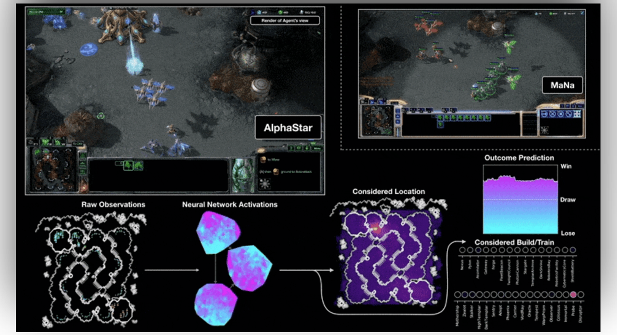
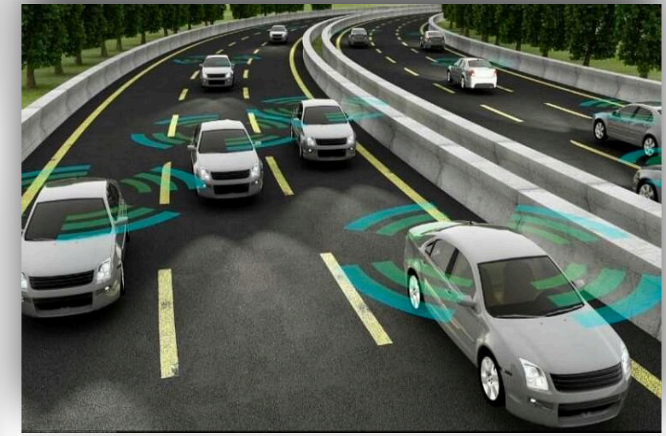
Reinforcement Learning with Human Values

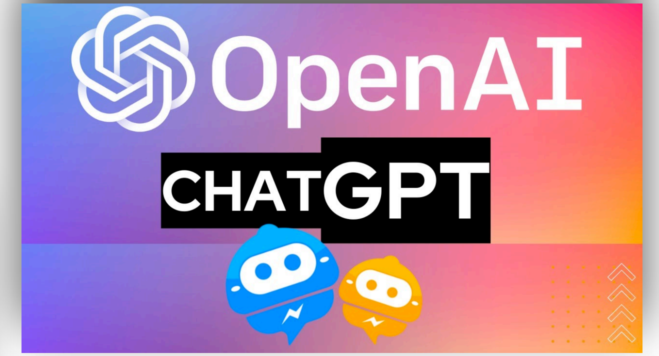
Yali Du
King's College London

1 Jun 2023

Interactive decision making



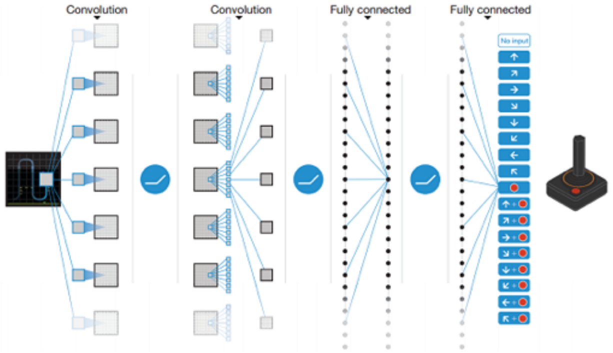




Decision Making

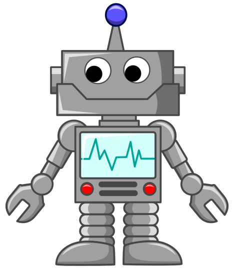
Deep Reinforcement Learning = DL + RL

DL



θ

Agent



$a \sim \pi_\theta(\cdot \mid s)$

Action a

State s,

Reward r

Environment

$p(s_{t+1} \mid s_t, a_t)$

RL

The RL objective: $\max_{\pi} \; E_{s_t,\; a_t,\; \ldots}[\sum_{t=0}^{\infty} r(s_t, a_t)]$

Challenges

- Existing success often comes with well-specified reward function
  - Go, Chess, StarCraft II, …
- However,
  - The quality of the designed reward function largely depends on the designer,
  - The agent may hack the reward function.
- Can we train reinforcement learning agent without well-specified reward function?

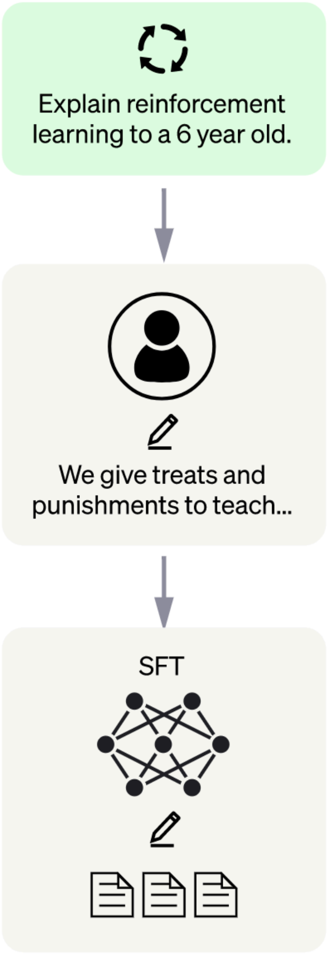
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



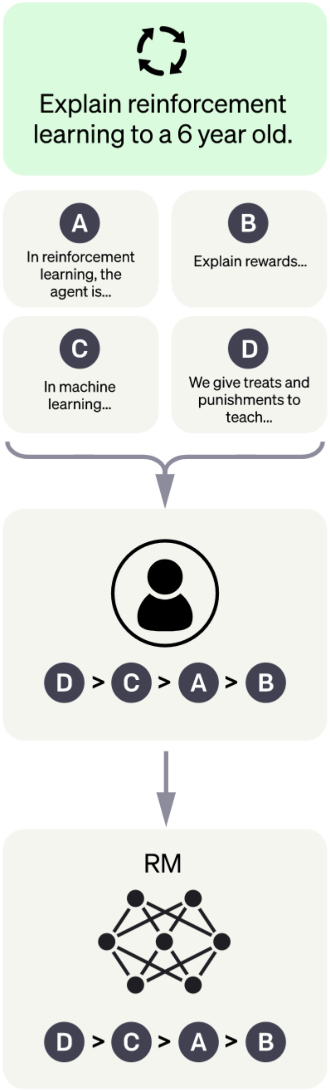
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

