on both **MuJoCo** control tasks and **Atari** games, using human preferences to train the reward model that guides learning. In MuJoCo we use **TRPO**, while in Atari we use **A2C**, both optimized with the learned reward instead of the environment's reward signal.

The A2C agent interacts with the Atari environment to generate gameplay trajectories, from which short clips are sampled and compared by human raters to express their preferences. A neural reward model is trained on these preference comparisons

7. *KEY RESULTS*

- **TRPO** on learned MuJoCo rewards reaches near–true–reward performance with only hundreds of human preference queries, demonstrating high sample efficiency.
- **A2C** on Atari achieves meaningful performance from learned rewards, though less consistently than MuJoCo, revealing sensitivity to reward model quality



8. *CONCLUSION*

- Human preference comparisons can effectively replace explicit reward functions, enabling deep RL in tasks with complex or subjective objectives;
- The approach reaches performance comparable to standard RL while requiring orders of magnitude less human feedback, enabling practical scalability;
- This framework opens the door to applying deep RL to real-world problems where goals are easier to judge qualitatively than to specify mathematically.

Technology
Arts Sciences
TH Köln

SOURCES: [Christiano et al., 2017]. Deep Reinforcement Learning from Human Preferences. 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA. https://th-koeln.sciebo.de/apps/onlyoffice/s/BpQaHRWrR7J7PfG?fileId=1059350931