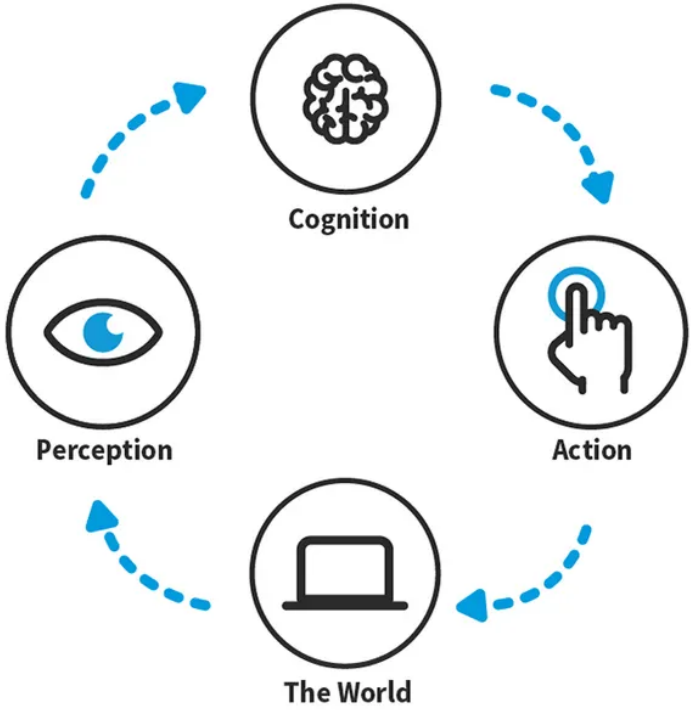
Interactive RL for Robotics

بردیا طوراندار

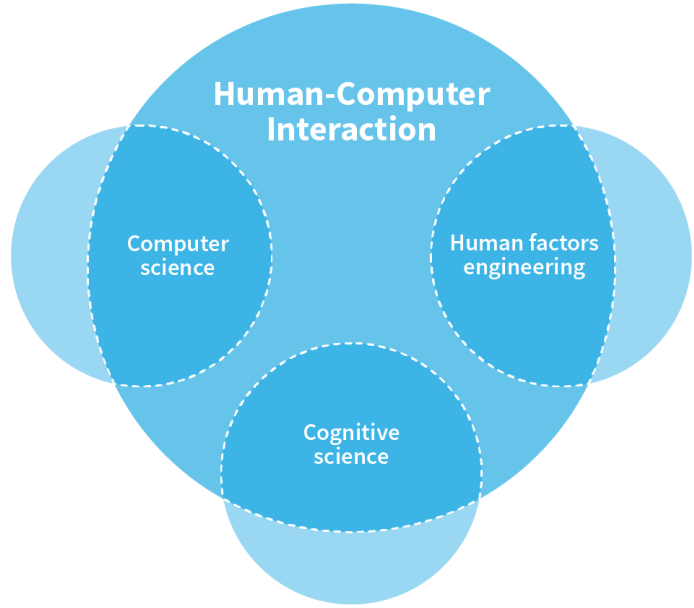
علیرضا عباسی

یادگیری تقویتی تعاملی برای رباتیک

- انسان در حلقه (Human-in-the-loop)
- تعامل انسان و کامپیوتر

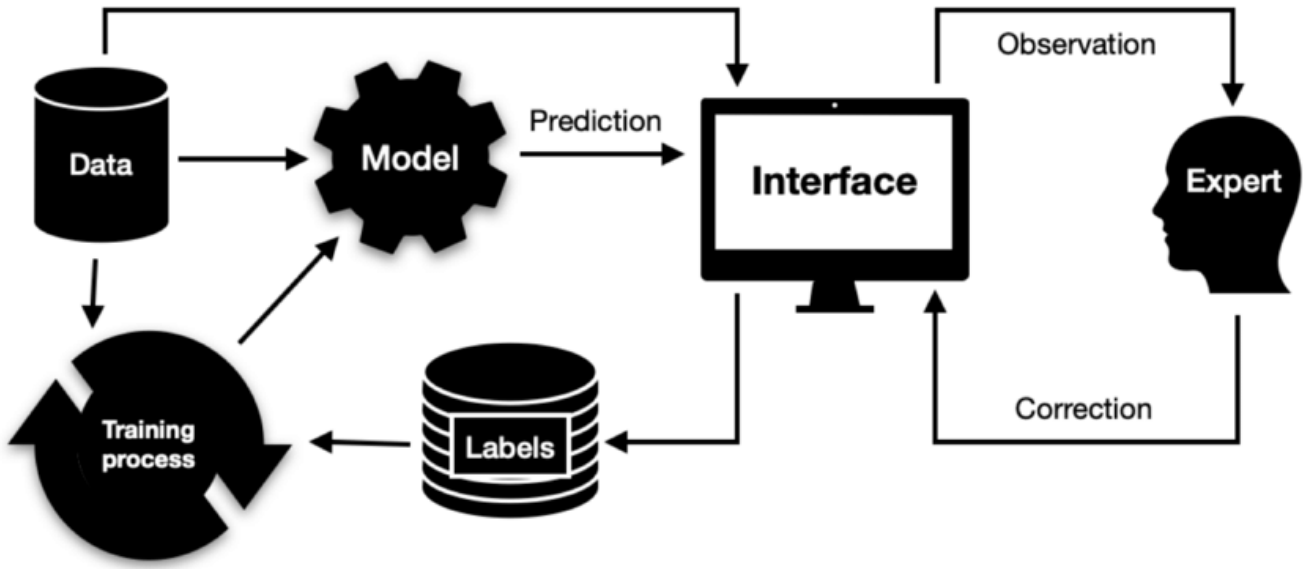


The Multidisciplinary Field of HCI