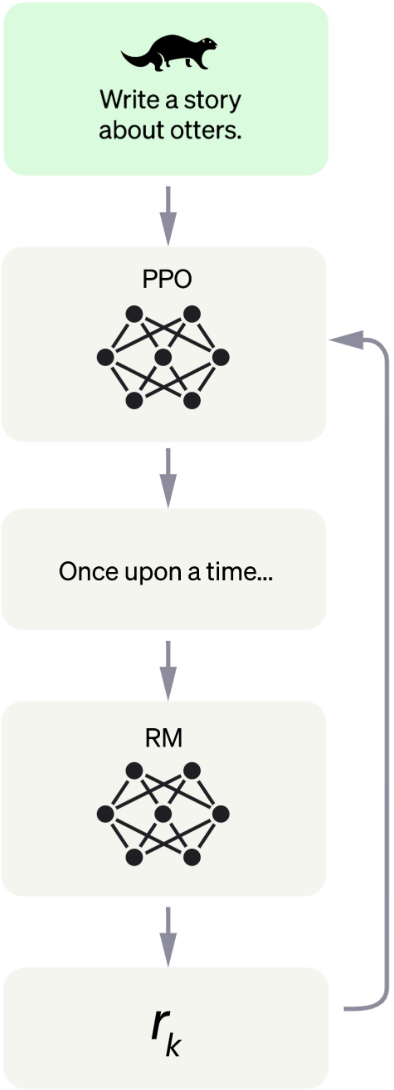
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

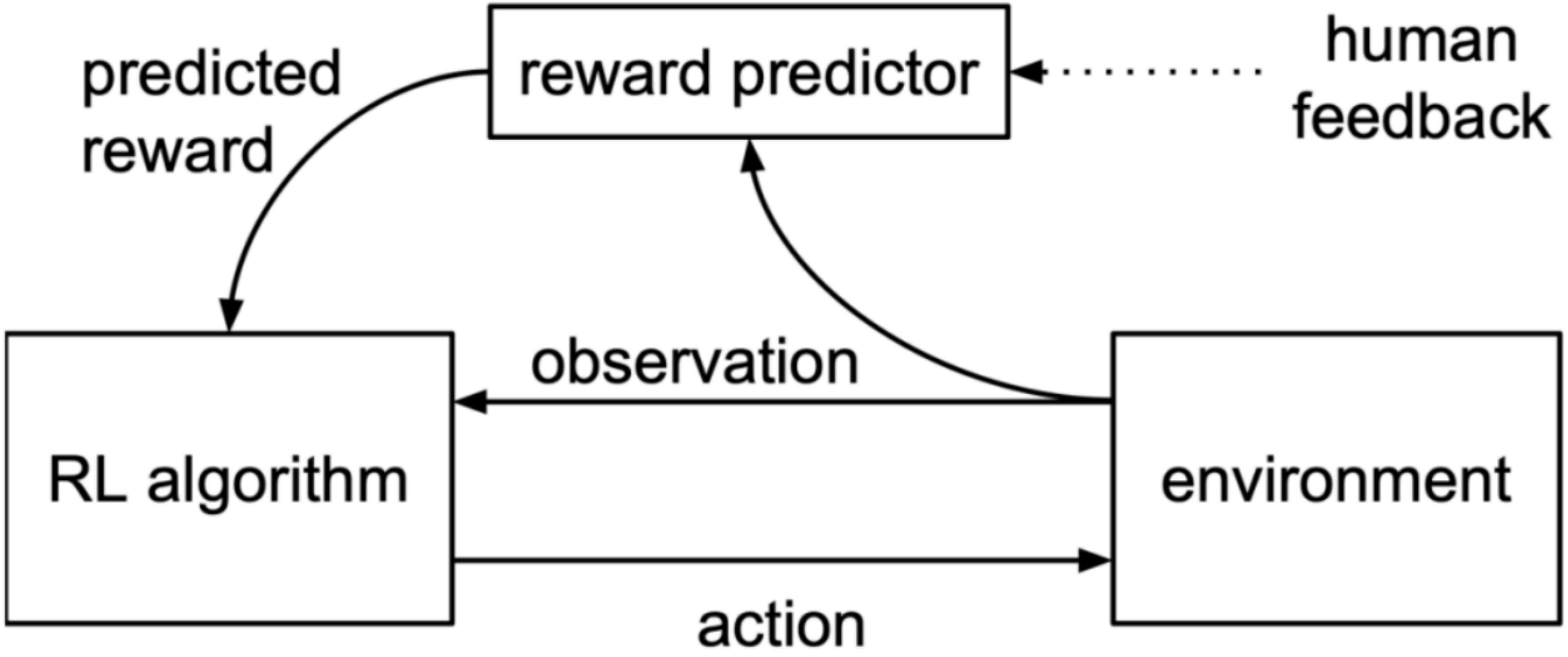
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



