Enhancing Human Experience in Human-Agent Collaboration: A Human-Centered Modeling Approach Based on Positive Human Gain

Tencent AI Lab & Tencent TiMi L1 Studio

Yiming Gao*, Feiyu Liu*, Liang Wang, Dehua Zheng, Zhenjie Lian, Weixuan Wang, Wenjin Yang, Siqin Li, Xianliang Wang, Wenhui Chen, Jing Dai, Qiang Fu, Wei Yang, Lanxiao Huang, Wei Liu

yatminggao@tencent.com

ICLR

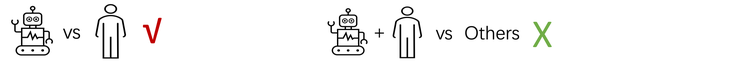
Background

- Recently, a variety of AI systems for MOBA game-playing have been developed.
  - Focus on enhancing agents' abilities to win the games.
  - Achieve or even exceed human-level performance in Human-Agent Competition.

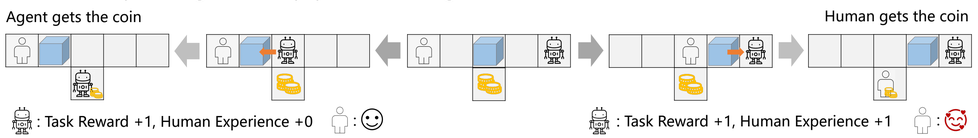
OpenAI Five (OpenAI et al. [2019])

Wukong (Ye et al. [2020a])

- However, they do not essentially care about the human experience in Human-Agent Collaboration.



- Humans reported greater enjoyment of the game when the AI assisted them more like a sidekick.



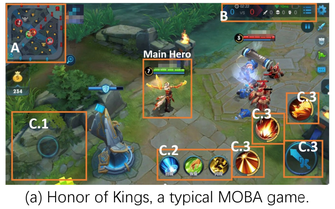
Contribution

- For guiding human-level agents to enhance the human experience in Human-Agent Collaboration,
  - we propose a human-centered agent modeling scheme.
- We gain insights into the challenge of human-agent credit assignment,
  - we address this challenge by presenting the Reinforcement Learning from Human Gain algorithm.
- We conducted human-agent tests in *Honor of Kings*, and both objective performance and subjective preference results show that the RLHG agent provides participants better gaming experience.

Human-centered Agent Modeling

- Our Key Idea: Include enhancing human experience into the optimization objectives of the agent.

1. We conceptualize the human experience as the human goals they expect to achieve during the task.



(a) Honor of Kings, a typical MOBA game.

Human experience	Task Goal: Game Victory	Human Goals: MVP Score	In-game Resource	Highlight Moment	…
Bittersweet	Win	Low	Few	Few	
Enjoyable	Win	**High**	**More**	**More**	

(b) An example of human experience in MOBA games.

2. We quantify human goals $\mathcal{G}^H = \{g_i\}_{i=1,\dots,M}$ as human rewards $R^H : \mathbf{S} \times \mathbf{A} \times \mathcal{G}^H \to \mathbb{R}$.
   - Human rewards measure the extent to which humans achieve these goals.
3. We combine agents' original (task-related) rewards with human rewards.
   - **Self-centered objective:** Optimize the agents' abilities to complete the task.

$$J(\theta) = V^{\pi_\theta,\pi_H}(s) = \mathbb{E}_{\pi_\theta,\pi_H}\left[G_t | s_t = s\right], \text{ where } G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

   - **Human-centered objective:** Optimize the agents' abilities to enhance human performance in achieving human goals.

$$J(\theta) = V^{\pi_\theta,\pi_H}(s) + \alpha \cdot V_H^{\pi_\theta,\pi_H}(s) = \mathbb{E}_{\pi_\theta,\pi_H}\left[G_t + \alpha \cdot G_t^H | s_t = s\right], \text{ where } G_t^H = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}^H$$

Human-Agent Credit Assignment

- Case: The human-centered agent follows the human throughout the entire game.

The red box contains human players.

- Insight: Human rewards are assigned to agents without any assistance behavior.
  - Humans possess the primitive abilities to achieve certain goals independently.
  - The agent can easily explore human goals that can be readily achieved by humans.

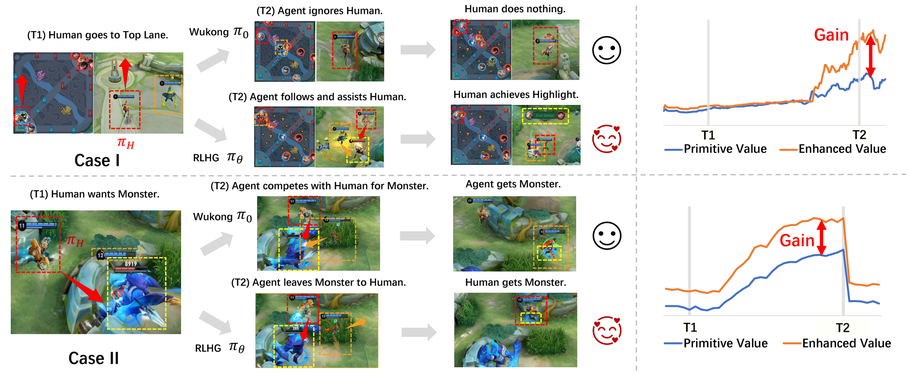
$$\nabla J(\theta) = \mathbb{E}_{\pi_\theta,\pi_H}\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|o_t)\left(A(s_t,a_t) + \alpha \cdot \boxed{A_H(s_t,a_t)}\right)\right]$$

  - Losing autonomy
  - Invalid enhancement behaviors

Effective Human Enhancement

- **Human Return** G_t^H: The discounted cumulative human rewards.
- **Human Primitive Value** $V_H^{\pi_0,\pi_H}(s)$: The human's own contribution to human return.
- **Human Gains**: The agent's actual contribution in enhancing human to achieve goals.

