CAROL: **C**ertifi**a**bly **Ro**bust Reinforcement **L**earning through Model-Based Abstract Interpretation

Chenxi Yang[1], Greg Anderson[2], Swarat Chaudhuri[1]
[1]The University of Texas at Austin, [2]Reed College

Background: RL in Safety-Critical Tasks

- Reinforcement learning (RL) is an established approach for various tasks, including safety-critical ones.

HOSPITAL HEALTHCARE EUROPE

News In depth Clinical Insight and Analysis Clinical Excellence Sponsored

Reinforcement learning AI model improves accuracy of skin cancer diagnoses

Article | Published: 22 March 2023

Dense reinforcement learning for safety validation of autonomous vehicles

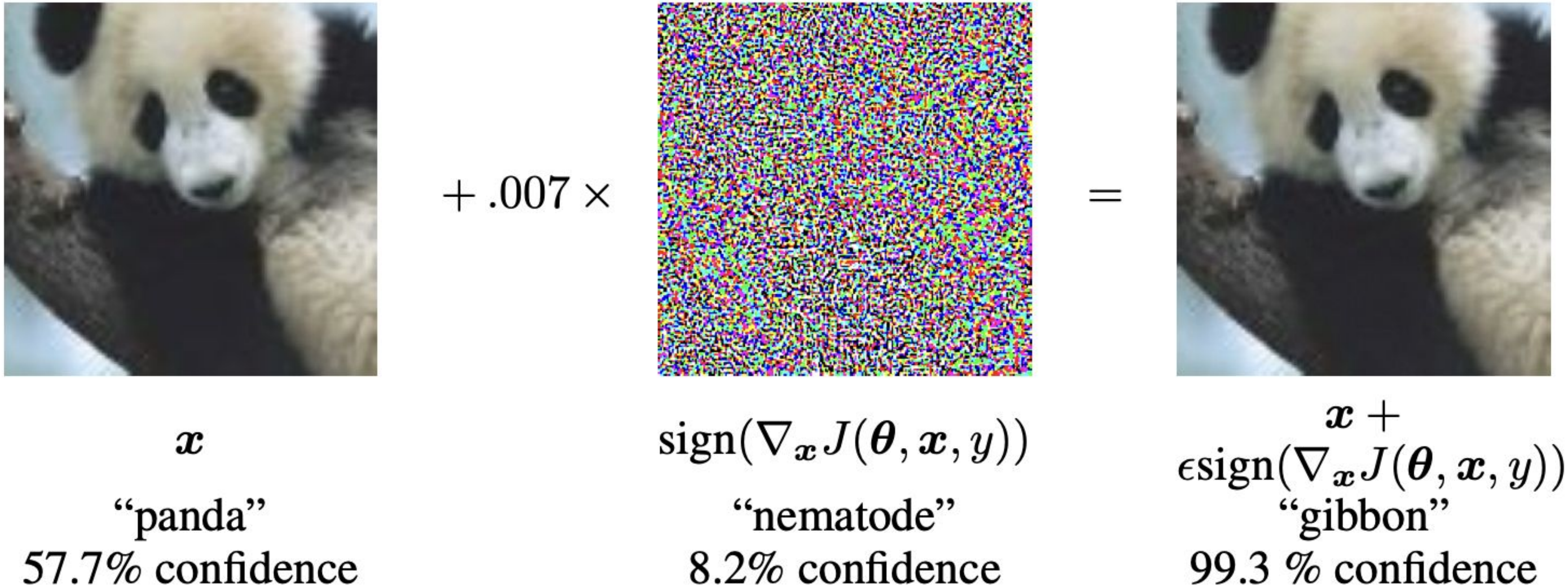
Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen & Henry X. Liu

Nature **615**, 620–627 (2023) | Cite this article

- State-of-the-art RL methods use neural networks as policy representations.

Background: RL with Neural Network Policies is Vulnerable

Neural networks are vulnerable.

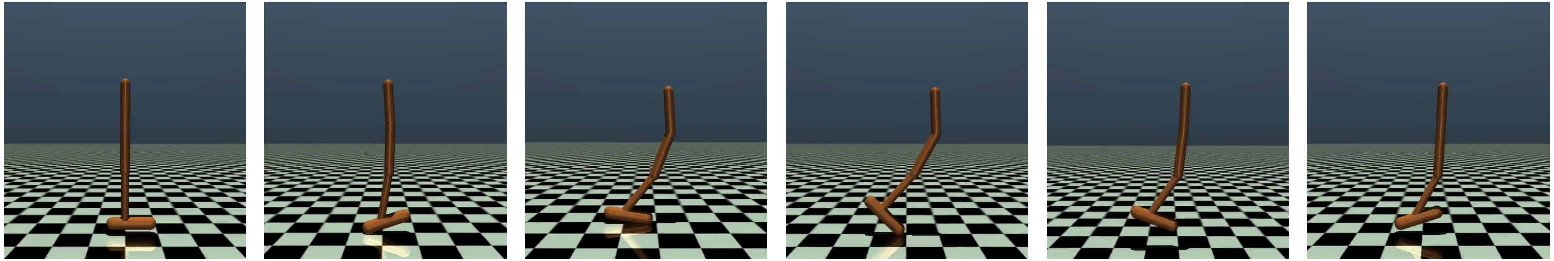


[1] Goodfellow et, al. Explaining and Harnessing Adversarial Examples. ICLR 2015.