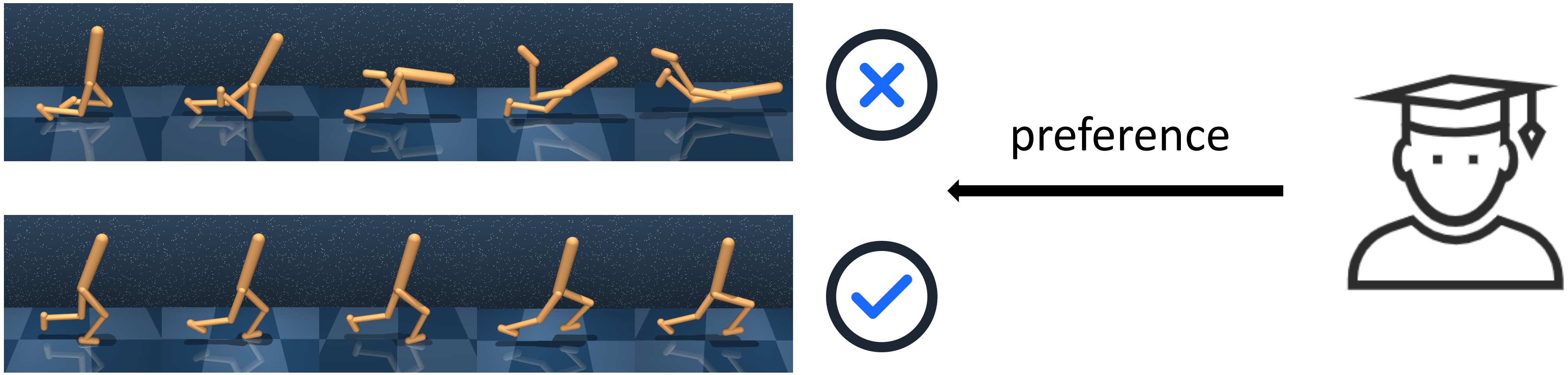
https://openai.com/blog/chatgpt

RLHF framework: Reinforcement learning from human feedback



-

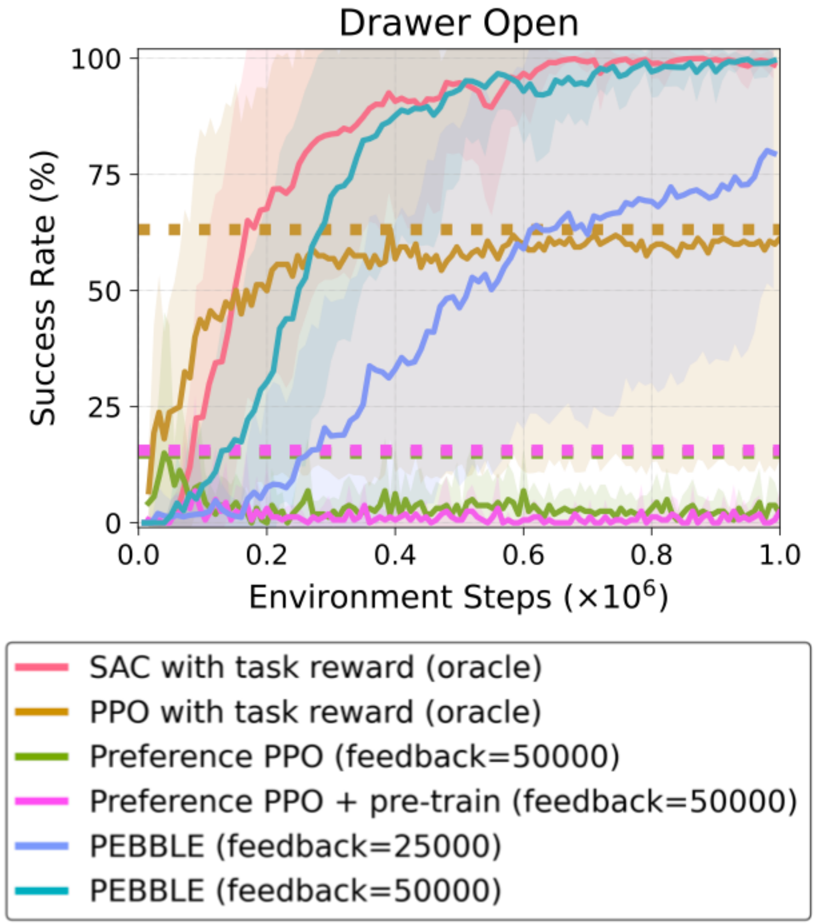
Preference-based RL



- Main reference:
  Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning. Runze Liu, Fengshuo Bai, Yali Du, Yaodong Yang. NeurIPS 2022

Preference-based RL

- Key challenge: feedback efficiency
  - Preference data is expensive.
  - Previous methods work badly given little feedback.
  - Confirmation bias, Q-function may overfit to the inaccurate outputs of the reward function.



Preference-based RL

- Construct a preference predictor by Bradley-Terry model:

$$P_\psi[\sigma^0 \succ \sigma^1] = \frac{\exp \sum_t \widehat{r}_\psi(s_t^0, a_t^0)}{\exp \sum_t \widehat{r}_\psi(s_t^0, a_t^0) + \exp \sum_t \widehat{r}_\psi(s_t^1, a_t^1)}$$

- Optimize the reward function through a classification task:

$$\mathcal{L}_{\text{supervised}}(\psi) = - \mathop{\mathbb{E}}_{(\sigma^0, \sigma^1, y) \sim \mathcal{D}} \Big[y(0) \log P_\psi[\sigma^0 \succ \sigma^1] + y(1) \log P_\psi[\sigma^1 \succ \sigma^0] \Big]$$

- Perform RL algorithms to learn a well-behaved policy.

[1] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952.
[2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In NeurIPS 2017.

Meta-Reward-Net

- Construct a preference predictor using the Q-function:

$$P_\theta[\sigma^0 \succ \sigma^1] = \frac{\exp Q_\theta(s_0^0, a_0^0)}{\exp Q_\theta(s_0^0, a_0^0) + \exp Q_\theta(s_0^1, a_0^1)}.$$

- Evaluate the Q-function on the preference data:

$$\mathscr{L}_{\text{meta}}(\theta(\psi)) = -\mathop{\mathbb{E}}_{(\sigma^0,\sigma^1,y)\sim\mathscr{D}}\Big[y(0)\log P_{\theta(\psi)}[\sigma^0 \succ \sigma^1] + y(1)\log P_{\theta(\psi)}[\sigma^1 \succ \sigma^0]\Big],$$

- Define the $Q-$loss: $J_Q(\theta) = \mathbb{E}_{\tau_t\sim\mathscr{B}}\Big[\Big(Q_\theta(s_t, a_t) - \hat{r}_\psi(s_t, a_t) - \gamma\bar{V}(s_{t+1})\Big)^2\Big].$

- The objective

$$\min_{\psi,\theta} \quad \mathscr{L}_{\text{meta}}(\theta(\psi)),$$

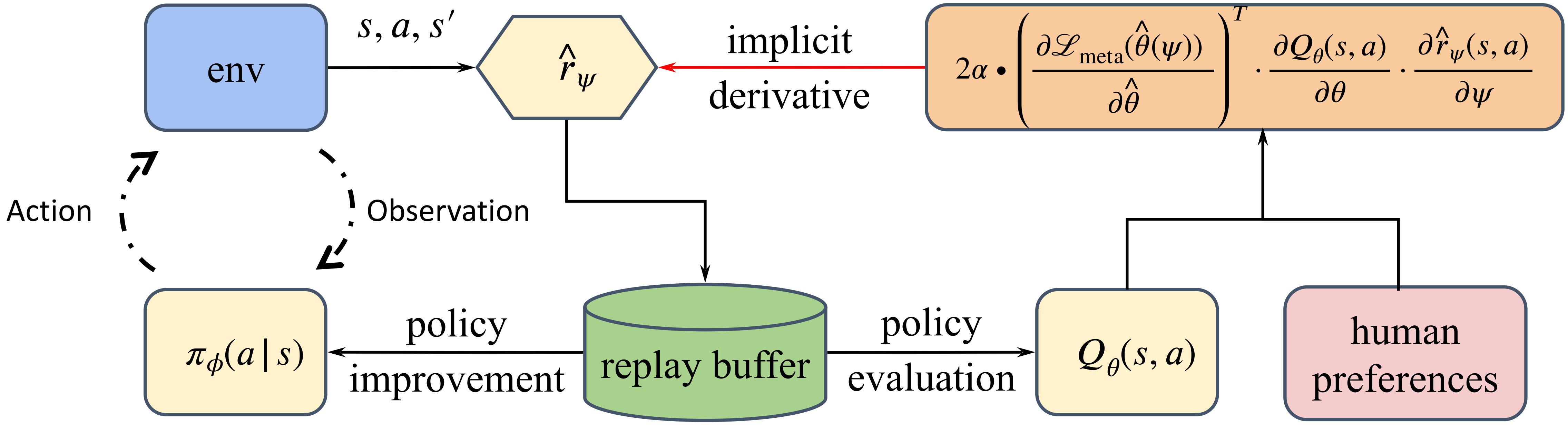
$$\text{s.t.} \quad \theta(\psi) = \arg\min_\theta J_Q(\theta, \psi).$$

