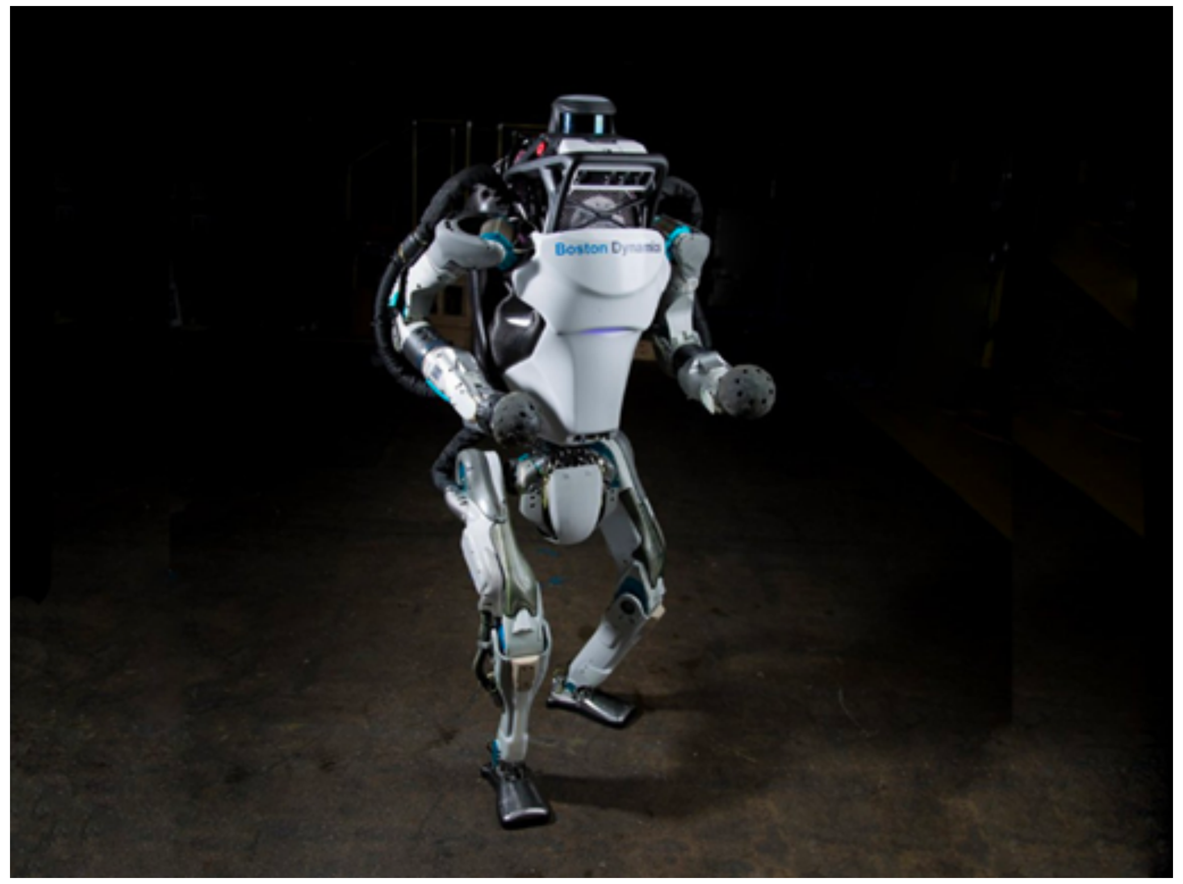
Human-in-the-loop: Provably Efficient Preference-based Reinforcement Learning with General Function Approximation

Xiaoyu Chen[1], Han Zhong[1], Zhuoran Yang[2], Zhaoran Wang[3], Liwei Wang[1]

1. Peking University, 2. Yale University, 3. Northwestern University

Reward-shaping Challenges

- Standard RL: The agent interacts with the unknown environment aiming to maximize cumulative rewards
- However, in many tasks, reward functions might not be readily available or difficult to design



