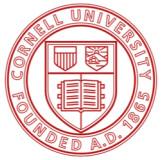
Adversarial Sequential Decision Making

Goran Radanović, Adish Singla, Wen Sun, Xiaojin Zhu

International Joint Conference on AI (IJCAI) 2022

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

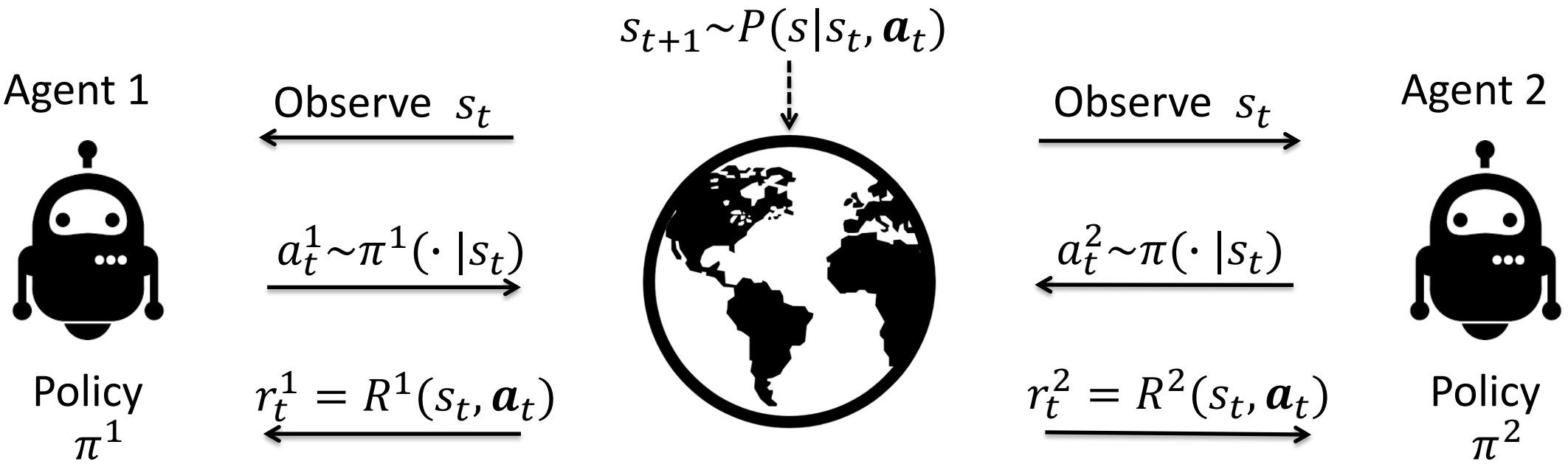


25th July, 2022

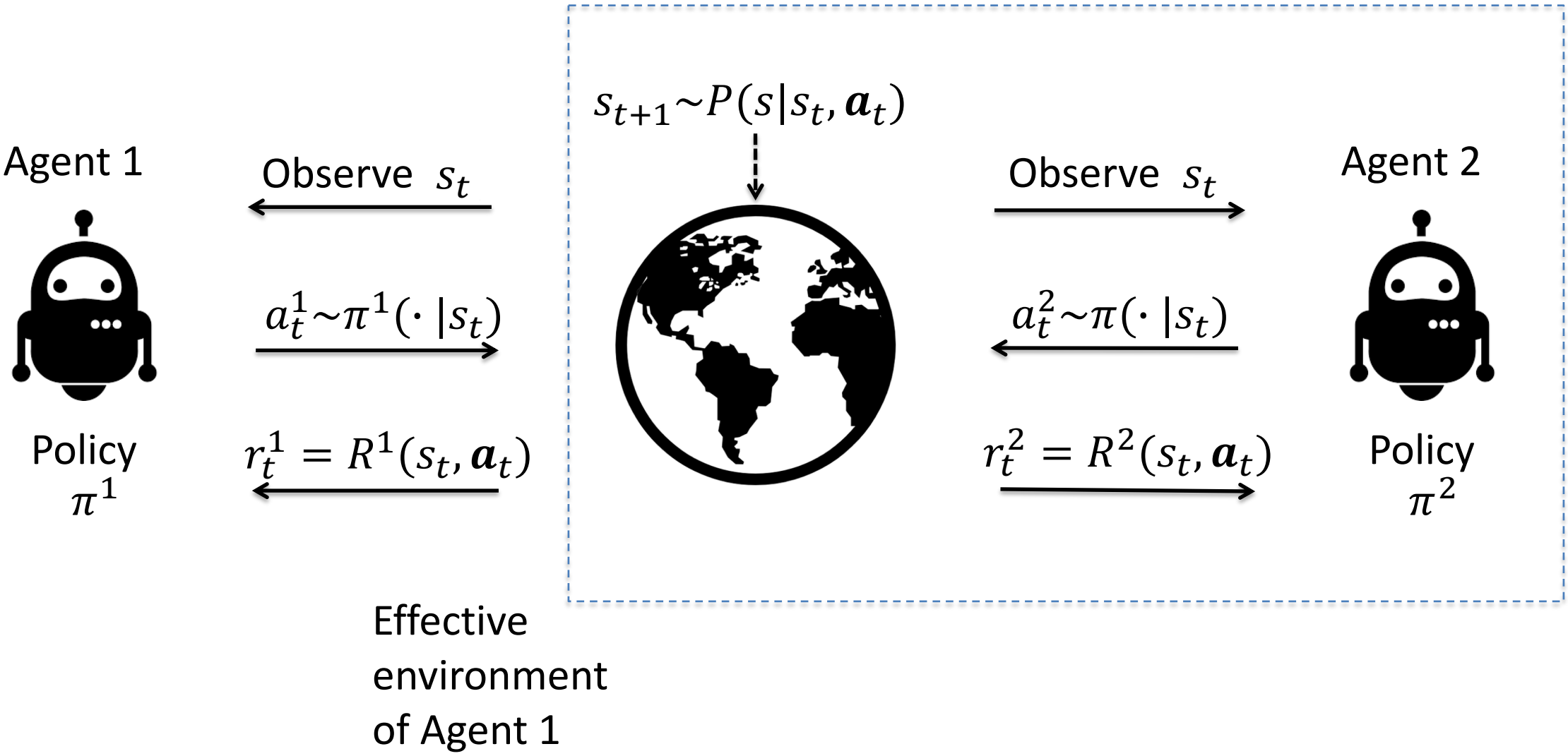
Outline

- Preliminaries
- Test-time Attacks and Defenses in RL
- Training-time Attacks in RL
- Training-time Defenses in RL
- **Adversarial Attacks in Multi-agent RL**
- Concluding Remarks