وظایف مهندسین تعامل انسان و کامپیوتر در زمینه IML

- تعریف تکنیک‌های تعامل انسان و کامپیوتر
- پیدا کردن کابردهای جدید



ویژگیهای IRL

- سریعتر از Automatic RL
- Generalized نبودن پاسخ‌ها
- پیچیدگی تعامل با انسان‌های متفاوت

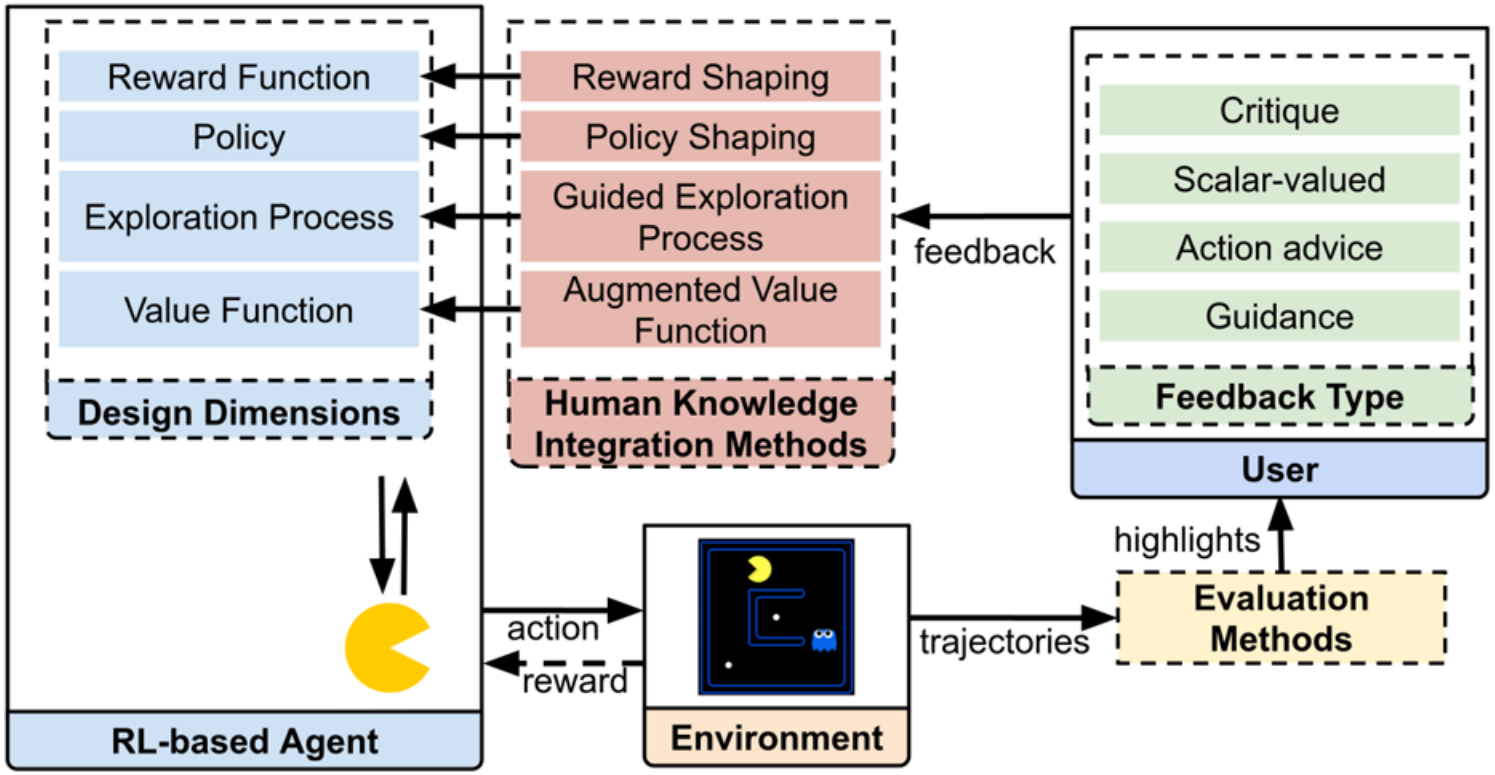
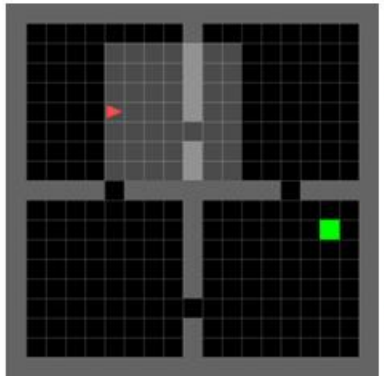


Figure 1. The interactive RL architecture.