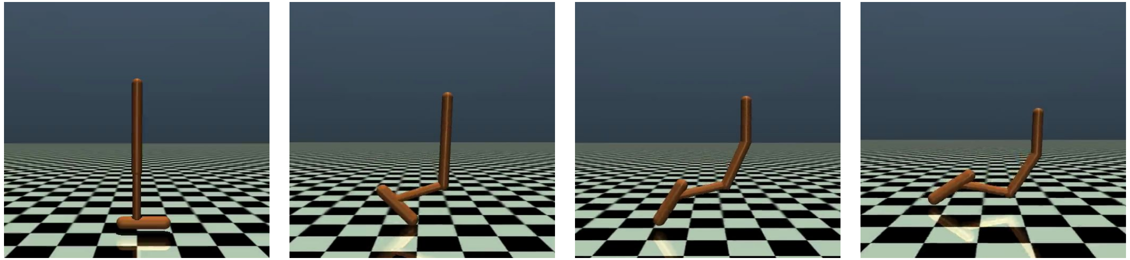
Background: RL with Neural Network Policies is Vulnerable

Problems are more severe in RL as mistakes can cascade.

A hopper moves forward



Under attacks



Background: Certified Defenses

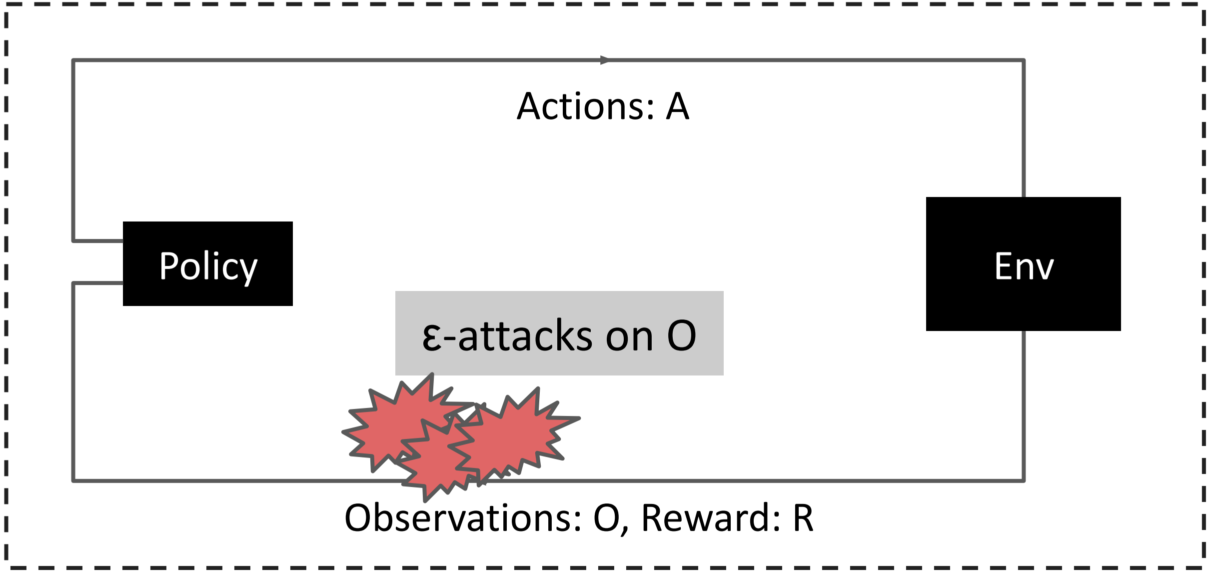
Certified Neural Networks in Supervised Learning
DiffAI (Mirman et al. 18),
k-ReLU (Singh et al. 19),
RNN Verification (Ryou et al. 21)

Defenses are still heuristic in RL
SA (Zhang et al. 20),
PA-AD (Sun et al. 22),
RADIAL (Oikarinen et al. 21)