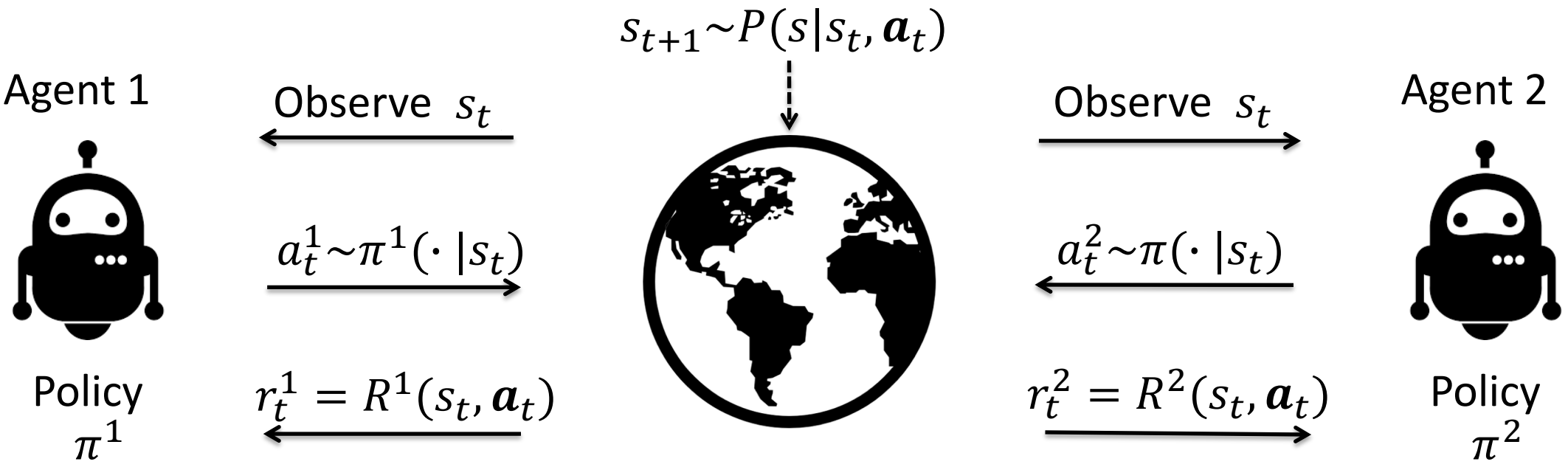
Markov Games



Markov Games

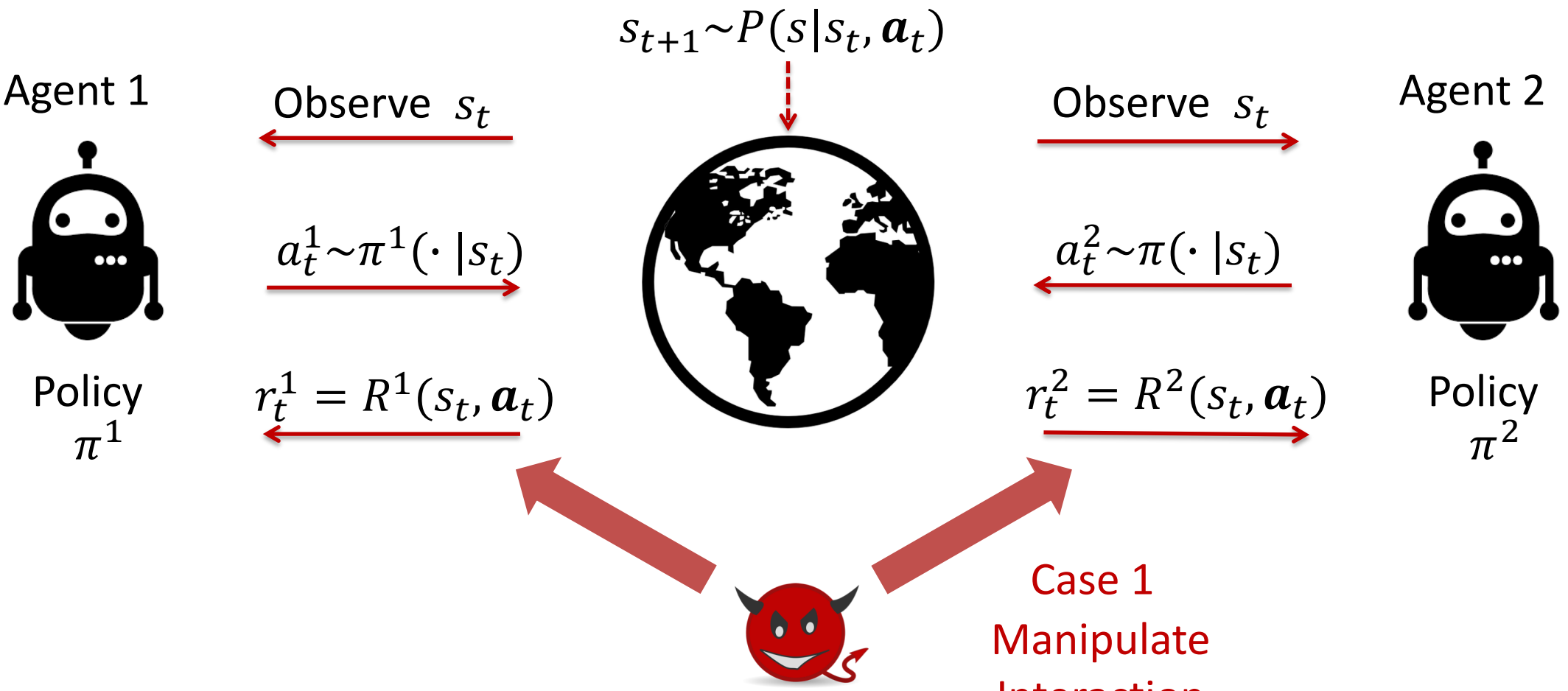


Markov Games

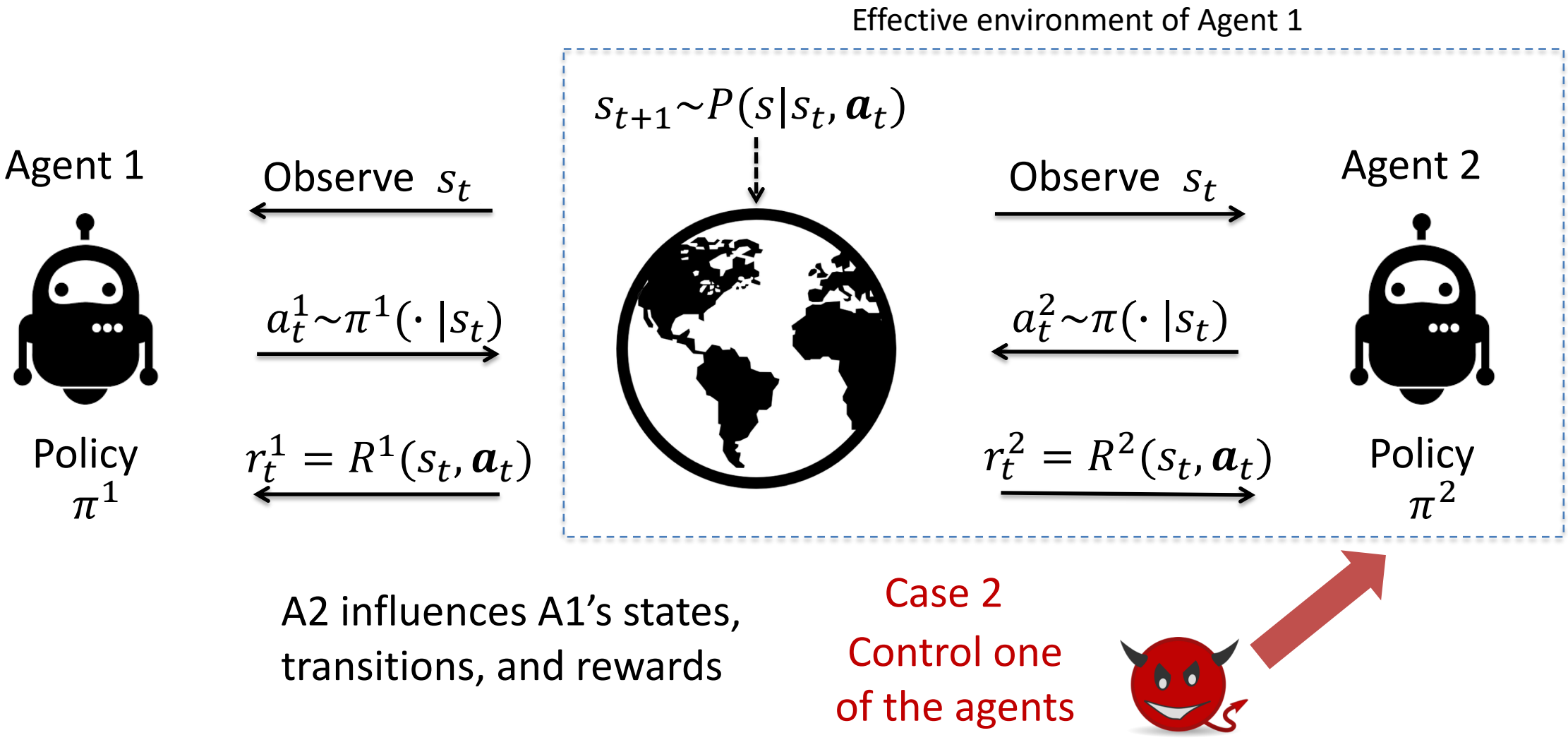


Agents optimize their returns ⇒ find an equilibrium