Meta-Reward-Net

- Inner-level updating:
  - $J_Q(\theta) = \mathbb{E}_{\tau_t \sim \mathscr{B}}\left[\left(Q_\theta(s_t, a_t) - \widehat{r}_\psi(s_t, a_t) - \gamma \bar{V}(s_{t+1})\right)^2\right].$
  - $\theta^{(k+1)} = \theta^{(k)} - \alpha \nabla_\theta J_Q(\theta)\Big|_{\theta^{(k)}},$
  - Update policy π based on critic $Q(s, a)$.
- Outer-level updating:
  - $g^{(k)}_{\text{meta}} = \nabla_{\hat{\theta}} \mathscr{L}_{\text{meta}}(\hat{\theta}(\psi))\Big|_{\hat{\theta}^{(k)}} \nabla_\psi \hat{\theta}^{(k)}(\psi)\Big|_{\psi^{(k)}} = h \cdot \nabla_\psi \widehat{r}(s_t, a_t; \psi)\Big|_{\psi^{(k)}},$
  - $\psi^{(k+1)} = \psi^{(k)} - \beta\, g^{(k)}_{\text{meta}}\Big|_{\psi^{(k)}},$

Our work: Meta-Reward-Net

- Main idea: consider the performance of the Q-function in reward learning



Theoretical Results

Theorem 1. *Assume the outer loss $\mathcal{L}_{\text{meta}}$ is Lipschitz smooth with constant L, and the gradient of $\mathcal{L}_{\text{meta}}$ and J_Q is bounded by ρ. Let $\widehat{r}_\psi$ be twice differential, with its gradient and Hessian respectively bounded by δ and $\mathcal{B}$. For some $c_1 > 0$, suppose the learning rate of the inner updating $\alpha_k = \min\{1, \frac{c_1}{T}\}$, where $c_1 < T$. For some $c_2 > 0$, suppose the learning rate of the outer updating $\beta_k = \min\{\frac{1}{L}, \frac{c_2}{\sqrt{T}}\}$, where $\frac{\sqrt{T}}{c_2} \geq L$, $\sum_{k=1}^{\infty} \beta_k \leq \infty$ and $\sum_{k=1}^{\infty} \beta_k^2 \leq \infty$. Meta-Reward-Net can achieve:*

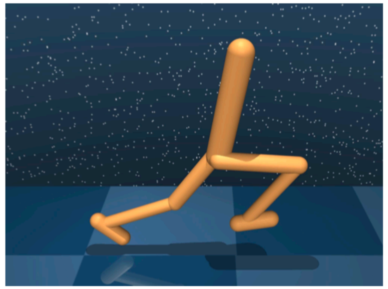
$$\min_{1 \leq k \leq T} \mathbb{E}\left[\left\|\nabla_\psi \mathcal{L}_{\text{meta}}(\hat{\theta}^{(k)}(\psi^{(k)}))\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Theorem 2. *Assume the outer loss $\mathcal{L}_{\text{meta}}$ is Lipschitz smooth with constant L, and the gradient of $\mathcal{L}_{\text{meta}}$ and J_Q is bounded by ρ. Let $\widehat{r}_\psi$ be twice differential, with its gradient and Hessian respectively bounded by δ and $\mathcal{B}$. For some $c_1 > 0$, suppose the learning rate of the inner updating $\alpha_k = \min\{1, \frac{c_1}{T}\}$, where $c_1 < T$. For some $c_2 > 0$, suppose the learning rate of the outer updating $\beta_k = \min\{\frac{1}{L}, \frac{c_2}{\sqrt{T}}\}$, where $\frac{\sqrt{T}}{c_2} \geq L$, $\sum_{k=1}^{\infty} \beta_k \leq \infty$ and $\sum_{k=1}^{\infty} \beta_k^2 \leq \infty$. Meta-Reward-Net can achieve:*

$$\lim_{k \to \infty} \mathbb{E}\left[\left\|\nabla_\theta J_Q(\theta^{(k)}; \psi^{(k+1)})\right\|^2\right] = 0. \tag{39}$$

Theoretically, the algorithms converge to local optimum, check more results in paper.

