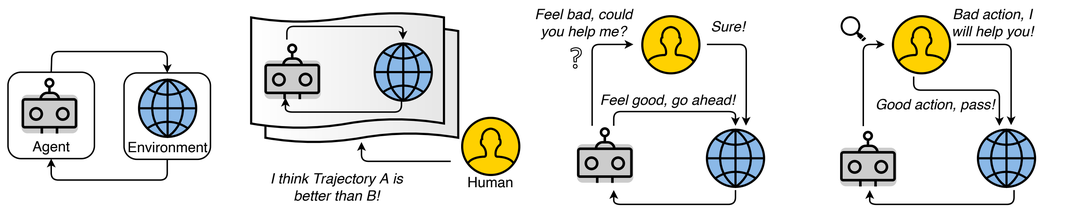
Learning from Active Human Involvement through Proxy Value Propagation

Zhenghao (Mark) Peng[1], Wenjie Mo[1], Chenda Duan[1], Quanyi Li[2], Bolei Zhou[1]
[1]UCLA, [2]University of Edinburgh

NEURAL INFORMATION PROCESSING SYSTEMS

🔍 Problem Formulation & Motivation

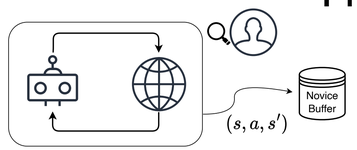


- Reward engineering is hard to encapsulte human intentions.
- **Human-in-the-loop** methods are promising to achieve alignment.
- To ensure safety, **active human involvement** enhances training-time safety.

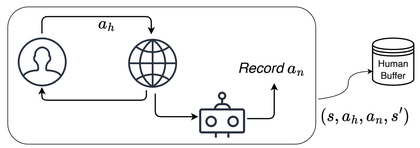
Can we transform a value-based RL algorithm to a reward-free policy trainer learning from online intervention & demonstration from human expert?

Policy learning in 10 minutes w/o reward via human-in-the-loop!

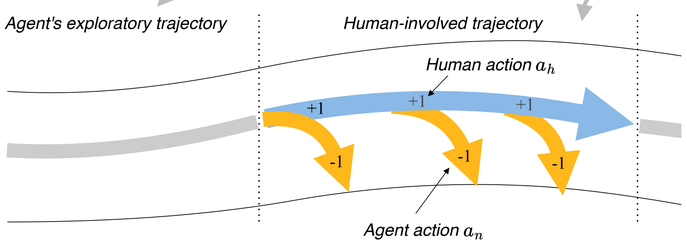
Proxy Value Propagation (PVP)

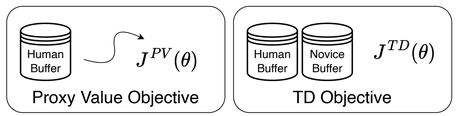


(1) Human oversees agent's exploration



(2) Human intervenes and provides demonstration





(3) Update & propagate proxy values

Proxy Value Objective: ← Apply loss to human-involved states

$$J^{PV}(\theta) = \mathop{\mathbb{E}}_{(s,a_n,a_h)} [\,|\,Q_\theta(s, a_h) - 1\,|^2 + |\,Q_\theta(s, a_n) + 1\,|^2]$$

Temporal Difference Objective: ← Drop the reward term!

$$J^{TD}(\theta) = \mathop{\mathbb{E}}_{(s,a,s')} |\,Q_\theta(s, a) - \gamma \max_{a'} Q_{\hat{\theta}}(s', a')\,|^2$$

Total Objective for Q network:

$$J(\theta) = J^{PV}(\theta) + J^{TD}(\theta)$$

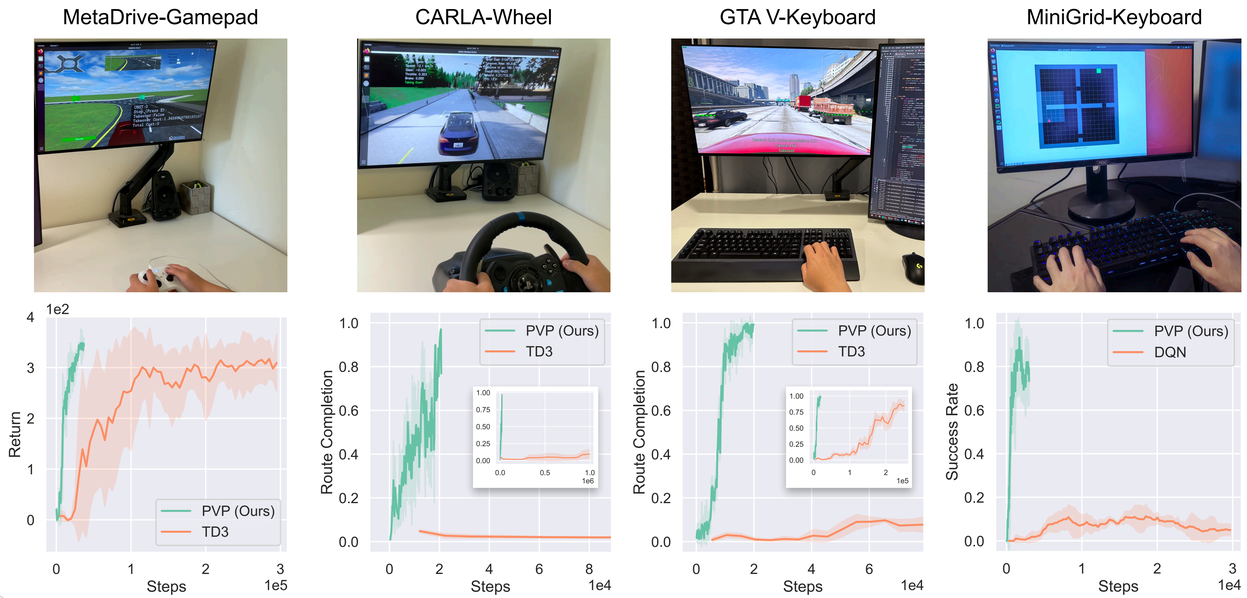
Objective for Policy:

$$J(\phi) = -\mathop{\mathbb{E}}_{s} Q_\theta(s, \pi_\phi(s))$$

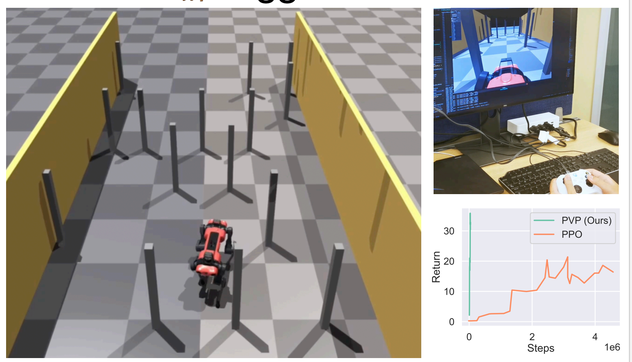
Q_θ - Q network
a_n - agent's action
a_h - human's action
s - state
π_ϕ - agent's policy

Code is available at metadriverse.github.io/pvp

👏 10x Faster than RL Baseline



🐕 Legged Robot



PVP can train a high-level controller for a goal-conditioned low-level policy.

😜 Human Friendly

User Study	HGDAgger	IWR	HACO	PVP
Compliance	3.0	4.0	3.0	**4.8**
Performance	2.2	3.7	3.3	**4.8**
Stress	3.2	4.5	2.3	**4.7**

PVP agents make human feels better (*compliance*), stronger (*performance*) and less stressful (*stress*) in shared control. It also makes human takes over less (right table).

🎉 Safety, Efficiency, Performance

MetaDrive-Keyboard Env	Training			Testing	
	Human Data	Total Data	Total SafetyCost	Episodic SafetyCost	Success Rate
SAC	—	1M	2.76K	0.73	0.82
PPO	—	1M	24.34K	3.41	0.69
TD3	—	1M	1.74K	**0.47**	0.70
SAC-Lag	—	1M	1.84K	0.72	0.73
PPO-Lag	—	1M	11.64K	1.18	0.51
CPO	—	1M	4.36K	1.71	0.21
HumanDemo	**30K**	**-**	**39**	**0.39**	**0.97**
BC	30K	-	—	2.17	0.07
GAIL	30K	2M	25.90K	1.30	0.0
HGDagger	39.0K	51K	56	1.97	0.04
IWR	35.8K	45K	**52**	1.45	0.46
HACO	19.2K	40K	130	1.64	0.13
PVP (Ours)	**14.6K**	40K	76.8	0.89	**0.85**

- Human-in-the-loop makes training much safer, even than Safe RL baselines.
- PVP learns most performant agent.

This work was supported by the National Science Foundation under Grant No. 2235012. The human experiment in this study is approved through the IRB#23-000116 at UCLA.