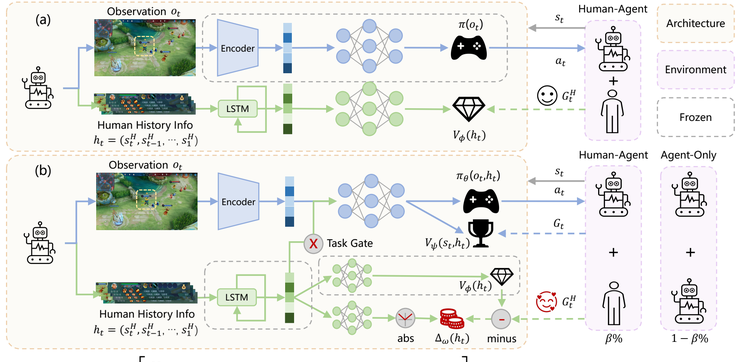
$$\Delta(s,a) = \mathbb{E}_{\pi_\theta,\pi_H}[G_t^H | s_t = s, a_t = a] - V_H^{\pi_0,\pi_H}(s)$$



- **Effective Enhancement Behaviors**: $A^+(s) = \{a | a \sim \pi_\theta, \Delta(s,a) > 0\}$
  - Actions that can help the human achieve human goals better than the primitive actions.
- **Invalid Enhancement Behaviors**: $A^-(s) = A(s) \setminus A^+(s) = \{a | a \sim \pi_\theta, \Delta(s,a) \leq 0\}$
  - Actions that provide no benefits or even hinders the human from achieving human goals.

Reinforcement Learning from Human Gain (RLHG)

- To ensure that the agent only learns effective enhancement behaviors, we divide the training process into two stages.
- Stage I: **Human Primitive Value Estimation**
  - We freeze the pre-trained agent policy and collect human-agent team trajectory samples to compute the human return.
  - The human primitive value network is updated by minimizing the TD errors.



$$\nabla J(\theta) = \mathbb{E}_{\pi_\theta,\pi_H}\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|o_t)\left(A(s_t,a_t) + \alpha \cdot \widehat{A}_H(s_t,a_t)\right)\right], \quad \widehat{A}_H(s,a) = \Delta(s,a) - \widehat{\Delta}(s), \tag{4}$$

Algorithm 1 Reinforcement Learning from Human Gain (RLHG)

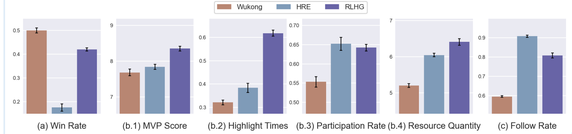
while not converged **do**
  Freeze π_0 and collect human-agent team $< \pi_0, \pi_H >$ trajectories;
  Compute human return G^H, and update human primitive value $V_\phi \leftarrow G^H$;
end while // Stage I: Human Primitive Value Estimation
while not converged **do**
  Freeze V_ϕ and collect human-agent team $< \pi_\theta, \pi_H >$ trajectories;
  Compute original return G, self-centered advantage A;
  Compute human return G^H, human gain Δ, and human-centered advantage $\widehat{A}_H$;
  Update agent policy network π_θ using Eq. 4, and value network $V_\psi \leftarrow G$;
  Update human gain network Δ_ω using Eq. 5;
end while // Stage II: Human Enhancement Training

- Stage II: **Human Enhancement Training**
  - We freeze the human primitive value network.
  - We collect human-agent team trajectory samples to compute the self-centered advantage and human return.
  - We estimate the expected positive human gain from effective enhancement behaviors. $\widehat{\Delta}(s) = \mathbb{E}_{a \sim A^+(s)}[\Delta(s,a)]$

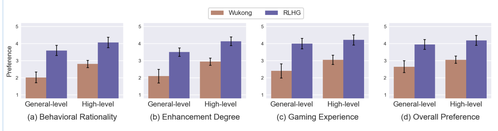
$$L(\omega) = \mathbb{E}_{s \in S, a \in A}\left[\mathbb{I}(\Delta(s,a)) \cdot \|\Delta(s,a) - \Delta_\omega(s)\|_2\right], \quad \mathbb{I}(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{5}$$

  - We replace the original human-centered advantage with the advantage of human gain over $\widehat{\Delta}(s)$.
  - The agent's policy and value network are fine-tuned using PPO.

Human Model-Agent Test



Human-Agent Test



High-level Participants

Agent \ Goals	Task Goal: Win Rate	Top 4 human Goals: MVP Score	Highlight Times	Participation Rate	Resource Quantity
Wukong	52%	8.86 (0.79)	0.53 (0.21)	0.46 (0.11)	5.3 (2.87)
RLHG	46.7%	**10.28** (0.75)	**0.87** (0.29)	**0.58** (0.09)	**6.28** (2.71)

General-level Participants

Agent \ Goals	Task Goal: Win Rate	Top 4 Human Goals: MVP Score	Highlight Times	Participation Rate	Resource Quantity
Wukong	34%	7.44 (0.71)	0.37 (0.349)	0.41 (0.11)	4.98 (2.73)
RLHG	30%	**9.1** (0.61)	**0.75** (0.253)	**0.59** (0.05)	**5.8** (2.78)