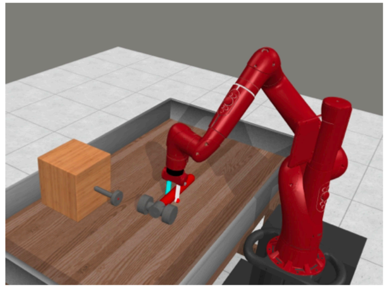
Experiments



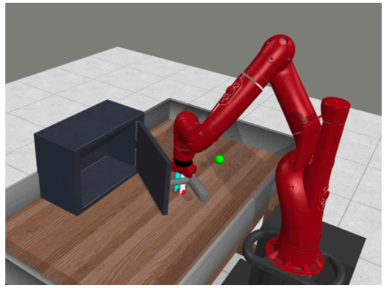
(a) Walker

(b) Cheetah

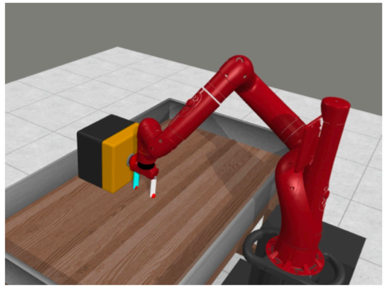
(c) Quadruped



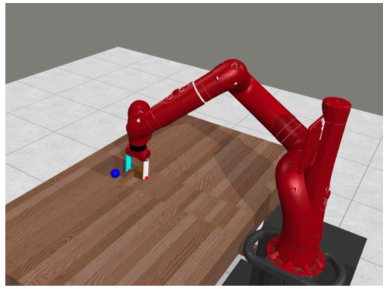
(d) Hammer



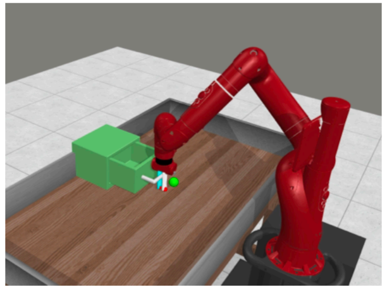
(e) Door Open



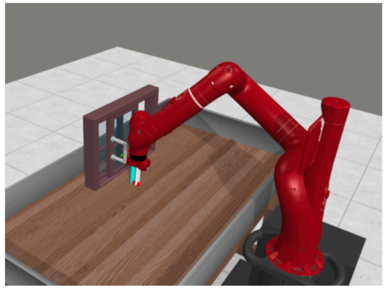
(f) Button Press



(g) Sweep Into



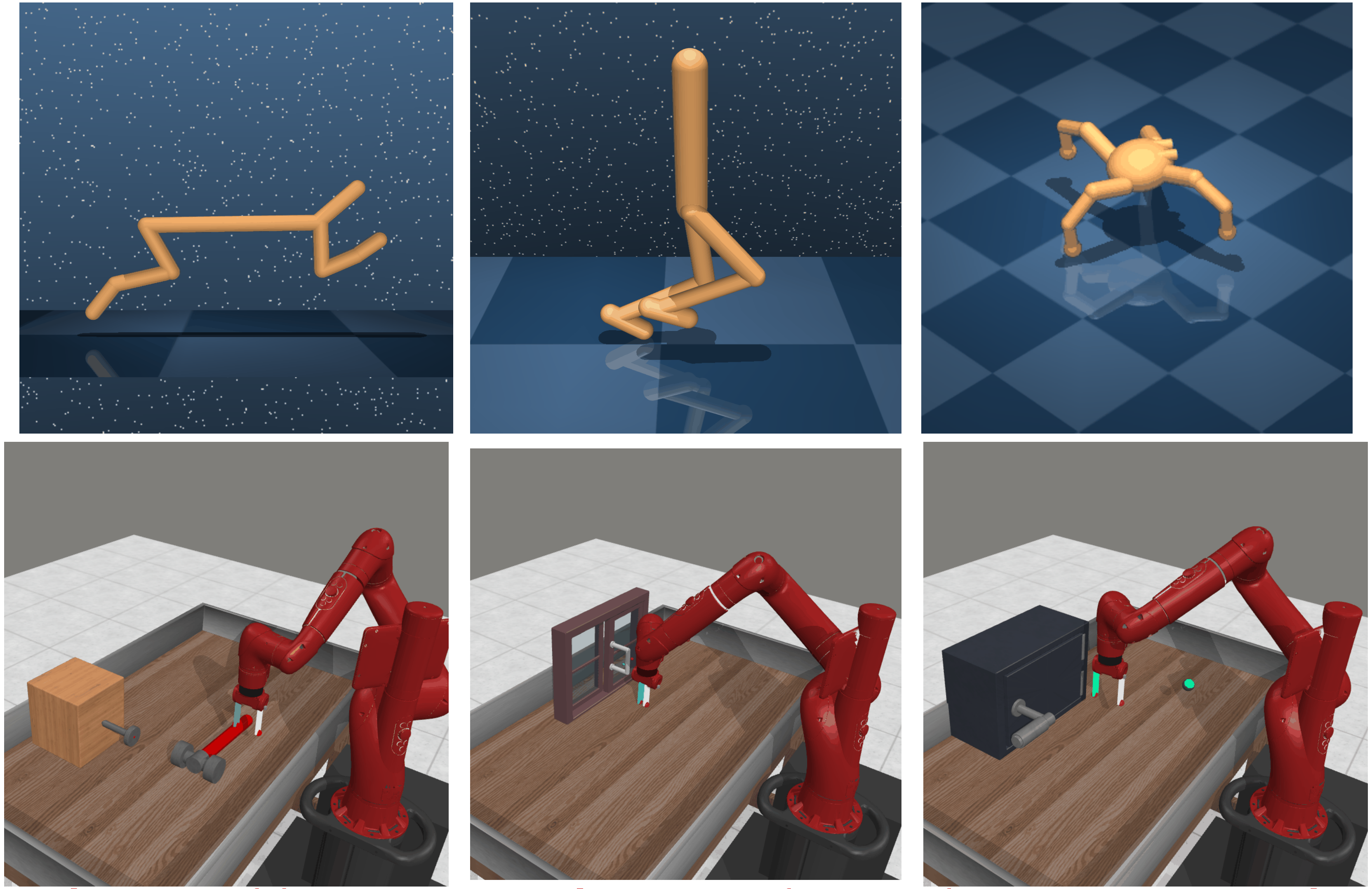
(h) Drawer Open



(i) Window Open

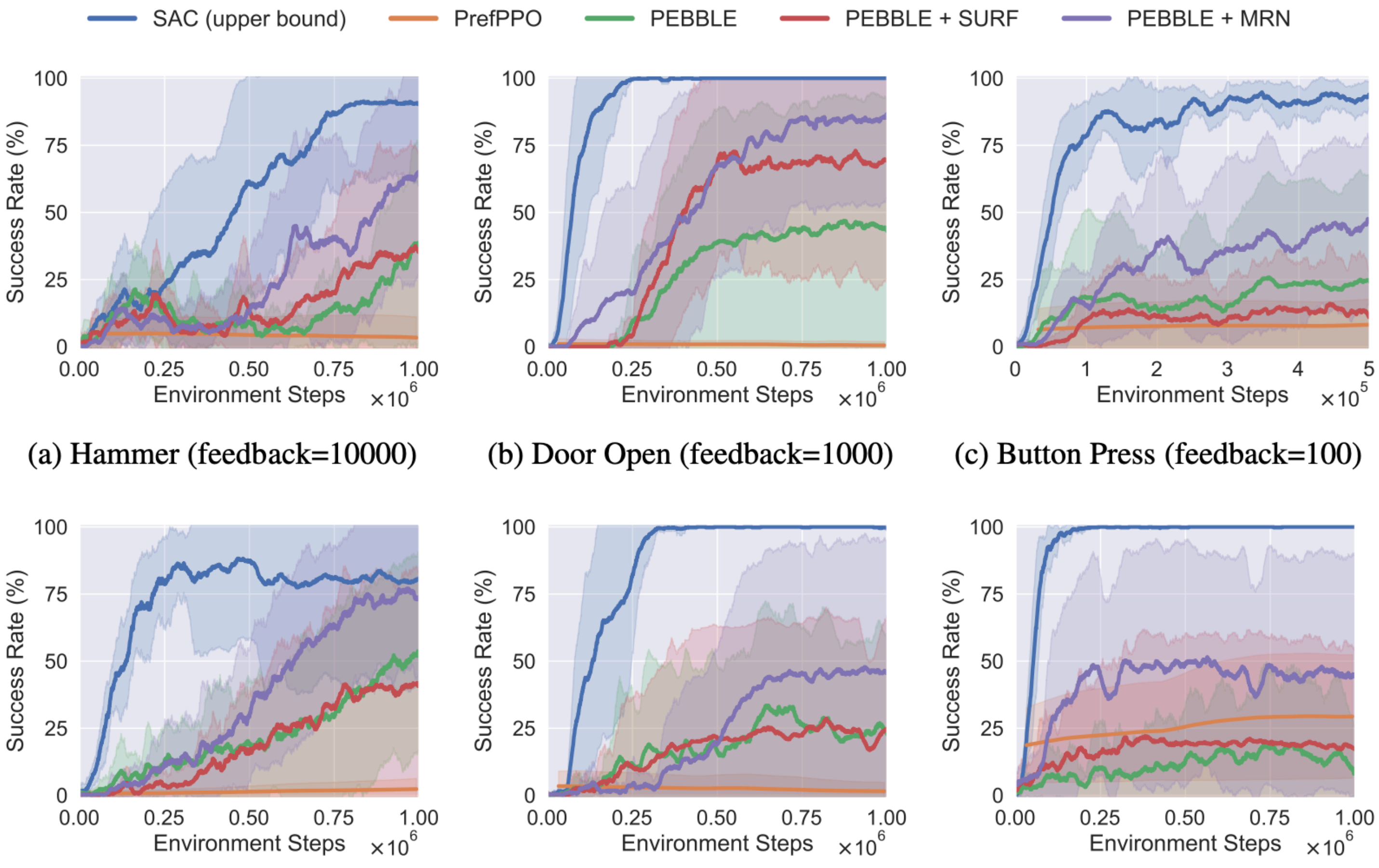
[1] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In CoRL 2020.
[2] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. arXiv preprint arXiv:1801.00690, 2018.
[3] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. Software Impacts, 6:100022, 2020.