Adversarial Multi-Agent Setting



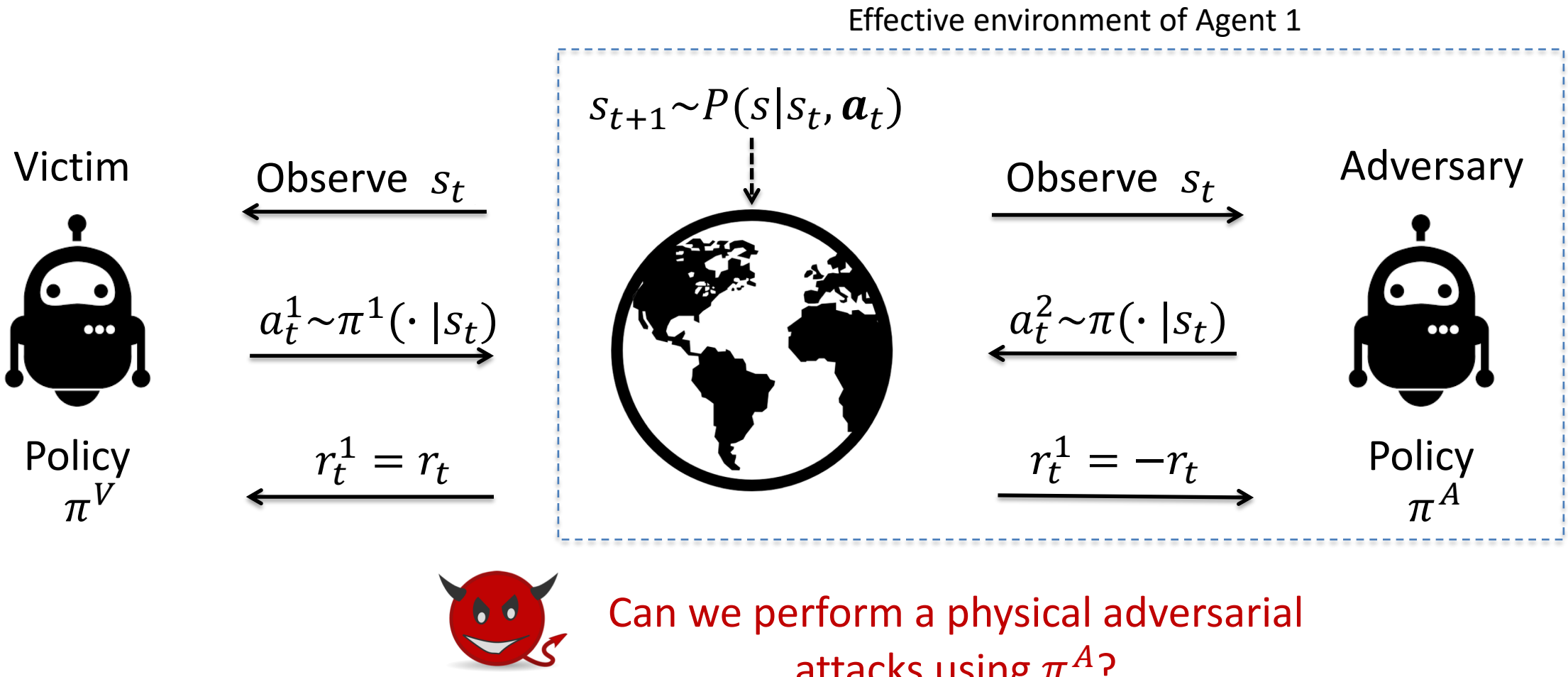
Adversarial Multi-Agent Setting



A2 influences A1's states, transitions, and rewards

Case 2
Control one of the agents