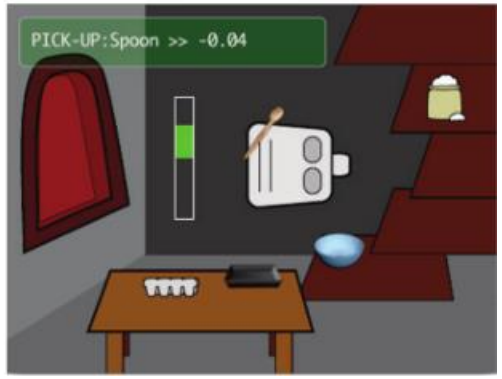
محیط‌های آزمایشی یادگیری تقویتی تعاملی

- رباتیک
- هوش مصنوعی در بازی‌ها
- تعامل انسان و کامپیوتر









ابعاد طراحی

- تابع پاداش: فیدبک انسانی (توسط شخص خبره)، جهت تسریع یادگیری
- سیاست بهینه
- کاوش
- تابع ارزش

$R' = R + F$, where $F : S \times A \times S \to \mathbb{R}$ is the *shaping reward*

Design Dimension	Testbed	Interaction	Initiative	HKI	Feedback	Algorithms
Reward Function	Robot in maze-like environment [10]	FE	Passive	RS using HF + ER	Critique	DQL [67, 77]
	Navigation simulation [12]	GUI	Passive	Advantage Function	Critique	DAC
	Sophie's Kitchen game [105, 103]	GUI	Passive, Active	RS using HF + EF	Critique	QL [109], HRL
	Bowling game [108]	GUI	Passive	RS + HF	Scalar-valued	DQL
	Shopping assistant, GridWorld [76]	GUI	Active	Active IRD	Queries	Model-based RL
	Mario, GridWorld, Soccer simulation [85]	Coding	Passive	PBRS	Heuristic Function	QL, QSL, QL(λ), QSL(λ)
	Navigation simulation [102]	VC	Passive	RS using HF + ER	AcAd	SARSA [86], SARSA(λ)
	Atari, robotics simulation [20]	GUI	Active	RS using HF	Queries	DRL
Policy	GridWorld, TurtleBot robot [71]	GUI, GC	Passive	PS	AcAd	AC(λ) [15, 91]
	GridWorld [56]	VC	Passive	PS	Critique, AcAd	BQL [26]
	Pac-Man, Frogger [37]	GUI	Passive	PS	Critique	BQL
Exploration Process	Pac-Man, Cart-Pole simulation [114]	GUI	Passive	GEP	AcAd	QL
	Simulated cleaning Robot [24, 23]	VC	Passive	GEP	AcAd	SARSA
	Pac-Man [7]	GUI	Active	GEP	AcAd	SARSA(λ)
	Pac-Man [32]	GUI	Active	Myopic Agent	AcAd	QL, QRL
	Sophie's Kitchen game [103]	GUI	Active	ACTG	Guidence	QL
	Street Fighter game [13]	Not apply	Passive	EB using Safe RL	Demonstration	HRL
	Nao Robot [93]	GUI	Passive	ACTG	Guidence	QL
	Nexi robot [54]	AT + CT	Passive	Myopic Agent	AcAd	SARSA(λ)
Value Function	Mountain Car simulation [52]	GUI	Passive	Weighted VF	Demonstration	SARSA(λ)
	Keepaway simulation [101]	GUI	Passive	Weighted VF	Demonstration	SARSA
	Mario, Cart Pole [18]	Not apply	Passive	Initialization of VF	Demonstration	QL(λ)

ابعاد طراحی

- طراحی پاداش:

The reward shaping (RS) method aims to mold the behavior of a learning agent by modifying its reward function to encourage the behavior the RL designer wants.

- طراحی سیاست:

The policy shaping (PS) approach consists of directly molding the policy of a learning agent to fit its behavior to what the RL designer envisions.

- فرآیند کاوش هدایت‌شده:

Guided exploration process methods aim to minimize the learning procedure by injecting human knowledge to guide the agent’s exploration to states with a high reward.

- تابع ارزش تقویت‌شده:

The procedure to augment a value function consists of combining the value function of the agent with one created from human feedback.