Experiments: DeepMind Control suite and Meta world tasks



Video demos https://sites.google.com/view/meta-reward-net

Experiments on Metaworld



(a) Hammer (feedback=10000) (b) Door Open (feedback=1000) (c) Button Press (feedback=100)

(d) Sweep Into (feedback=4000) (e) Drawer Open (feedback=1000) (f) Window Open (feedback=100)

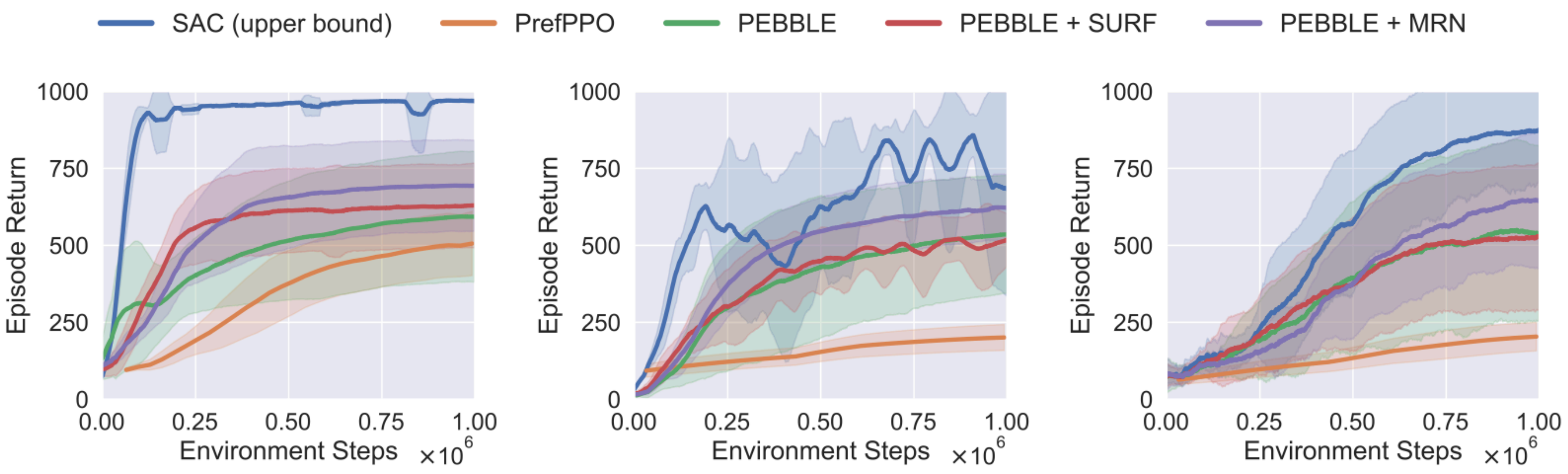
[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In ICML 2018.

[2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In NeurIPS 2017.

[3] Kimin Lee, Laura M Smith, and Pieter Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In ICML 2021.

[4] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In ICLR 2022.

