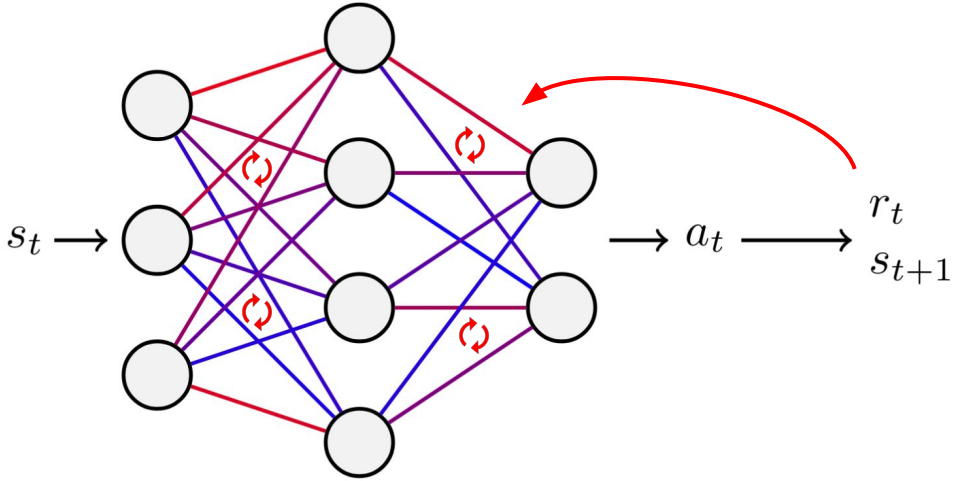
MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

Robust In-Context Reinforcement Learning Under Reward Poisoning Attacks

Paulius Sasnauskas [†] [1] [2] Yiğit Yalın [3] Goran Radanović [3]

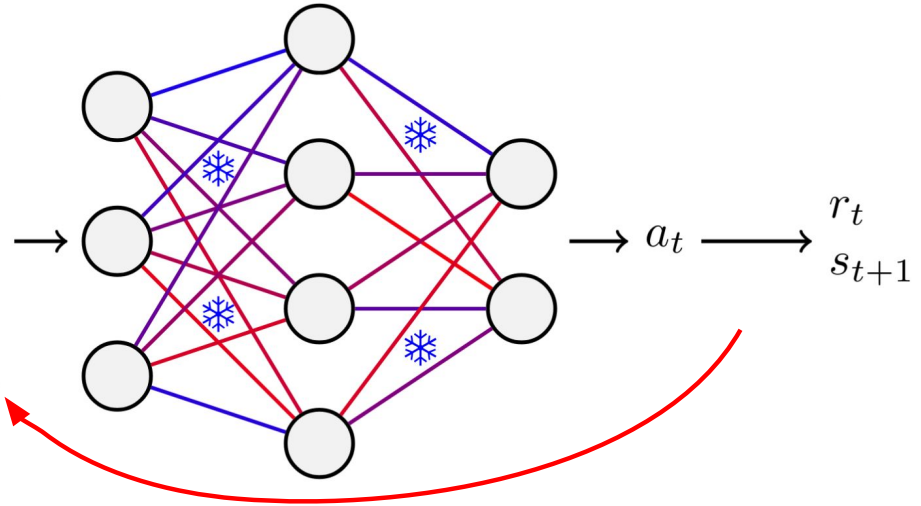
[†] This work was done as a part of an internship project at MPI-SWS
[1] Department of Computing Science, University of Alberta, Edmonton, Canada
[2] Alberta Machine Intelligence Institute (Amii), Edmonton, Canada
[3] MPI-SWS, Saarbrücken, Germany
Correspondence to: Paulius Sasnauskas <sasnausk@ualberta.ca>

Single-Task RL:

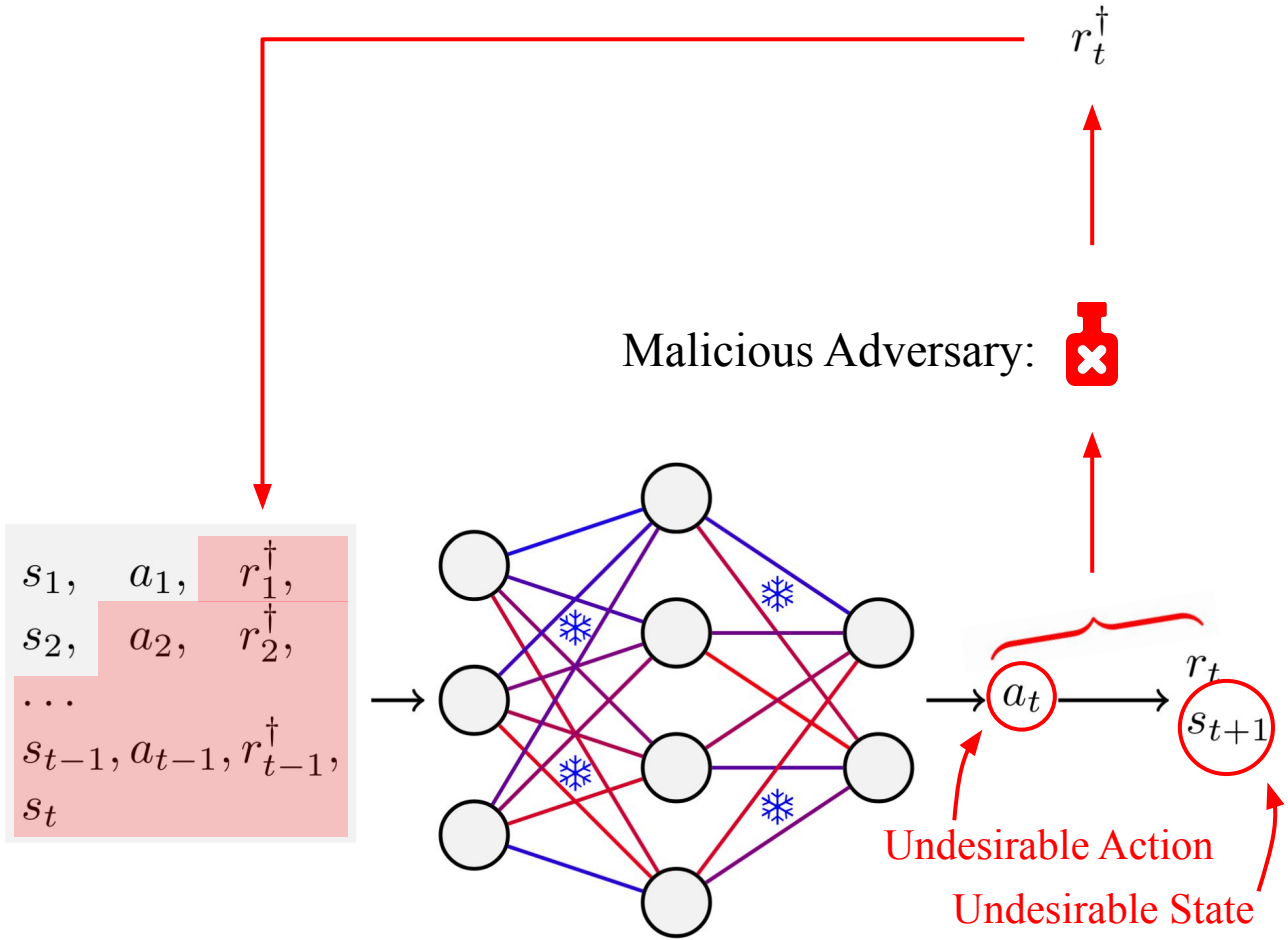


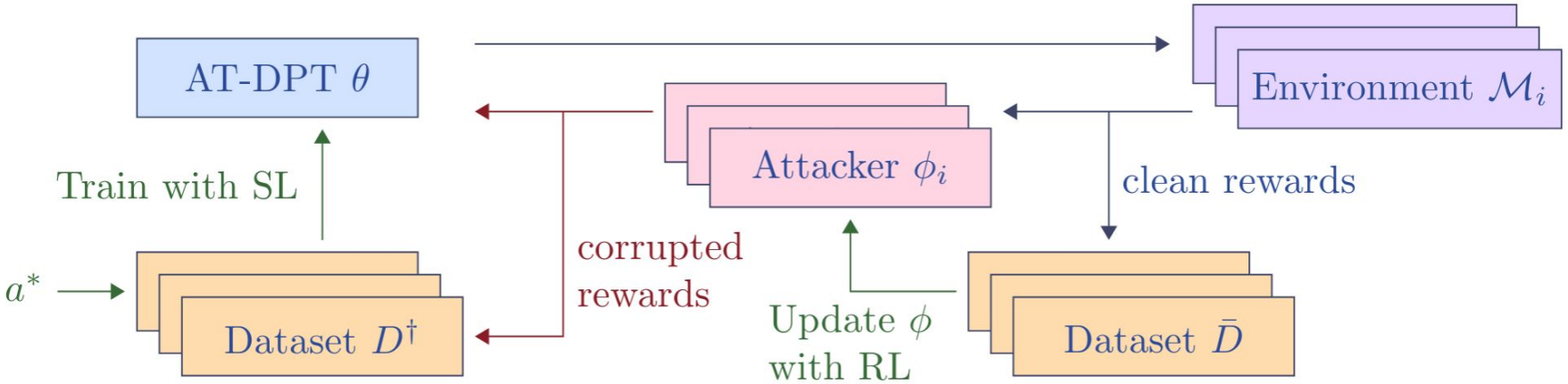
In-Context RL:

$$s_1, \quad a_1, \quad r_1,$$
$$s_2, \quad a_2, \quad r_2,$$
$$\dots$$
$$s_{t-1}, a_{t-1}, r_{t-1},$$
$$s_t$$



Poisoned In-Context RL:




AT-DPT θ
Environment $\mathcal{M}_i$
Attacker ϕ_i
Train with SL
clean rewards
corrupted
rewards
a^*
Dataset $D^\dagger$
Update ϕ
with RL
Dataset $\bar{D}$

Cumulative regret in the **bandit** setting (lower is better)

Algorithm	Attacker Target		Unif. Rand. Attack	Clean Env.
	AT-DPT	TS		
AT-DPT	24.2 ± 1.2	29.8 ± 3.0	38.7 ± 1.7	13.0 ± 0.9
AT-DPT (sub. 30%)	41.2 ± 2.9	45.7 ± 3.3	47.8 ± 3.5	25.9 ± 3.3
DPT	63.6 ± 8.6	62.0 ± 8.6	37.2 ± 1.2	11.5 ± 0.5
TS	106.3 ± 3.8	94.3 ± 3.8	34.2 ± 1.6	8.7 ± 0.6
RTS*	102.9 ± 4.5	90.4 ± 4.4	33.9 ± 1.6	10.2 ± 0.4
UCB1.0	104.1 ± 3.6	90.6 ± 4.1	38.1 ± 2.2	16.0 ± 0.5
crUCB*	86.0 ± 4.4	82.0 ± 4.4	31.8 ± 1.6	15.8 ± 0.5

* We fine-tune parameters for RTS and crUCB outperforming the default ones.

bandit
bandit, adaptive attacker
linear bandit
MDP gridworld (darkroom)
MDP high-dimensional (miniworld)