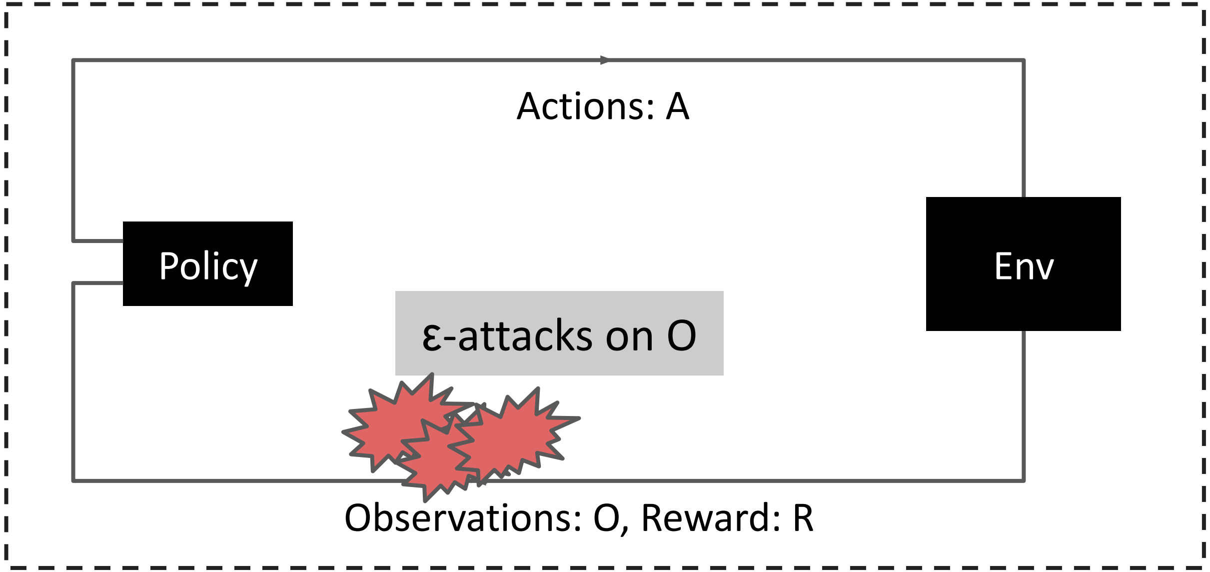
Heuristic defenses are defeated by **counter** attacks.

Can we train a **certifiable** RL policy against **arbitrary** attacks?

Goal: Train Certifiable Robust Reinforcement Learning Policies



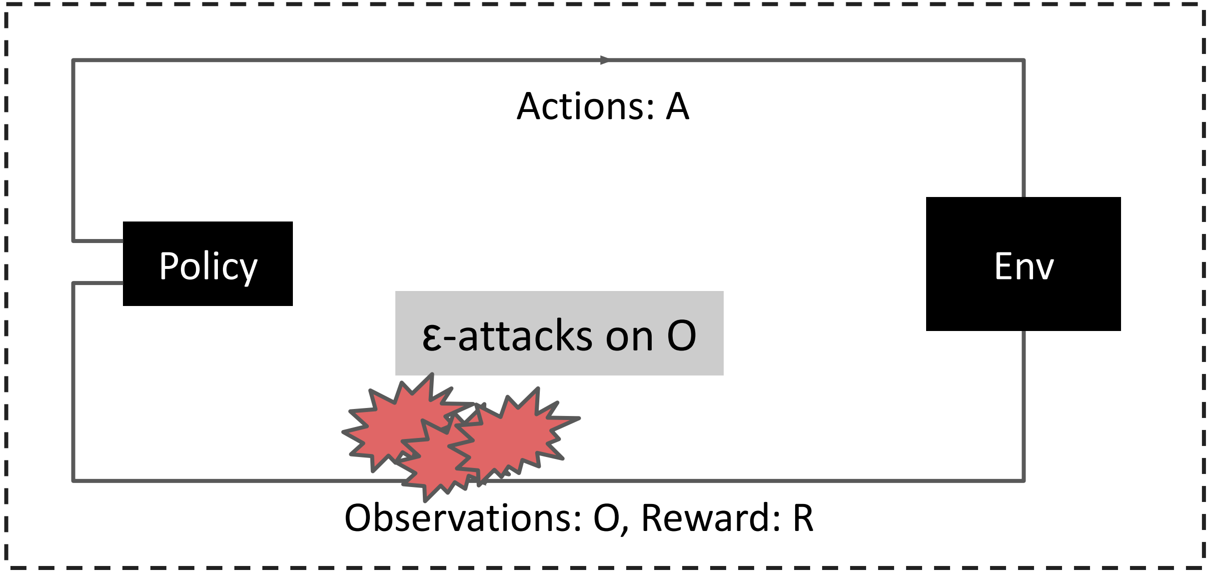
Goal: Train Certifiable Robust Reinforcement Learning Policies



Challenges

- How to represent and quantify worst-case attacks?