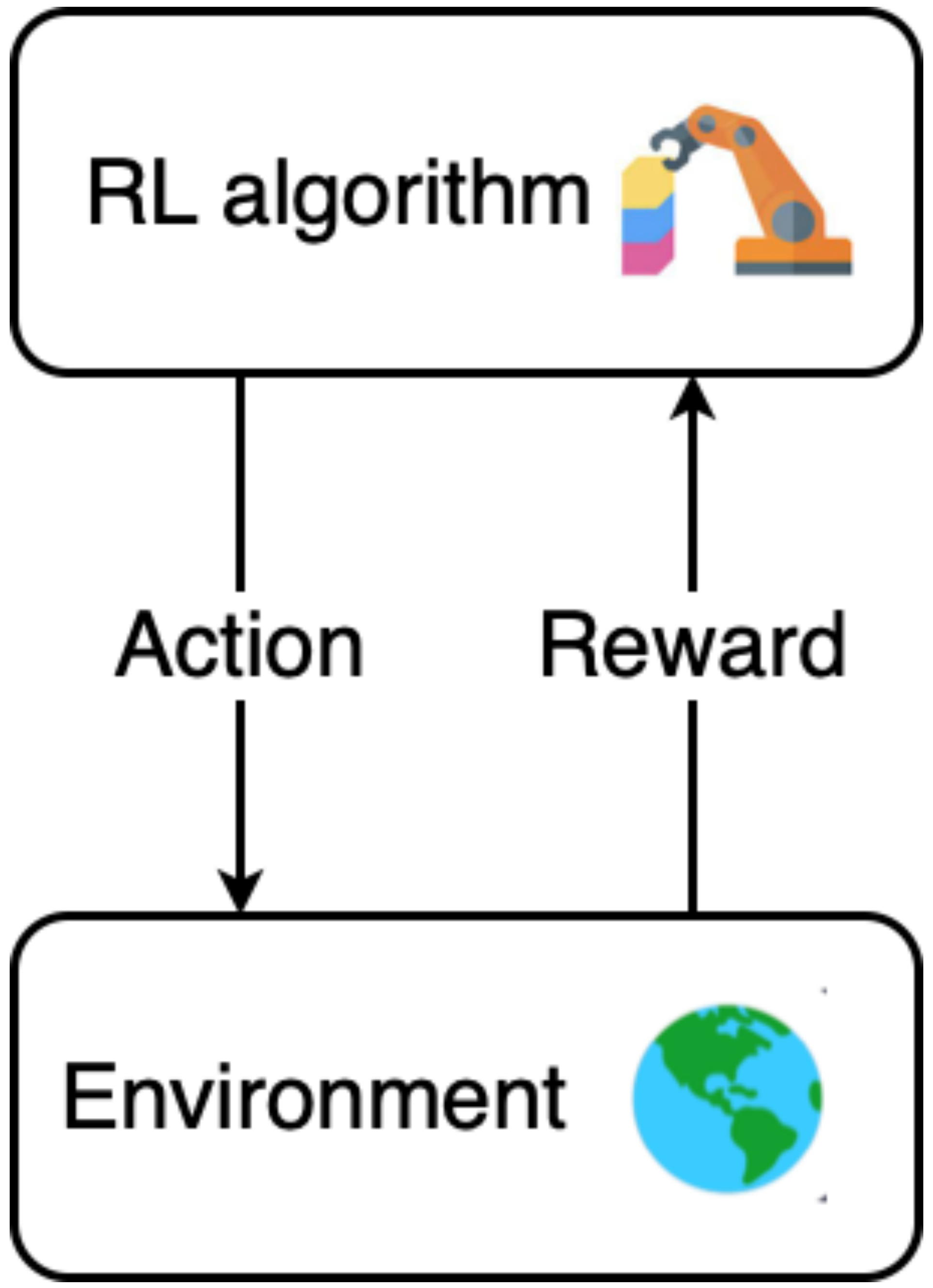
Robotics

Autonomous Driving

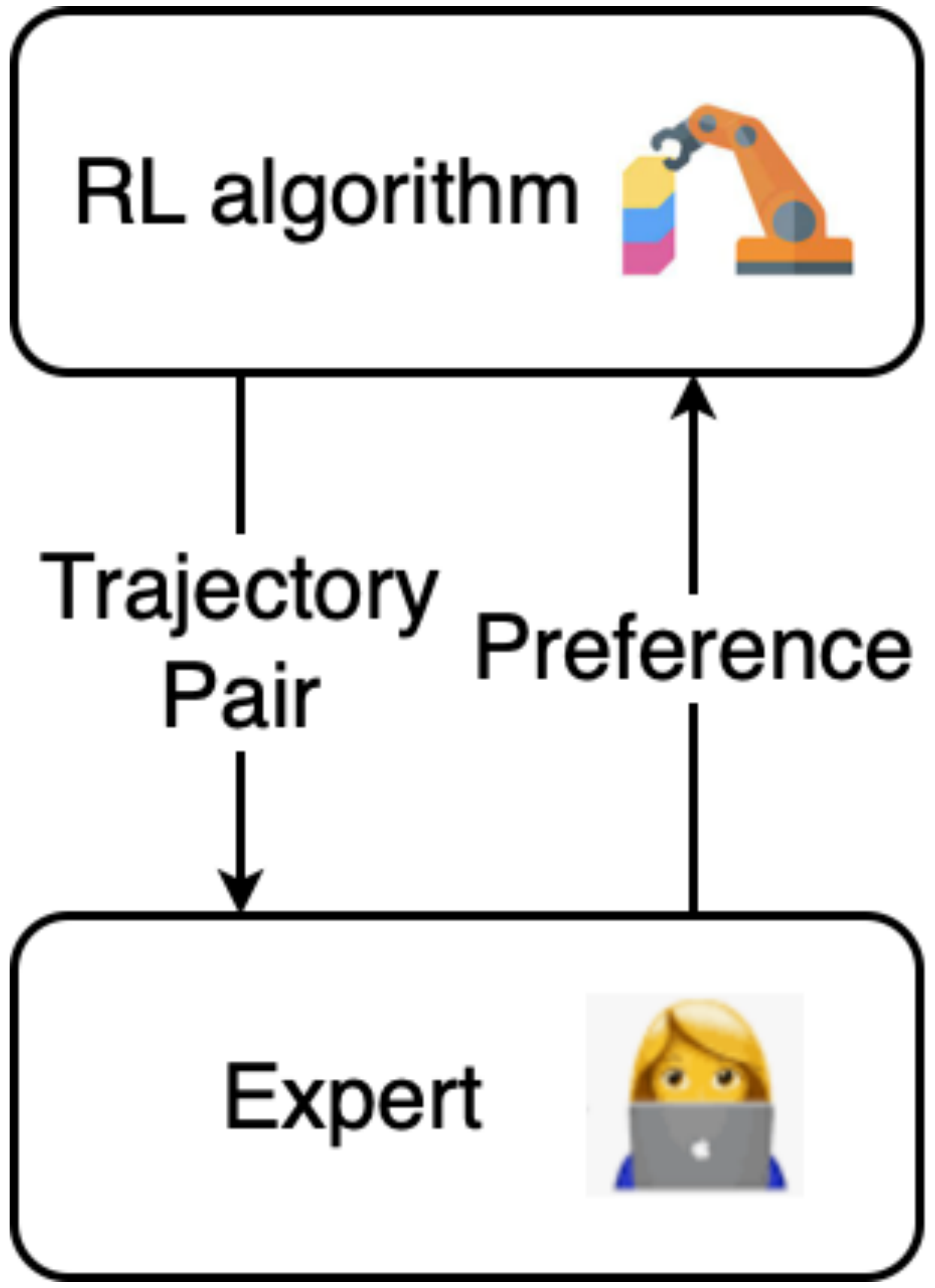
Healthcare

Preference-based RL (PbRL)

- Basic Idea: The agent only receives preference feedback over trajectory pairs instead of reward signals.



Standard RL



PbRL

- Previous results mainly study efficient exploration in tabular RL setting [1,2].

[1] Xu et al. Preference-based reinforcement learning with finite-time guarantees. NeurIPS 2022
[2] Pacchiano et al. Dueling rl: Reinforcement learning with trajectory preferences. arXiv pre-print

Our Formulation

- Overview: Efficient exploration for RbRL with general function approximation.

- Preference-based RL formulation (Episodic setting with K episodes):
  - In episode k, the agent executes two policies $\pi_{k,1}, \pi_{k,2}$, obtains the trajectories $\tau_{k,1}, \tau_{k,2}$, and asks for their preferences between $\tau_{k,1}, \tau_{k,2}$
  - $\mathbb{T}\left(\tau_1, \tau_2\right) = \Pr\left(\tau_1 > \tau_2\right)$: the probability that τ_1 is preferred compared with τ_2
  - $\mathbb{T}\left(\pi_1, \pi_2\right)$: The expected preference over two policies π_1, π_2
  - We aim to minimize the regret compared with the optimal policy π^*:

$$\mathrm{Reg}(K) = \sum_{k=1}^{K} \sum_{i=1}^{2} \left(\mathbb{T}\left(\pi^*, \pi_{k,i}\right) - \frac{1}{2} \right)$$

Our Formulation

- Overview: Efficient exploration for RbRL with general function approximation.

- Function Approximation formulation:
  - We assume the transition and preference function space satisfies bounded Eluder dimension and log-covering number.