مسائل تحقیقاتی

High-dimensional Environments
محیط‌های با ابعاد بالا

Lack of Evaluation Techniques
کمبود روش‌های ارزیابی

Lack of Human-like Oracles
فقدان مشاوران شبیه به انسان

Modeling Users
مدل‌سازی کاربران

Combining Different Design Dimensions
ترکیب ابعاد مختلف طراحی

Safe Interactive RL
یادگیری تقویتی تعاملی ایمن

Fast Evaluation of Behaviors
ارزیابی سریع رفتارها

Explainable Interactive RL
یادگیری تقویتی تعاملی قابل توضیح

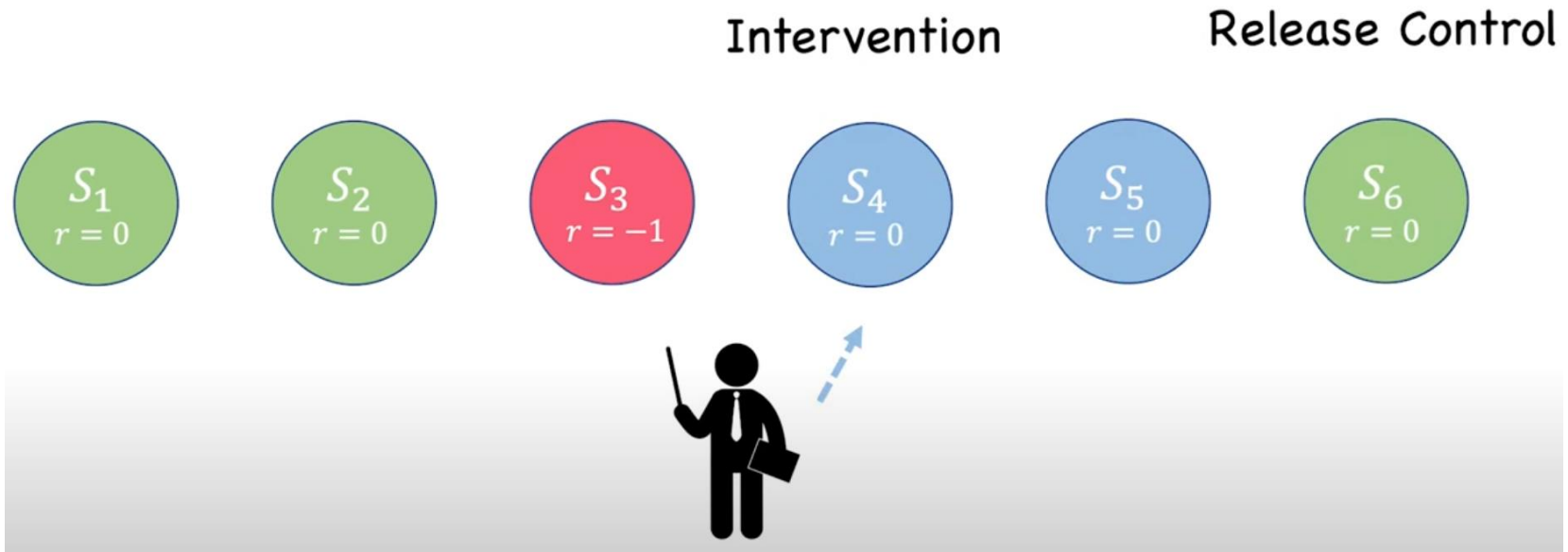
یادگیری تقلیدی تعاملی در یادگیری تقویتی

- اعمال فیدبک توسط انسان درصورت بهینه نبودن پالیسی
- اعمال یادگیری تقویتی با فیدبک تعاملی بعنوان سیگنال ریوارد

Algorithm 2 RLIF

Require: π, π^{exp}, D

for trial $i = 1$ to N **do**
  Train π on D via reinforcement learning.
  for timestep $t = 1$ to T **do**
    if π^{exp} intervenes at t **then**
      label (s_{t-1}, a_{t-1}, s_t) with -1 reward, append to D_i
    else
      label (s_{t-1}, a_{t-1}, s_t) with 0 reward, append to D_i
    end if
  end for
  $D \leftarrow D \cup D_i$
end for