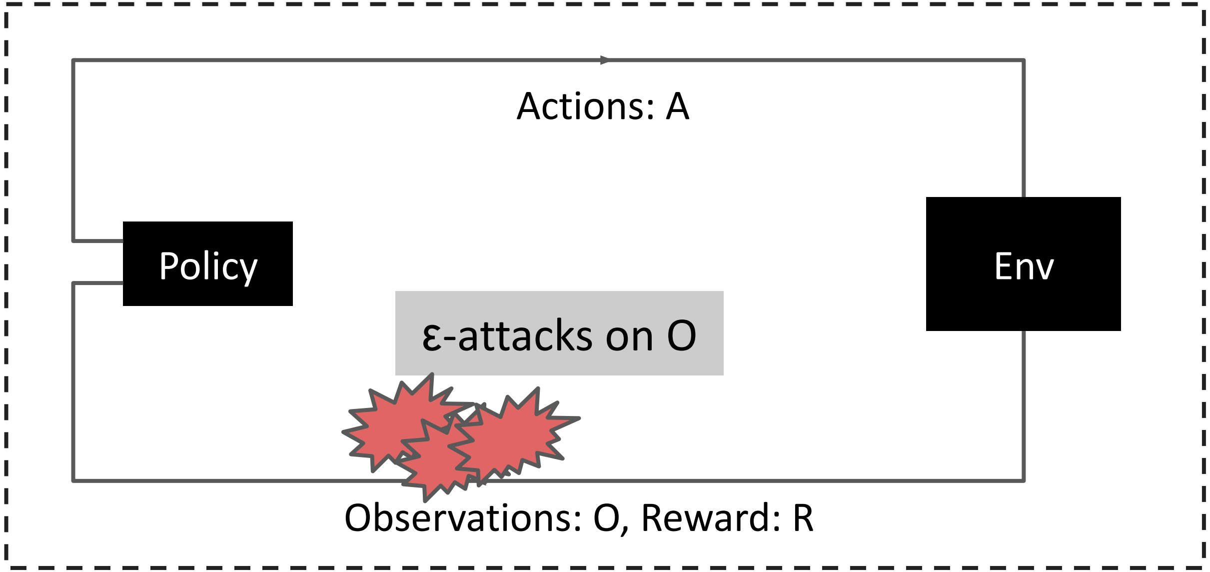
Goal: Train Certifiable Robust Reinforcement Learning Policies



Challenges

- How to represent and quantify worst-case attacks?
- How to reason over the black-box environment?

Goal: Train Certifiable Robust Reinforcement Learning Policies



Challenges

- How to represent and quantify worst-case attacks?

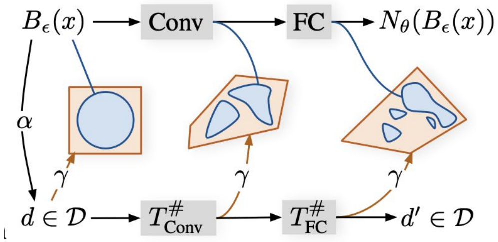
We use abstract interpretation, covering all the attacks.

- How to reason over the black-box environment?

Goal: Train Certifiable Robust Reinforcement Learning Policies

Abstract Interpretation[1]: A well-established method to effectively compute bounds over functions.

It can be used to certify neural networks[2].



Challenges

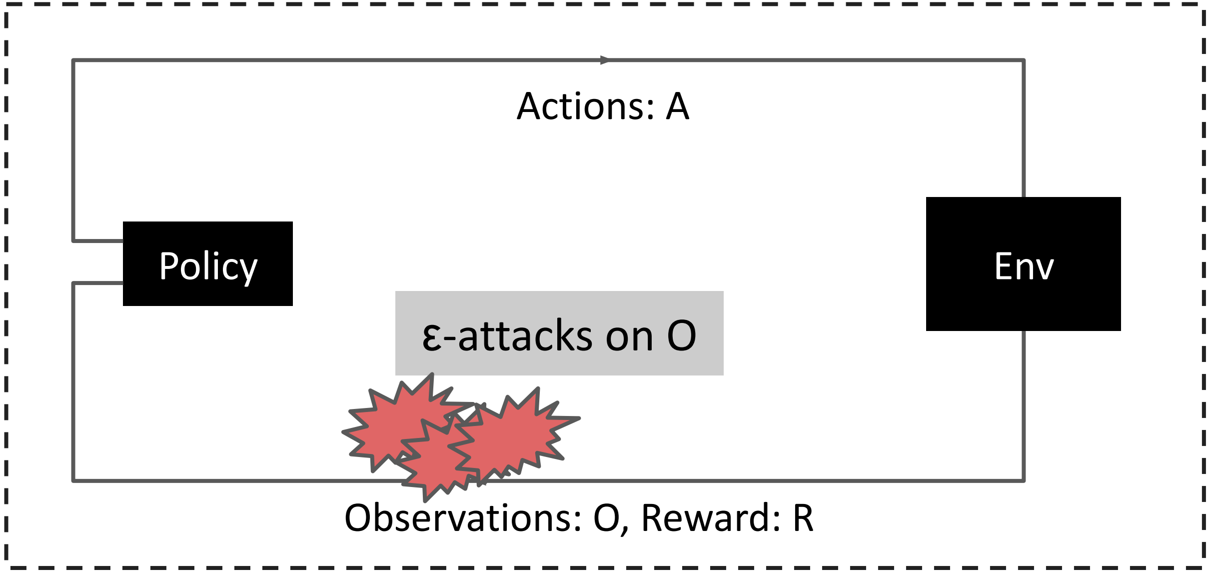
- How to represent and quantify worst-case attacks?