Adversarial Policies



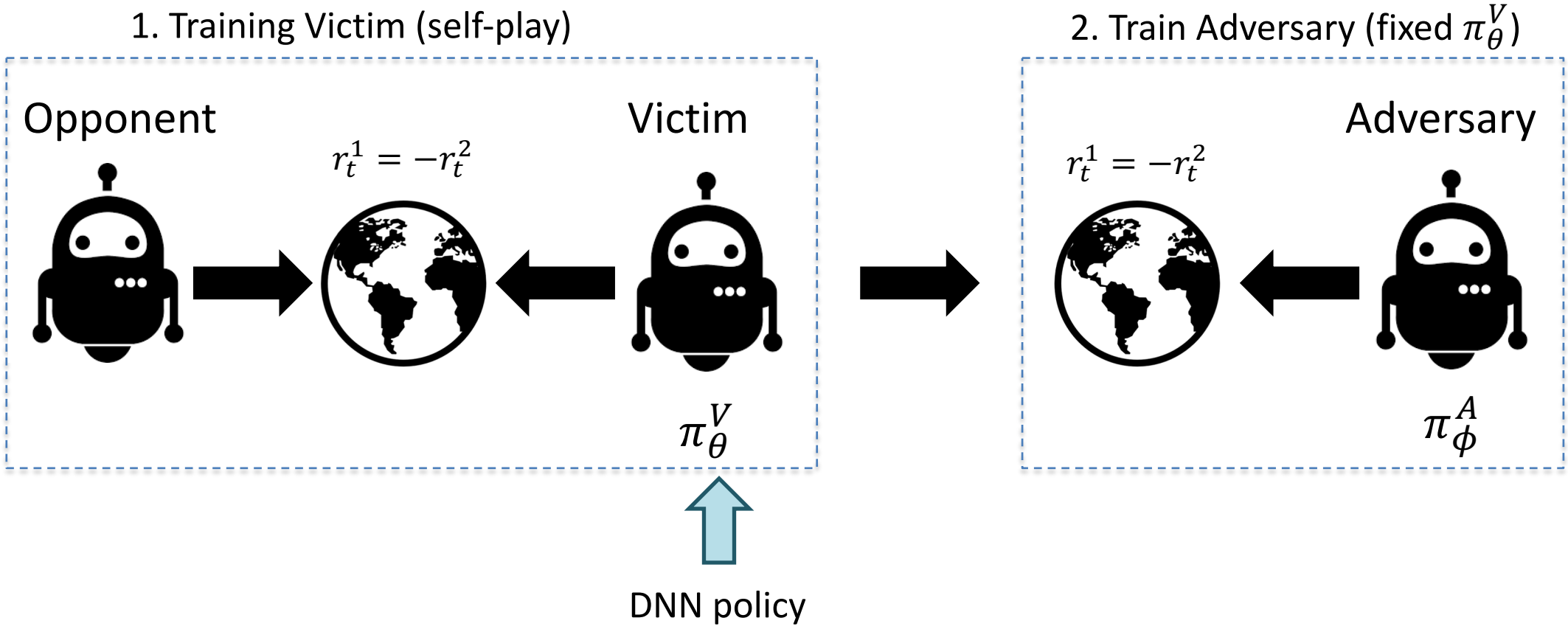
Can we perform a physical adversarial attacks using π^A?

[Gleave et al., 2020]

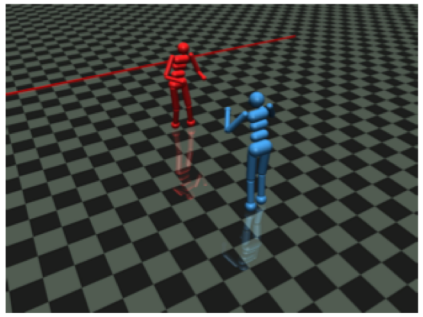
See also [Guo et al., 2021]

