Exploiting Demonstration Vulnerabilities: Malicious Agents in Inverse Reinforcement Learning

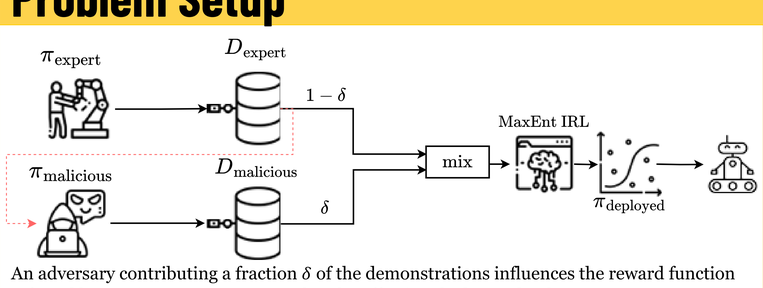
Arezoo Alipanah
aalipanah@uwaterloo.ca

Yash Vardhan Pant
yash.pant@uwaterloo.ca

Abstract

- Inverse Reinforcement Learning (IRL) infers reward functions from demonstrations.
- Real-world demonstrations are often imperfect.
- We investigate: Can a small fraction of adversarial demonstrations degrade IRL performance?
- Key finding: 10% malicious demonstrations distort the learned reward & reduce policy performance.

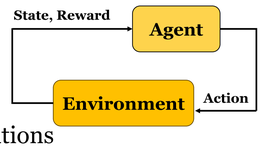
Problem Setup



An adversary contributing a fraction δ of the demonstrations influences the reward function inferred by Maximum Entropy IRL, thereby affecting the learned policy $\pi_{deployed}$.

Introduction

- Reinforcement Learning (RL) depends on carefully designed rewards, which are often difficult to specify.
- IRL infer rewards from expert demonstrations
- Real-world demonstrators are suboptimal, increasing IRL's vulnerability
- Research Question: Can adversaries who subtly perturb demonstrations corrupt IRL training?
- IRL assumes that demonstrations reflect the expert's genuine intent, but such assumptions break under adversarial manipulation



UNIVERSITY OF WATERLOO

Methodology

- We model adversarial behavior as an optimization problem:

$$\min_{x(s,a)^t} \sum_{s,a,t} \gamma^t x(s,a)^t \phi(s)^T \omega$$

Subject to

$$\left\| \sum_{s,a,t} x(s,a)^t \phi(s) - \hat{\phi}^E(s) \right\|_2$$

$$\sum_a x(s,a)^t = \sum_{s',a'} P(s|s',a') x(s',a')^{t-1}$$

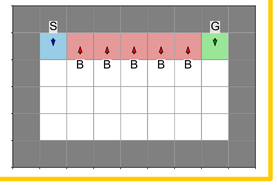
$$\forall s \in S, \forall t \in \{1,\dots,T\}$$

$$\sum_a x(s,a)^0 = \mathbb{1}\{s = s_0\}, \qquad \forall s \in S.$$

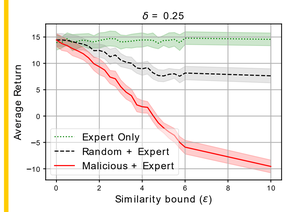
- The adversary generates deceptive trajectories indistinguishable (up to ϵ) from expert demonstrations but minimizes the true return.
- We compare against a random noisy policy constrained to the same ϵ-bounded feature counts.

Experiments

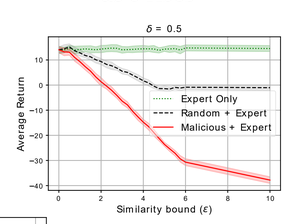
- Cliff World Environment was used for experiments



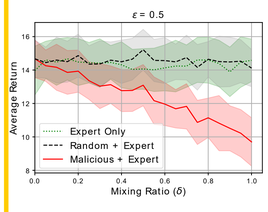
Results



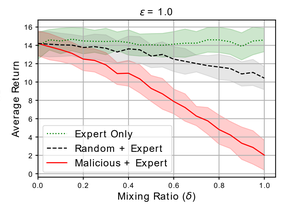
- Return decreases near-linearly with δ & ϵ.
- Fixed δ=0.25 → Return declines as ϵ rises



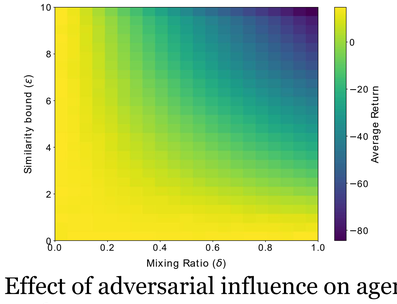
- Fixed δ=0.5 → Return declines as ϵ rises



- Fixed ϵ=0.5→ Return degrades as δ increases.



- Fixed ϵ=1.0→ Return degrades as δ increases.



- Effect of adversarial influence on agent performance

Discussion

Cliff World Experiments:

- Dataset: 1,000 demonstrations (mix of expert + adversary).
- Mixing ratio (δ) = fraction of adversarial demonstrations.
- Perturbation budget: ϵ.
- Key findings:10% malicious demonstrations results in ~10 unit drop in average return.
- Larger ϵ increases the impact.
- When adversary mimics expert features closely, degradation is mitigated even at high mix-ratios.

Conclusions

- Even small adversarial inputs can corrupt learned rewards.
- IRL algorithms need robustness mechanisms for deployment in real-world, adversarial settings.

Future Work

- Study adversarial defenses (e.g., filtering demonstrations, robust IRL variants).
- Explore other domains beyond Cliff World.

CPI Graduate Student Conference (CPI GradConf 2025)