Experiments on DMControl

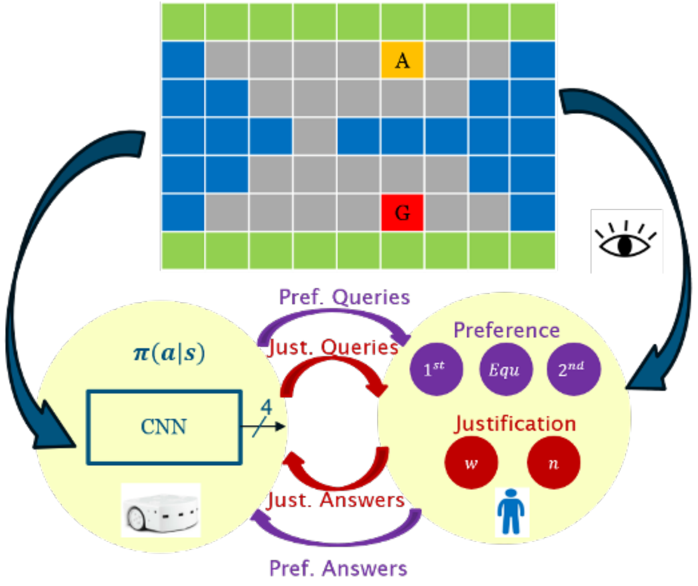


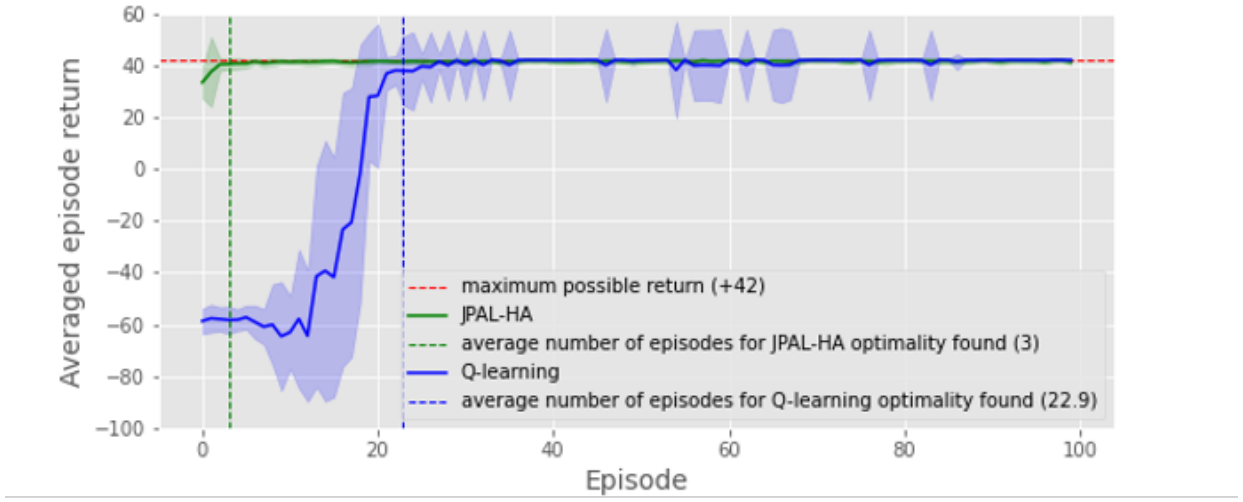
(a) Walker (feedback=100) (b) Cheetah (feedback=100) (c) Quadruped (feedback=700)

[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In ICML 2018.
[2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In NeurIPS 2017.
[3] Kimin Lee, Laura M Smith, and Pieter Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In ICML 2021.
[4] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In ICLR 2022.

Human-in-the loop safe RL [Kazantzidis et al., 2022]

- Safe exploration
  - Safe RL-> Human-in-the-loop safe RL
- Agent alignment
  - Human-in-the-loop RL—> Human-in-the-loop safe RL

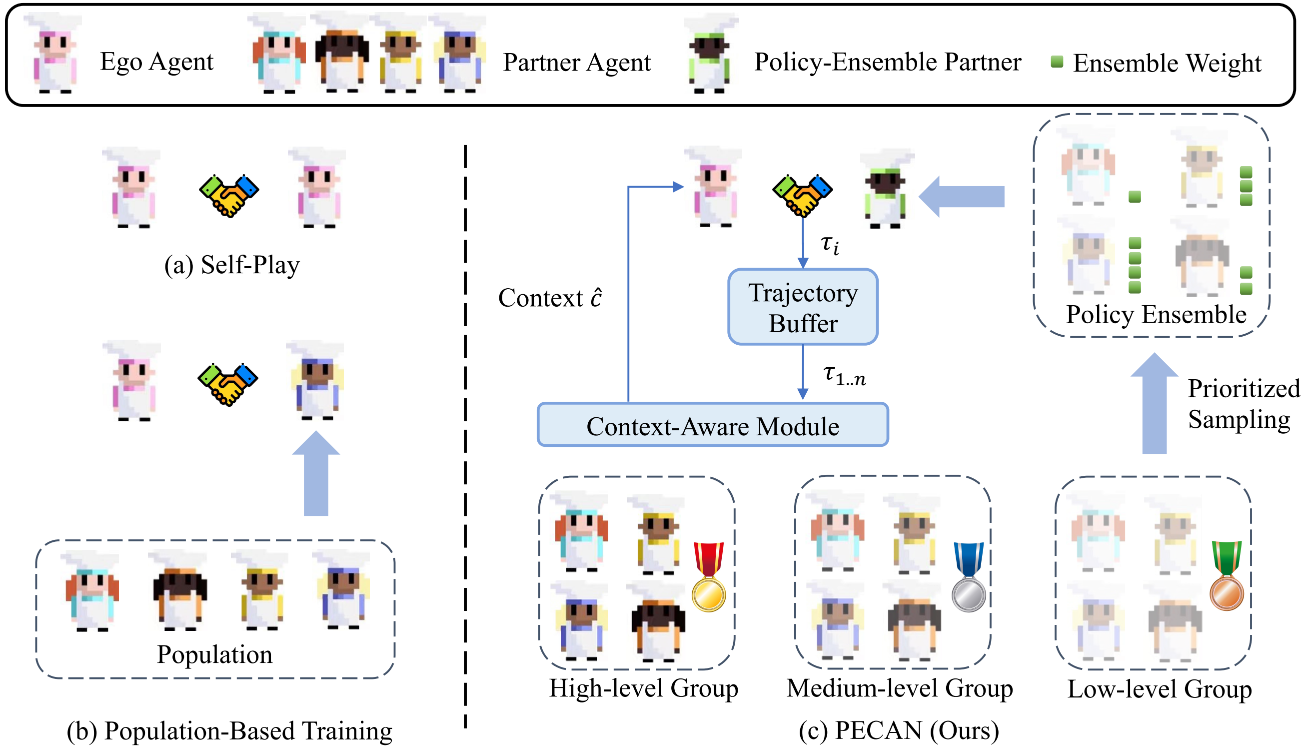


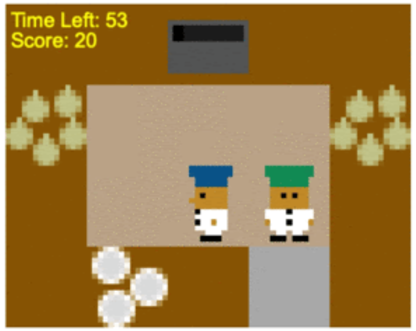


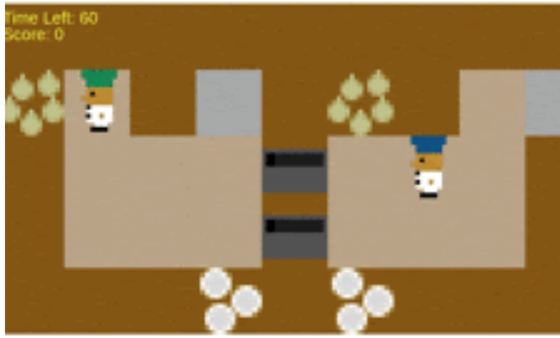
How to train your agent: Active learning from human preferences and justifications in safety-critical environments. AAMAS 2022. Ilias Kazantzidis, Tim Norman, Yali Du, Christopher Freeman