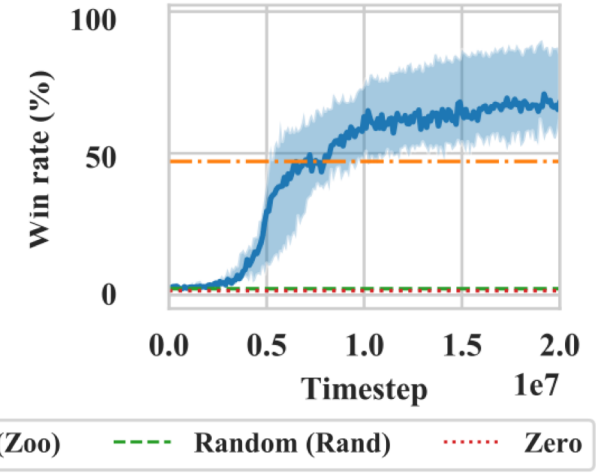
Adversarial Policies

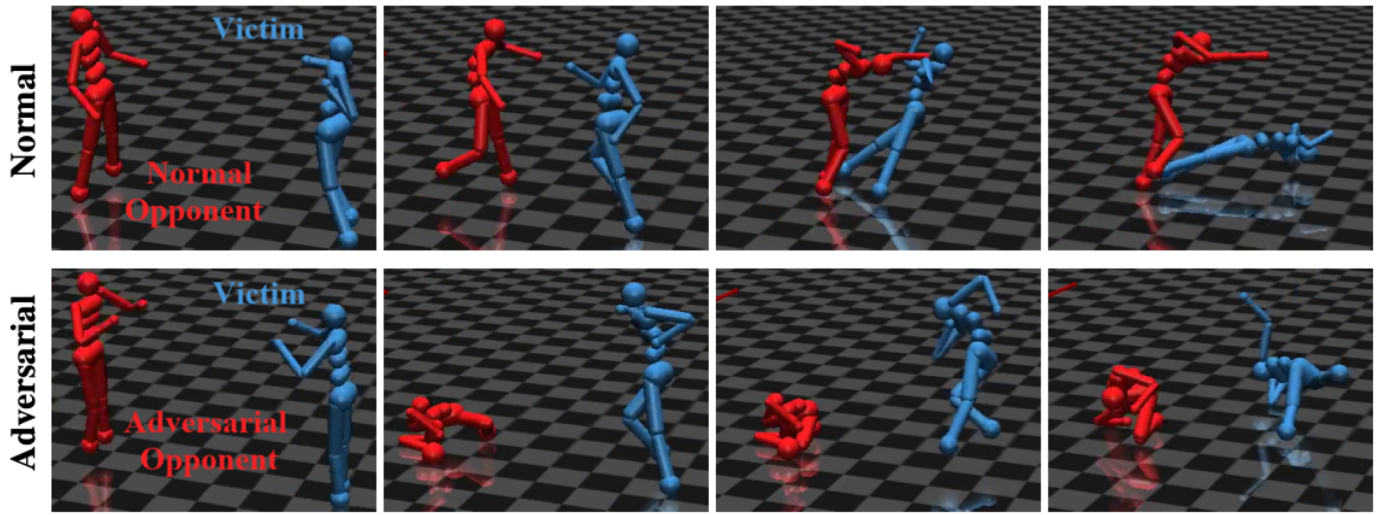
Test-time attack



Adversarial Policies





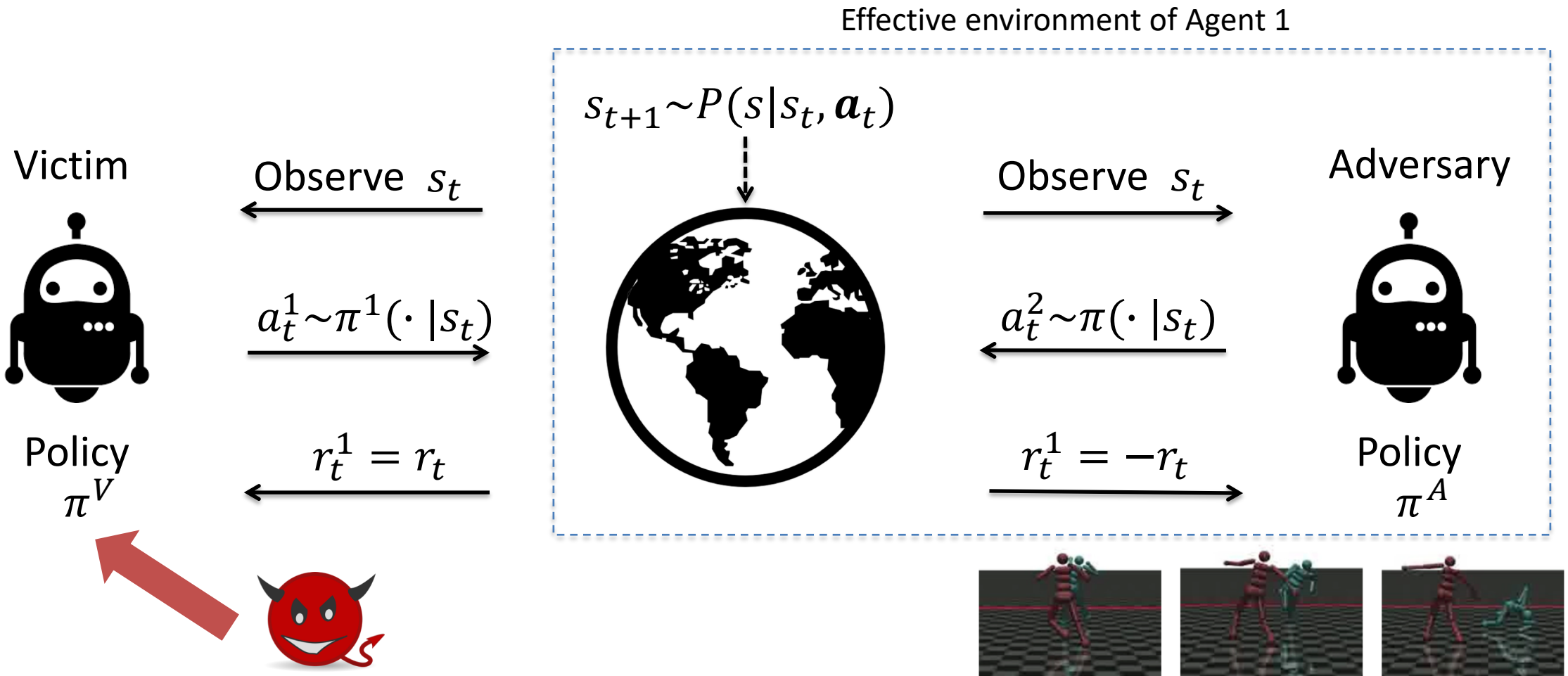


Adversary trained with < 3% or time-steps used for training Victim.

Masking adversary's position helps: the victim's win rate increases.