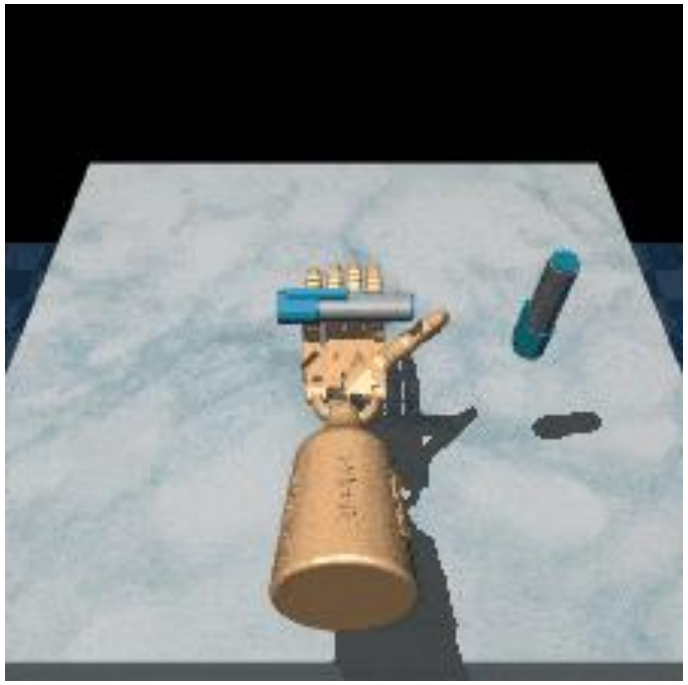


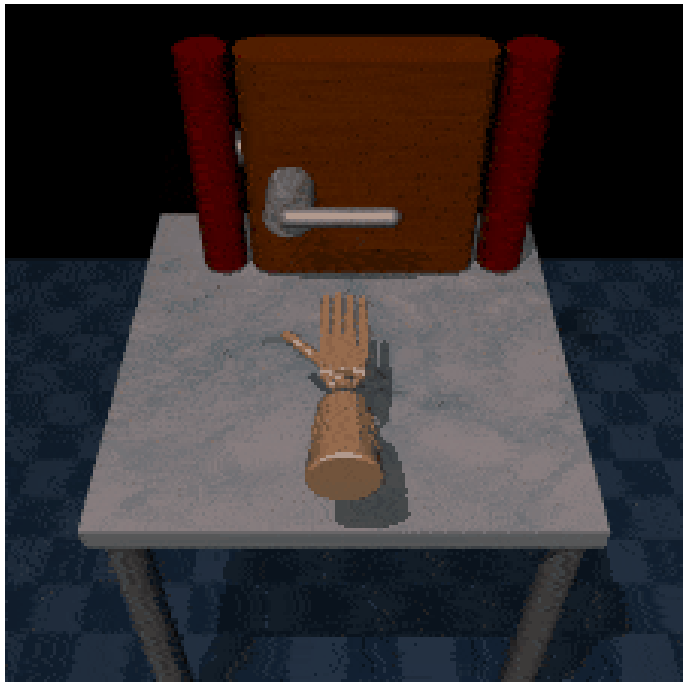
International Conference on Learning Representations (ICLR) 2024

https://rlif-page.github.io

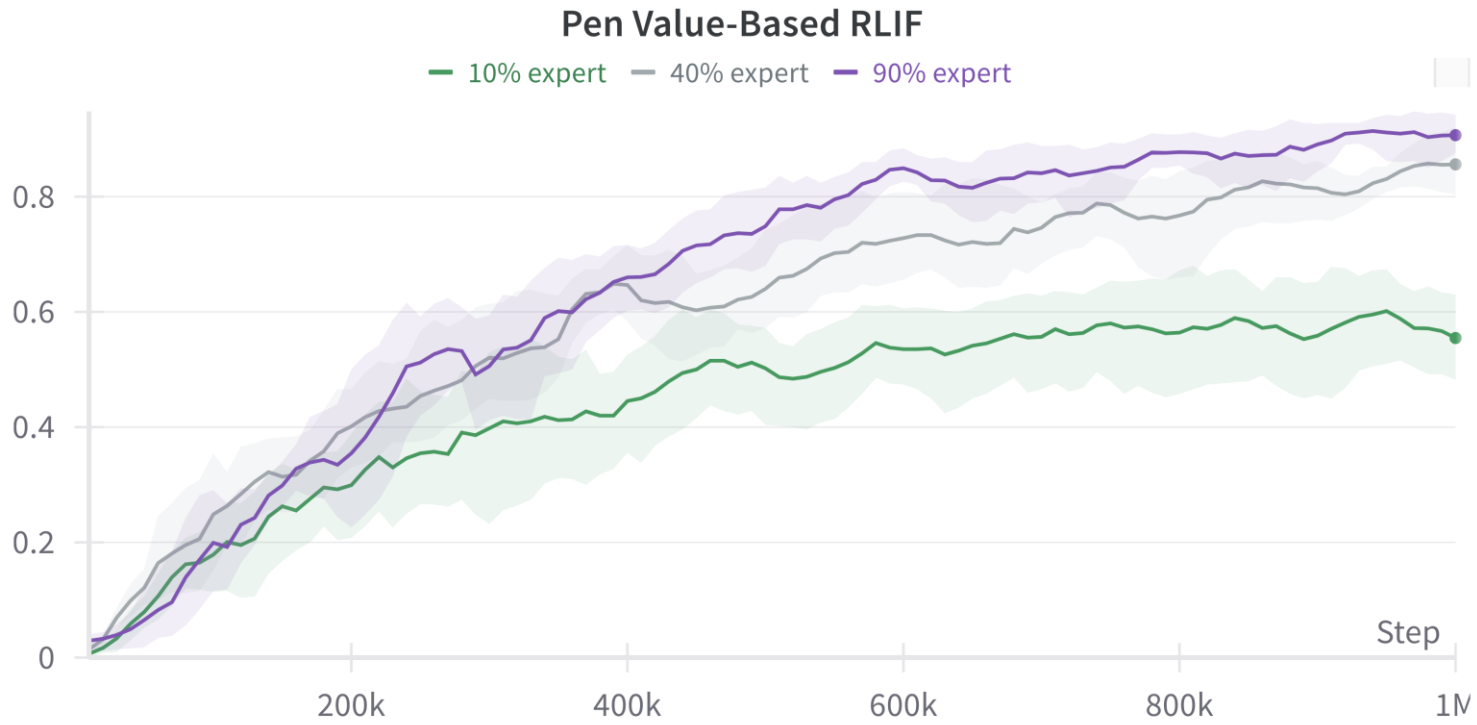
منبع:

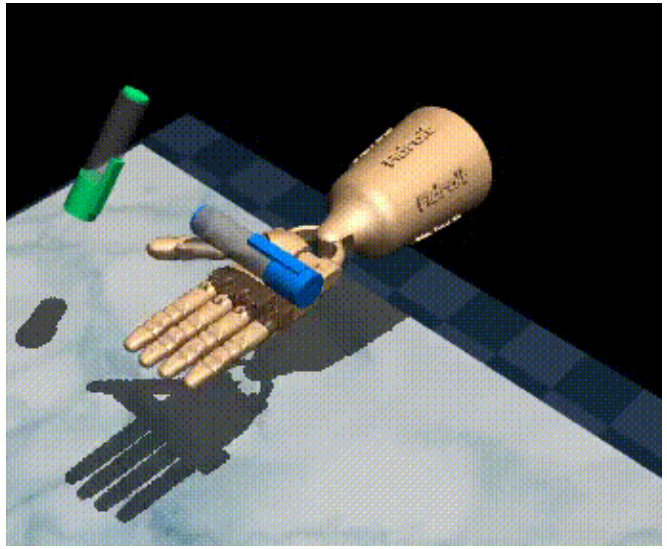
یادگیری تقلیدی تعاملی در یادگیری تقویتی





- D4RL : ابزار متن–باز برای بنچ–مارک الگوریتم‌های آفلاین یادگیری تقویتی
- محیط Adroit Pen





یادگیری تقلیدی تعاملی در یادگیری تقویتی

- استراتژی مداخله:
  - random : مداخله در ۳۰ و ۵۰ و ۸۵ درصد زمان‌ها
  - بر پایه value : مقایسه با پالیسی خبره (expert)

$$\mathbb{P}(Intervention|s) = \begin{cases} \beta, & \text{if } Q^{\pi^{\text{ref}}}(s, \pi^{\text{exp}}(s)) * \alpha > Q^{\pi^{\text{ref}}}(s, \pi(s)) \\ 1-\beta, & \text{otherwise.} \end{cases}$$

We choose a value for β close to 1 such as 0.95 and a value of α close to 1 such as 0.97.

- پالیسی خبره:
  - ترین روی دیتاست D4RL در ابعاد کوچک‌تر (۱۰ و ۴۰ و ۹۰ درصد)