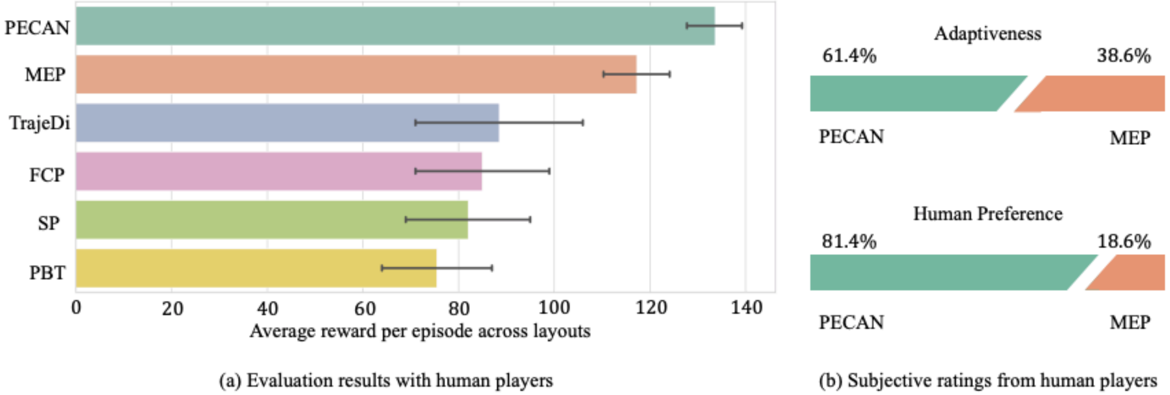
Zero-shot human-AI coordination: Overcooked AI [Lou et al., 2023]

- Improved Population-based training







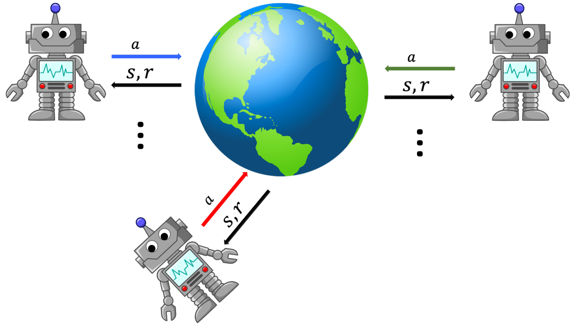


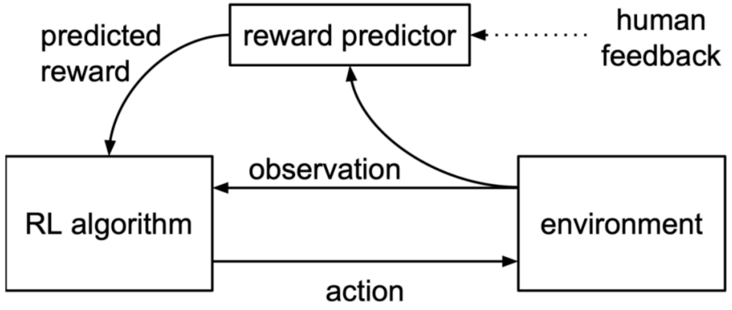
PECAN: Leveraging Policy Ensemble for Context-Aware Zero-Shot Human-AI Coordination. Xingzhou Lou, Jiaxian Guo, Junge Zhang, Jun Wang, Kaiqi Huang and Yali Du. AAMAS 2023

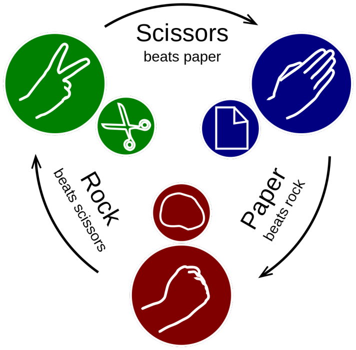
Our lab

- Aim: enable machines to exhibit cooperative and responsible behavior in intelligent decision making tasks.

Cooperative AI Lab



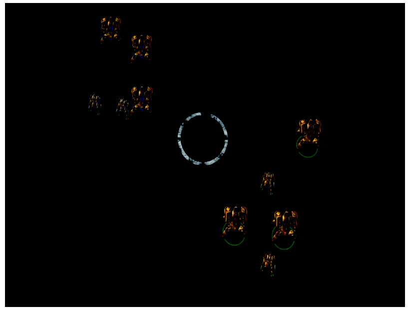


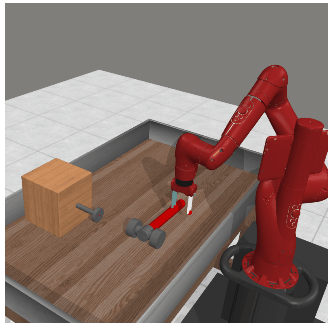


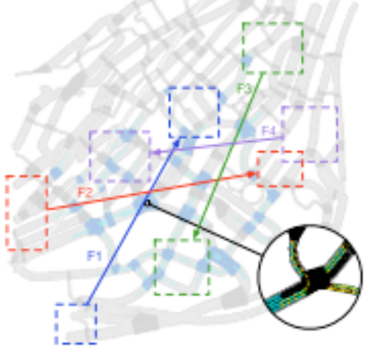
Collaborative Multi-agent learning：
- Cooperation [ICML2019,AAMAS2021-23]
- Credit assignment [NeurIPS 2019]
- Communication [AAMAS2021,2022]

Agent alignment and Safe control:
- Safety control [AAMAS2022, AIJ 2023]
- Morality [NeurIPS2022,ICLR2023]

Efficient evaluation:
- Efficient sampling [ICML2021, AAAI22]
- Capacity of cooperation [ICML2023]







Summary

- This talk
  - Human preferences serves as good alternatives to reward signals.
  - Human-AI teaming has great potential but yet to be explored.
- Next steps
  - Feedback efficiency
  - Potential conflicts among humans
  - Generalisation to new tasks
  - Grounding to physical world
  - …