[Gleave et al., 2020]

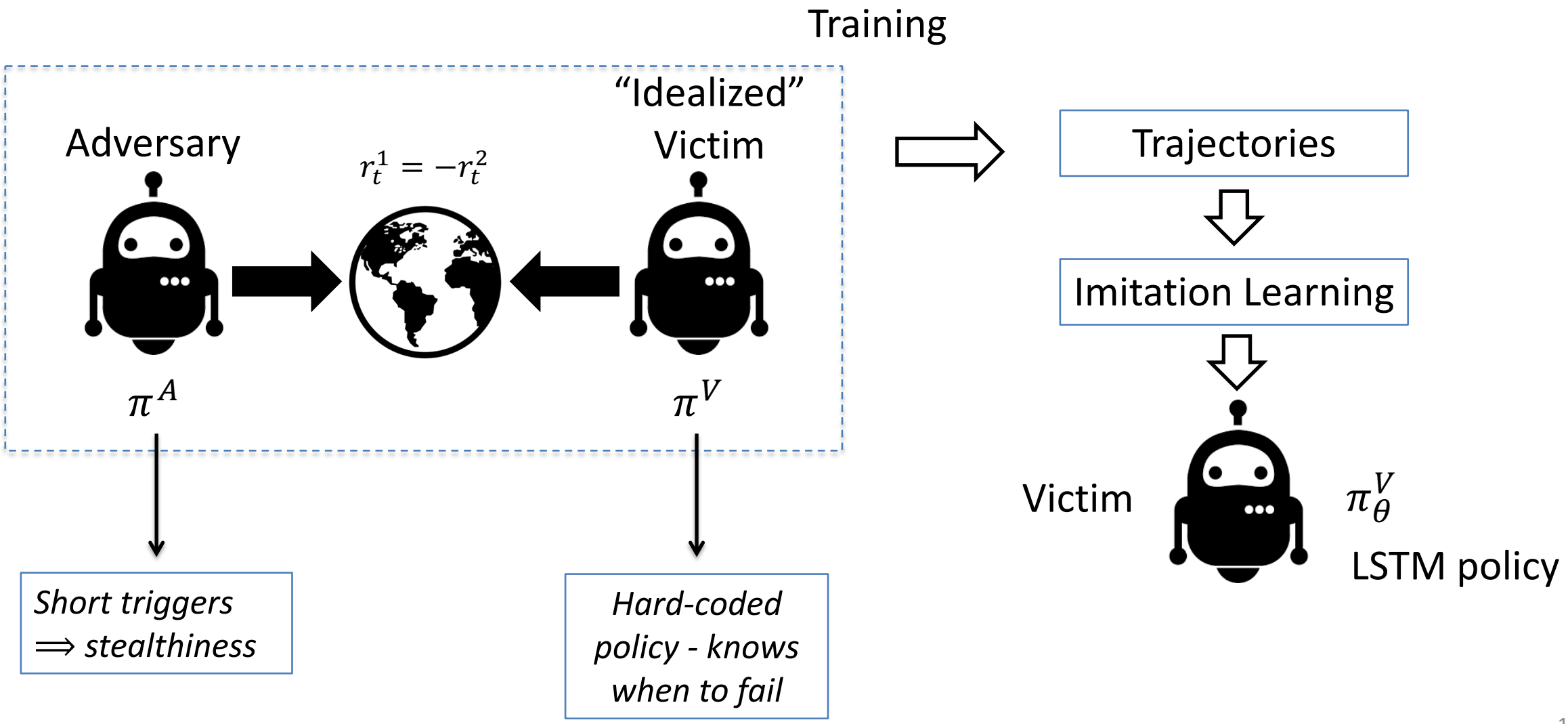
Backdoor Attacks



Inject a backdoor that can be triggered by π^A

[Wang et al., 2021]

Backdoor Attacks



Backdoor Attacks

Victim's winning rate reduces by 17%-37% when the backdoor is triggered

Environment	Triggered			Not Triggered			Benign Baseline		
	Failing	Tie	Winning	Failing	Tie	Winning	Failing	Tie	Winning
Run To Goal (Ants)	73.8%	2.4%	23.8%	45.0%	5.1%	49.9%	46.0%	3.0%	51.0%
Run To Goal (Humans)	20.8%	69.3%	9.9%	52.2%	0.7%	47.1%	51.2%	1.4%	47.4%
You Shall Not Pass (Humans)	83.0%	0.0%	17.0%	50.1%	0.0%	49.9%	50.5%	0.0%	49.5%
Sumo (Humans)	34.4%	54.7%	10.9%	29.7%	42.2%	28.1%	30.1%	34.4%	35.5%

Table 2: The failing/tie/winning rate of the victim agent when the backdoor is triggered (or not). Benign baselines are measured on two normal agents.

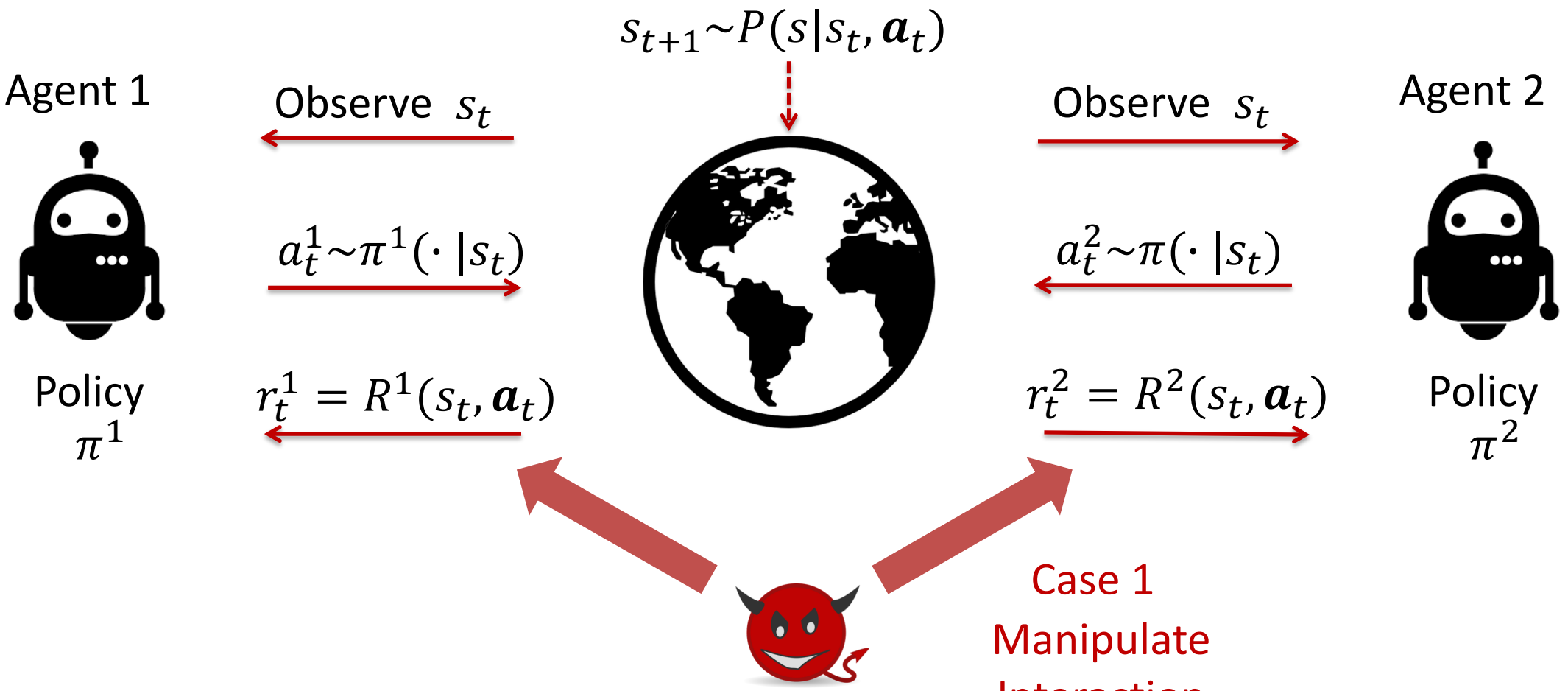
Fine tuning defense not fully successful