We use abstract interpretation, covering all the attacks.

- How to reason over the black-box environment?

[1] Cousot et, al. Abstract Interpretation. POPL 1977.
[2] Mirman et, al. Differentiable Abstract Interpretation for Provably Robust Neural Networks. ICML 2018.

Goal: Train Certifiable Robust Reinforcement Learning Policies



Challenges

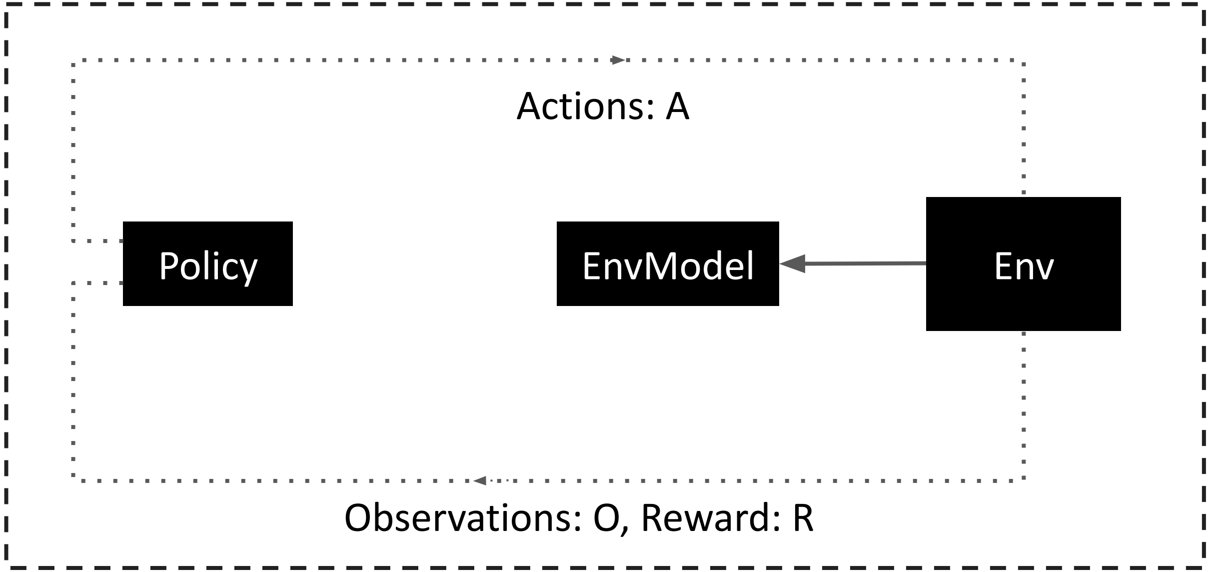
- How to represent and quantify worst-case attacks?

We use abstract interpretation, covering all the attacks.

- How to reason over the black-box environment?

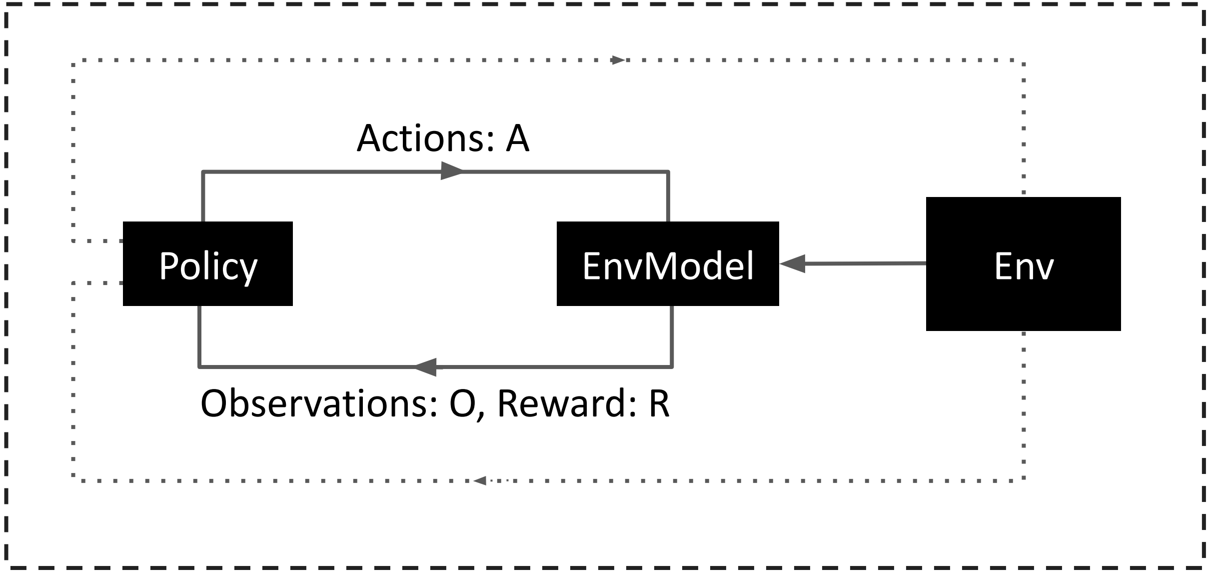
Learn a white-box transition representation of the environment with the policy.