  - Covers linear and generalized linear preference function space.
  - Covers linear mixture MDPs and tabular MDPs.

Our Results

- We propose an efficient learning algorithm with $\tilde{O}(\mathrm{poly}(dH)\sqrt{K})$ regret
  - d: Eluder dimension or log-covering number of the transition and preference space
  - H: horizon in episodic setting
  - K: Total number of episodes

- Main techniques:
  - Construct a near-optimal policy set and execute the most exploratory policy [2]
  - A refined confidence bonus inspired from the reward-free setting [3]

- Our lower bound indicates that the upper bound is near-optimal when specialized to linear setting

[2] Pacchiano et al. Dueling rl: Reinforcement learning with trajectory preferences. arXiv pre-print
[3] Chen et al. Near-optimal reward-free exploration for linear mixture mdps with plug-in solver. ICLR 2022

Other results

- Connection with the setting of *RL with Once-per-episode feedback* [4]
  - Upper bound: our Algorithm for PbRL can be almost directly applied to this setting with near-optimal regret.
  - Lower bound: The lower bound for PbRL is derived by reduction from the problem of RL with once-per-episode feedback.

- New setting: RL with n-wise Comparisons
  - Basic idea: multiple trajectories are sampled and compared with each other
  - We propose efficient algorithm with near-optimal regret guarantee

[4] Chatterji et al. On the theory of reinforcement learning with once- per-episode feedback. NeurIPS 2021