Tasks	Dataset	Subsampled Size
Adroit Pen	pen-expert-v1	50 trajectories

RLIF: Interactive Imitation Learning as Reinforcement Learning

یادگیری تقلیدی تعاملی در یادگیری تقویتی

Jianlan Luo* Perry Dong* Yuexiang Zhai Yi Ma Sergey Levine

International Conference on Learning Representations (ICLR) 2024

Vienna, Austria

- نتیجه شبیه‌سازی

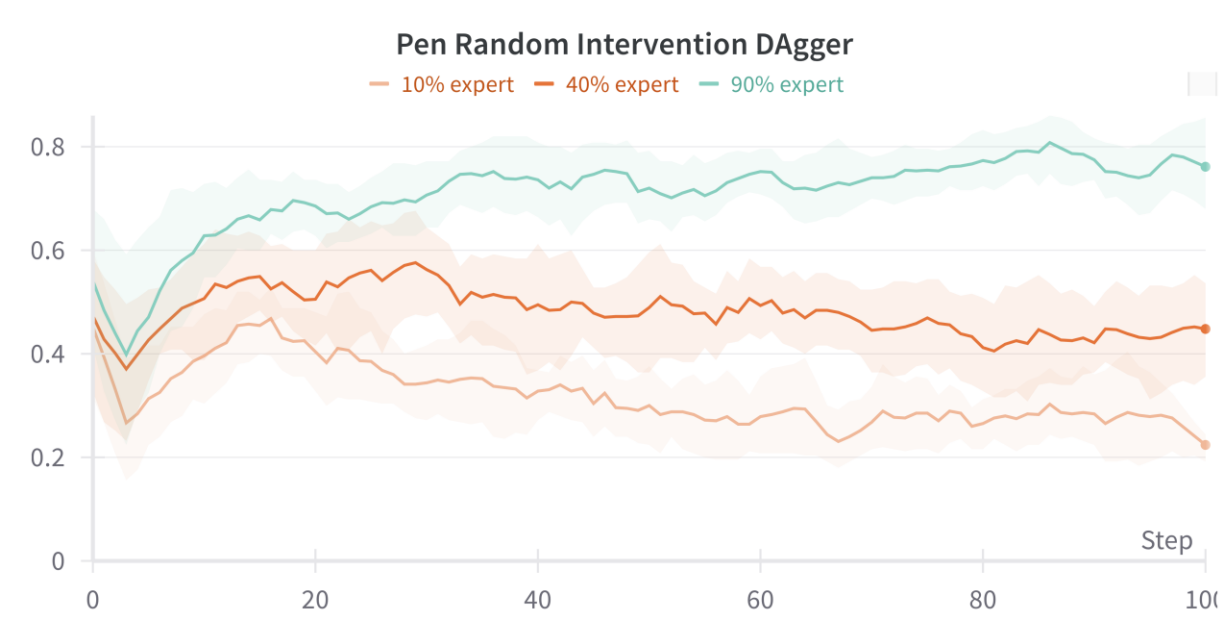
Paper Code

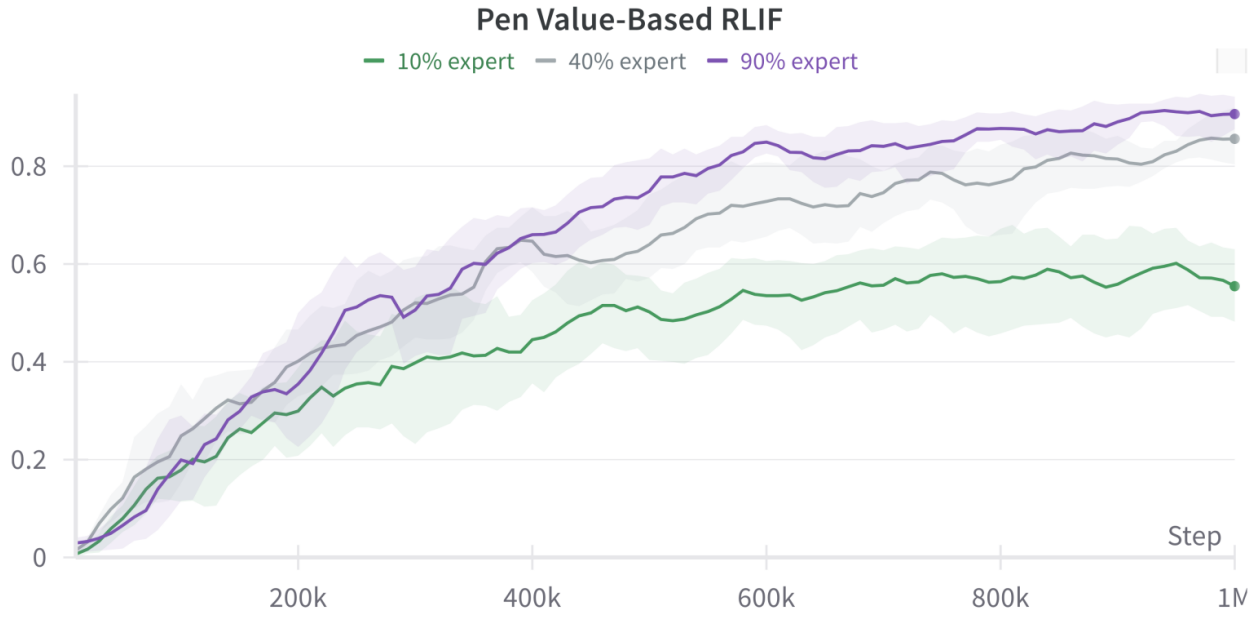
Domain	Expert Level	RLIF with Value Based Intervention	RLIF with Random Intervention	HG-DAgger	HG-DAgger with 85% Random Intervention	DAgger	DAgger with 85% Random Intervention	BC
adroit-pen	∼90%	**88.47**	42.87	73.47	74.27	78.13	79.07	
	∼40%	**80.87**	34.13	60	29.33	35.73	38.67	54.13
	∼10%	**64.04**	28.33	28.53	9.47	8.93	12.8	
	average	**77.79**	35.11	54	37.69	40.93	43.51	54.13
locomotion-hopper	∼110%	108.99	106.51	53.55	**112.7**	57.94	76.13	
	∼70%	**99.66**	75.62	44.75	69.73	20.49	43.59	44.46
	∼20%	**102.85**	19.11	11.94	19.66	12.37	20.1	
	average	**103.83**	67.08	36.75	67.36	30.27	46.61	44.46
locomotion-walker	∼110%	**109.17**	93.76	80.3	86.93	70.58	61.64	
	∼40%	**108.42**	103.9	40.66	42.65	38.7	19.63	64.77
	∼15%	**108.01**	75.12	25.2	24.37	19.54	10.29	
	average	**108.53**	90.93	48.72	51.32	42.94	30.46	64.77