Carol: Certifiably Robust Reinforcement Learning



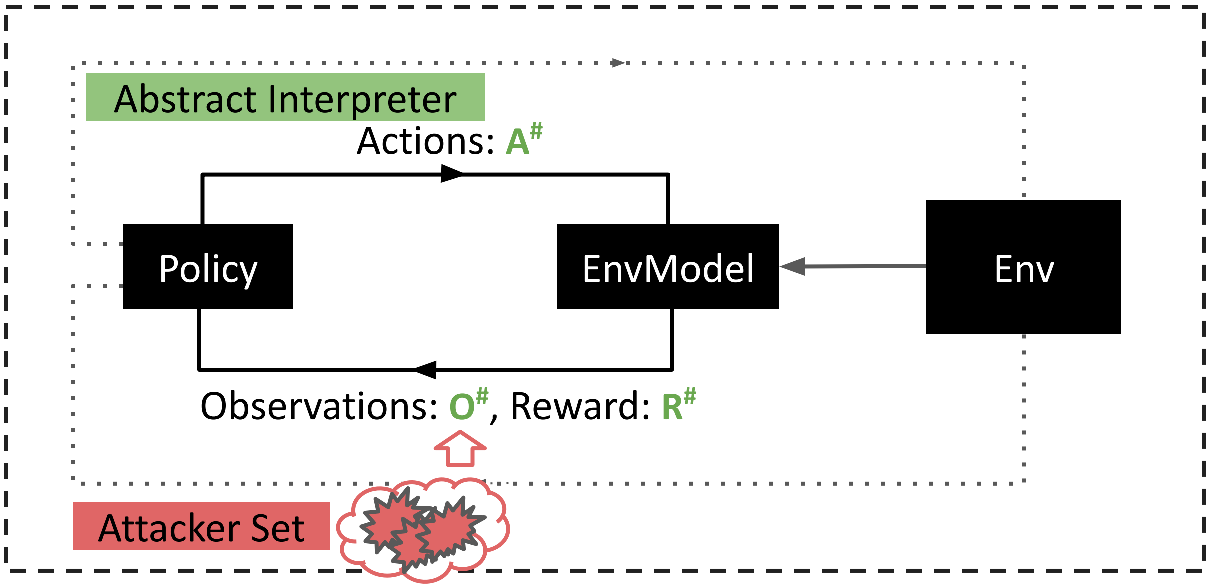
Step 1: Train a NN represented **model** (verifiable) for the black-box environment during normal training.