Environment/Failing Rate	BACKDOORL	Fine-tuned
Run to Goal (Ants)	23.8%	**39.0%**
Run to Goal (Humans)	**9.9%**	5.0%
You Shall Not Pass	17.0%	**23.8%**
Sumo	10.9%	**22.5%**

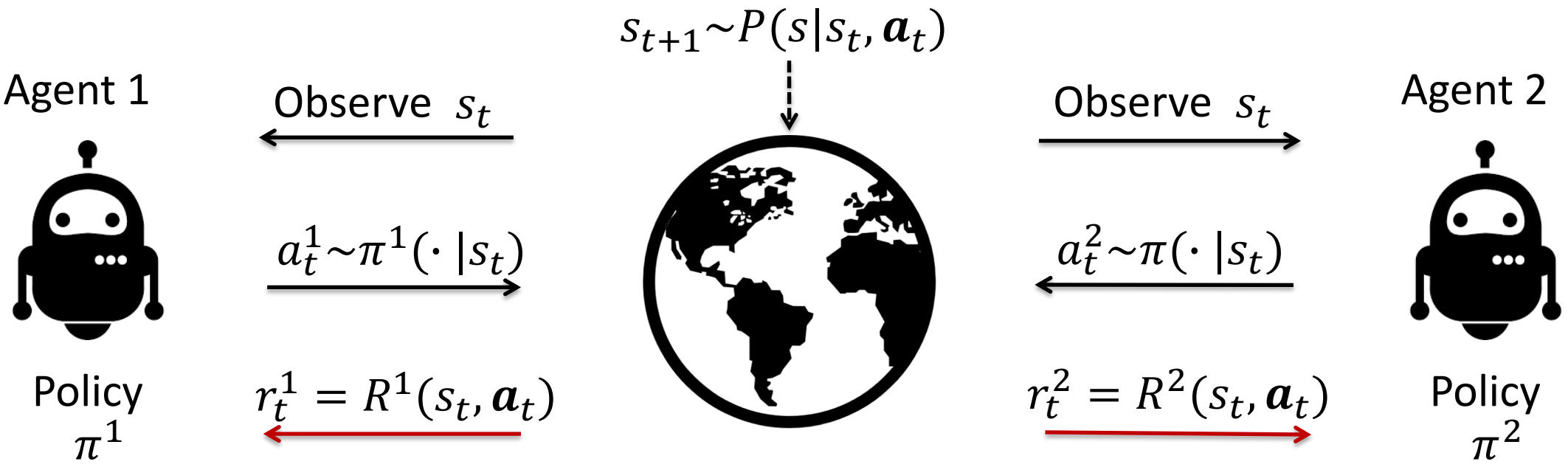
Table 6: Winning rate before/after fine-tuning when facing the trigger. Bolded numbers are the higher winning rates.

[Wang et al., 2021]

Adversarial Multi-Agent Setting



Reward Poisoning Attacks



Force a joint target policy $\pi_\dagger$:
Minimally change R^i s.t. $\pi_\dagger$ is an ε-strict Markov perfect **dominant strategy equilibrium**

[Wu et al., 2022]

Reward Poisoning Attacks

- Offline finite-horizon setting:
  - Attacker: Modifies the rewards in a given dataset
  - Agents/Learners: Estimate the parameters of Markov game from the poisoned data

- Q-confidence bound backward induction – minimize cost $C(r, r_0)$ while satisfying

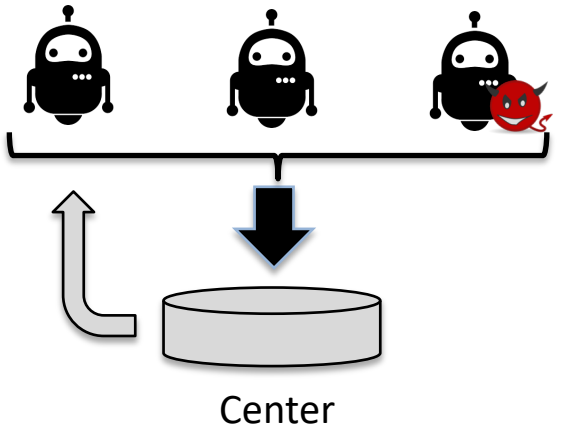
$$\underline{Q}_{i,h}(s, (\pi_\dagger^{i,h}(s), a^{-i})) \geq \overline{Q}_{i,h}(s, (a^i, a^{-i})) + \varepsilon \quad \longleftarrow \quad \textit{Equilibrium condition}$$

  - Exponential in #Agents → Greedy backward-induction

[Wu et al., 2022]

Summary

- MARL as a framework for physical adversarial attacks
- Attacks and defense in MARL largely unexplored
- Byzantine Attacks in Distributed RL
  - Fault-tolerant Federated RL [Fan et al., 2021]
  - Byzantine-Robust Distributed RL [Chen et al., 2022]