یادگیری تقلیدی تعاملی در یادگیری تقویتی

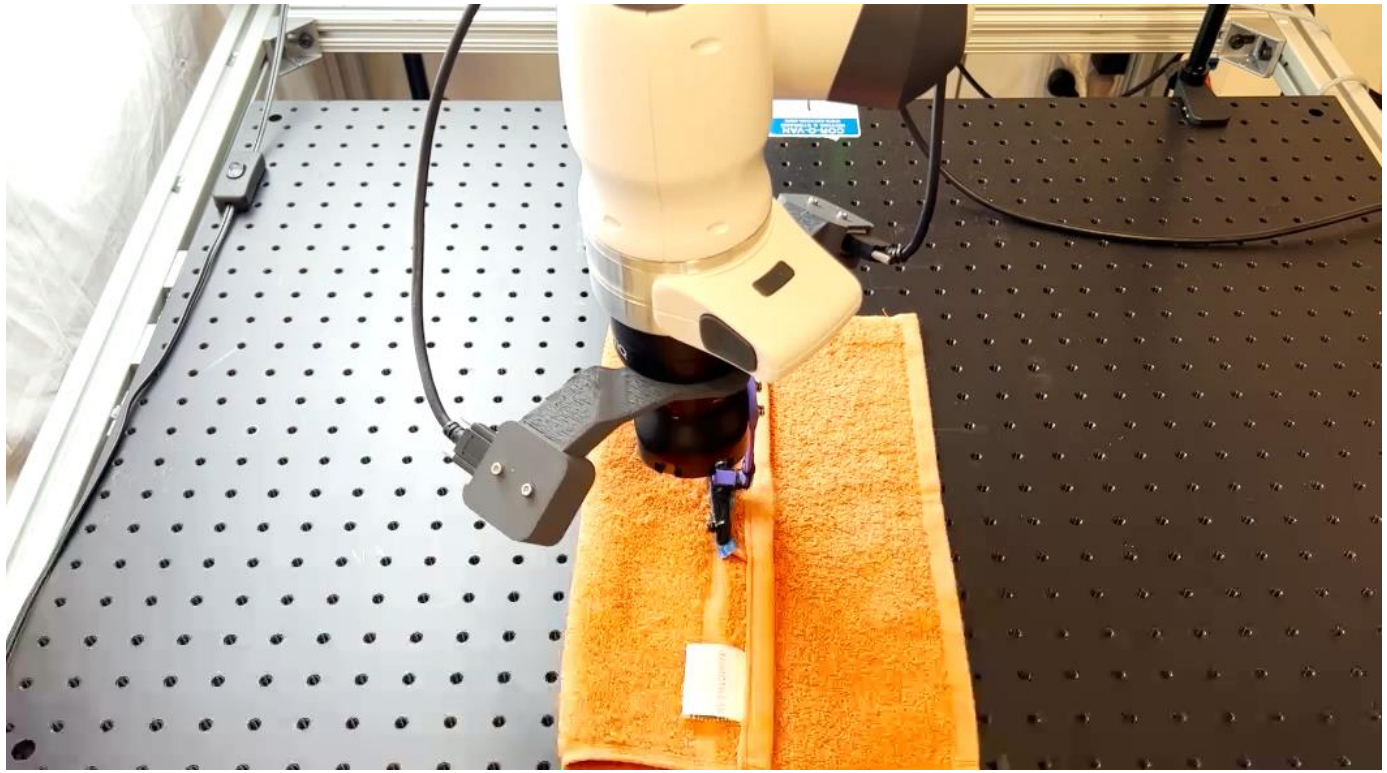
```
bash scripts/run_pen_hgdagger.sh
bash scripts/run_pen_value_based_rlif.sh
```

- نتیجه شبیه‌سازی





یادگیری تقلیدی تعاملی در یادگیری تقویتی



- اعمال روی ربات حقیقی
  - داینامیک غیرپیوسته و غیرقابل پیش‌بینی برخورد
  - ورودی سنگین (تصویر دوربین)
  - اعمال روی پلتفرم با شکل متغیر (حوله)
  - اپراتور مداخله‌گر (انسان) غیر ایده‌آل
- انتخاب ریوارد بصورت دستی بسیار دشوار است.
- نتیجه: ۹۵درصد نرخ موفقیت در ۲۰ مسیر سنجش