Carol: Certifiably Robust Reinforcement Learning



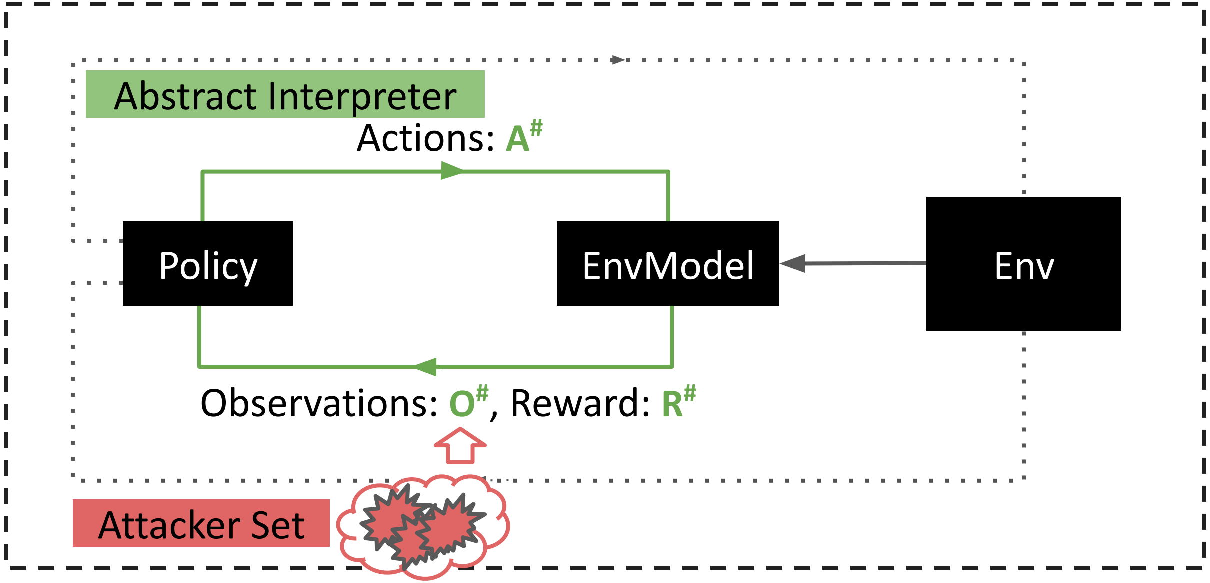
Step 1: Train a NN represented **model** (verifiable) for the black-box environment during normal training.
Step 2: Train the **policy** over the NN model of the real environment.

Carol: Certifiably Robust Reinforcement Learning



Step 1: Train a NN represented **model** (verifiable) for the black-box environment during normal training.
Step 2: Train the **policy** over the NN model of the real environment.
Step 3: A **symbolic** RL algorithm: **RL#** : with the learnt symbolic reward **R#**.

Carol: Certifiably Robust Reinforcement Learning



Step 1: Train a NN represented **model** (verifiable) for the black-box environment during normal training.
Step 2: Train the **policy** over the NN model of the real environment.
Step 3: A **symbolic** RL algorithm: **RL#** : with the learnt symbolic reward **R#**.
Step 4: In each iteration: we use the accumulative reward lower bound to guide the training: $\hat{R}^{\#}$ = LowerBound[**RL#**(**A#**, **O#**, **R#**)]

Theoretical Bound of Reward

With probability 1 - δ, the reward (R) under the worst attack is bounded by,

$$R \geq \hat{R}^{\#} - \frac{1}{\sqrt{\delta}}\sqrt{\frac{Var[R^{\#}]}{N}} - \left(1 - (1 - \delta_E)^T\right) C.$$

Theoretical Bound of Reward

With probability 1 - δ , the reward (R) under the worst attack is bounded by,

$$R \geq \hat{R}^{\#} - \frac{1}{\sqrt{\delta}}\sqrt{\frac{Var[R^{\#}]}{N}} - \left(1 - (1 - \delta_E)^T\right) C.$$

1. The bound grows as the δ shrinks.
   ⇨ We pay the price of a looser bound as we consider higher confidence levels.
2. The bound depends on $Var[R^{\#}]$ and N in an intuitive way.
   ⇨ Higher variance makes it harder to measure the true reward, more samples make the bound tighter.
3. As δ_E increases, the last term grows.
   ⇨ A less accurate environment model leads to a looser bound.
4. The bound grows with T.
   ⇨ Over longer time horizons, our reward measurement gets less accurate.

Results: Certifiable Accumulative Reward Bound

Reward
Bound under
Worst-case
Attack

Time Horizon (T)