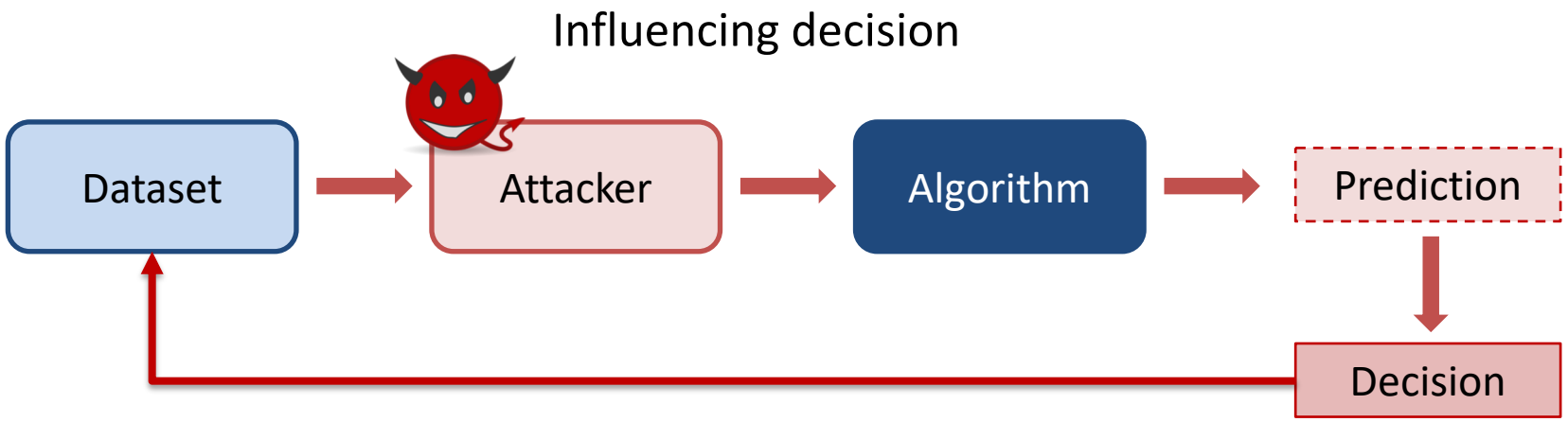
Outline

- Preliminaries
- Test-time Attacks and Defenses in RL
- Training-time Attacks in RL
- Training-time Defenses in RL
- Adversarial Attacks in Multi-agent RL
- Concluding Remarks

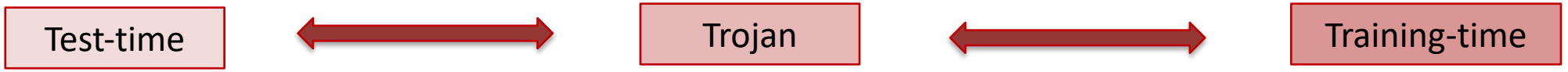
Outline

- Preliminaries
- Test-time Attacks and Defenses in RL
- Training-time Attacks in RL
- Training-time Defenses in RL
- Adversarial Attacks in Multi-agent RL
- Concluding Remarks

Concluding Remarks

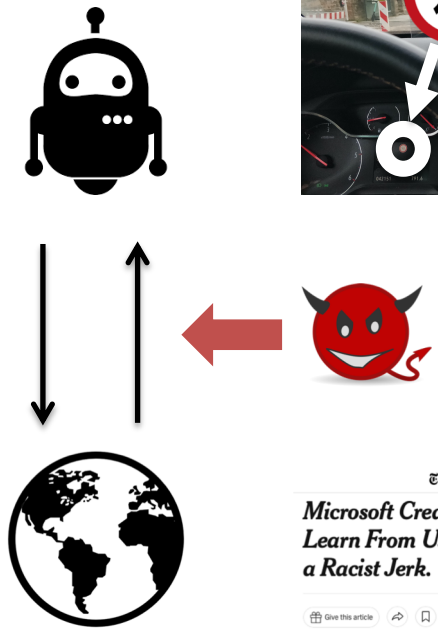


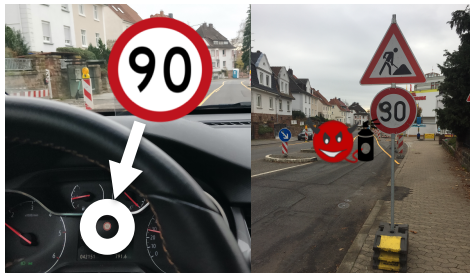
Attack modalities



Concluding Remarks

Attacks/Defenses





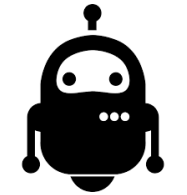
The New York Times

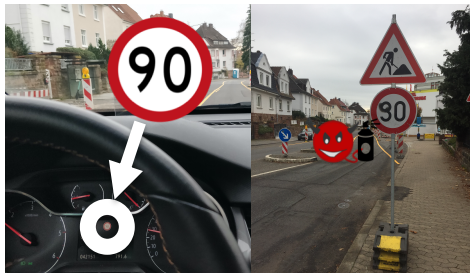
Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Give this article

Concluding Remarks

Attacks/Defenses

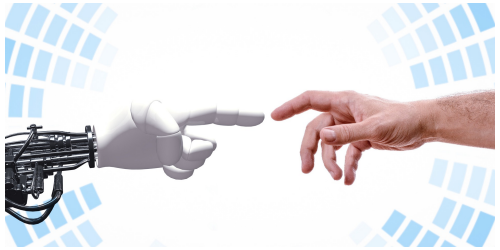


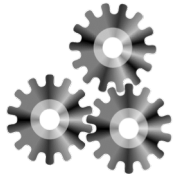


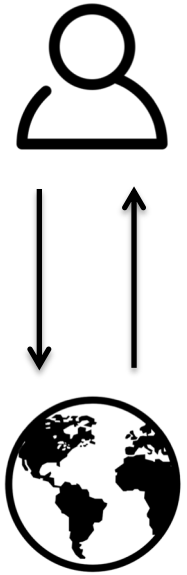
The New York Times

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Steering/Teaching







References

- Gleave et al., Adversarial Policies: Attacking Deep Reinforcement Learning, 2020.
- Guo et al., Adversarial Policy Learning in Two-player Competitive Games, 2021.
- Wang et al., BACKDOORL: Backdoor Attack against Competitive Reinforcement Learning, 2021.
- Fan et al., Fault-Tolerant Federated Reinforcement Learning with Theoretical Guarantee, 2021.
- Wu et al., Reward Poisoning Attacks on Offline Multi-Agent Reinforcement Learning, 2022.
- Chen et al., Byzantine-Robust Online and Offline Distributed Reinforcement Learning, 2022.