Results: Certifiable Accumulative Reward Bound

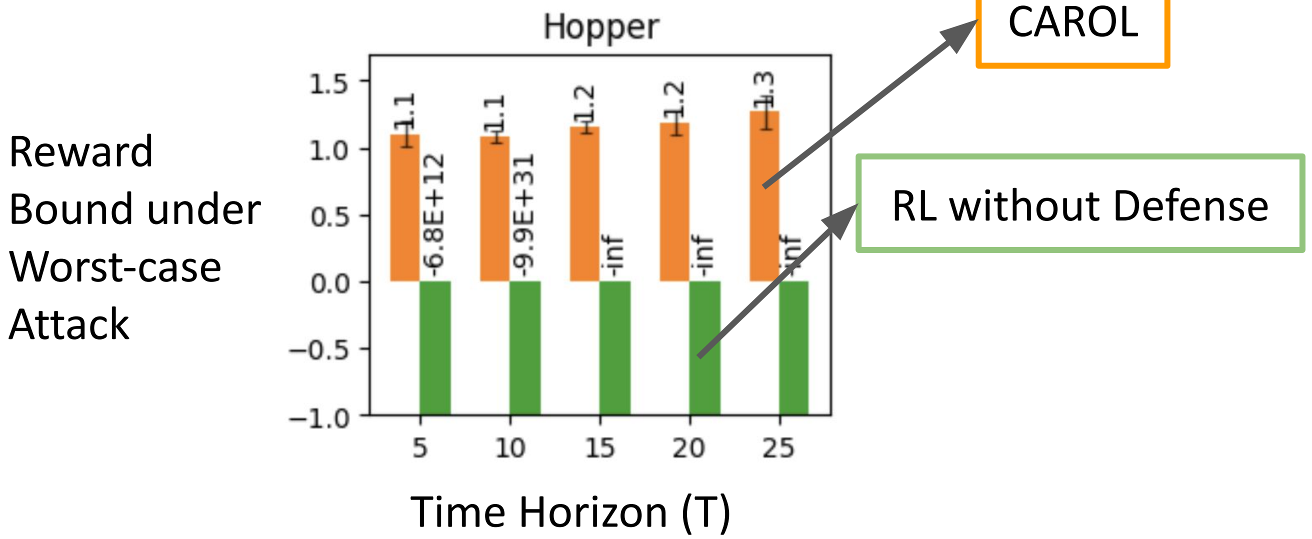
CAROL

Reward
Bound under
Worst-case
Attack

RL without Defense

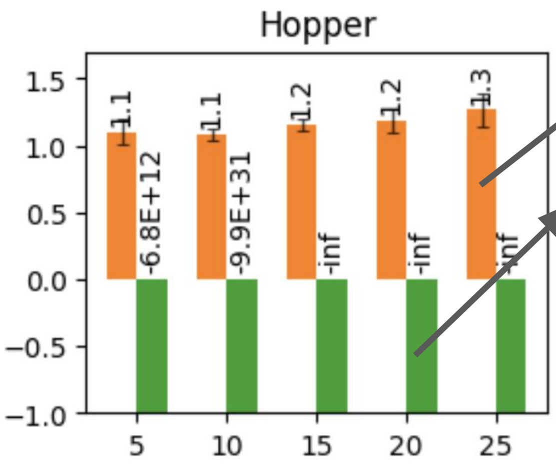
Time Horizon (T)

Results: Certifiable Accumulative Reward Bound



Results: Certifiable Accumulative Reward Bound

Reward Bound under Worst-case Attack



Time Horizon (T)

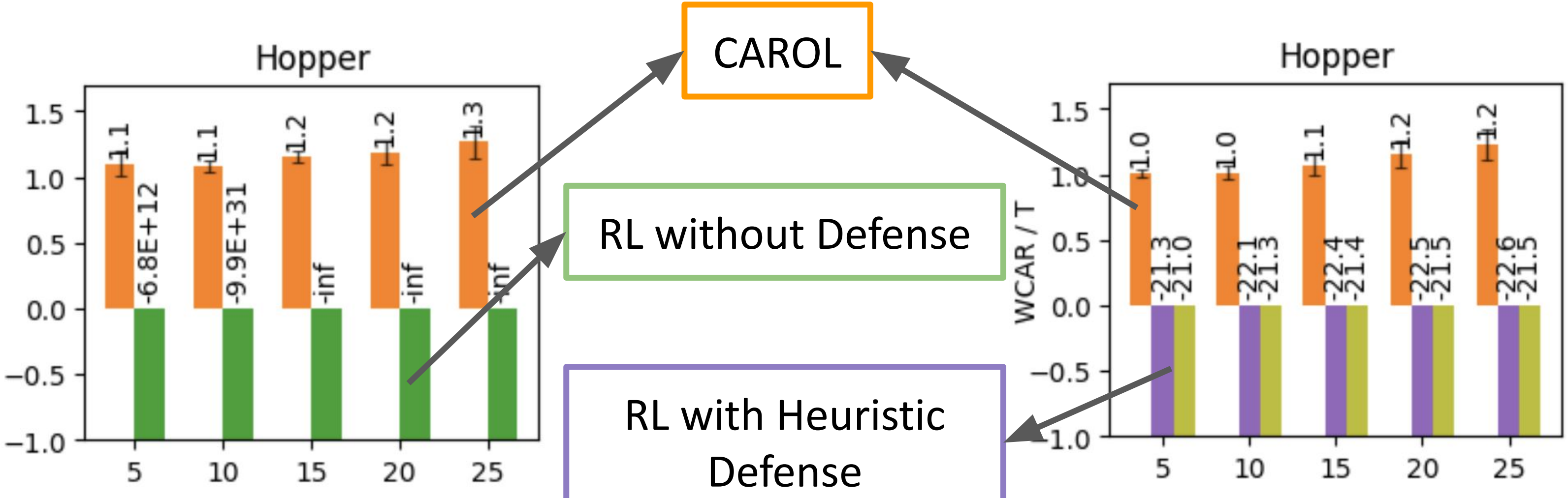
CAROL

RL without Defense

RL with Heuristic Defense

Results: Certifiable Accumulative Reward Bound

Reward Bound under Worst-case Attack



Time Horizon (T)

Time Horizon (T)

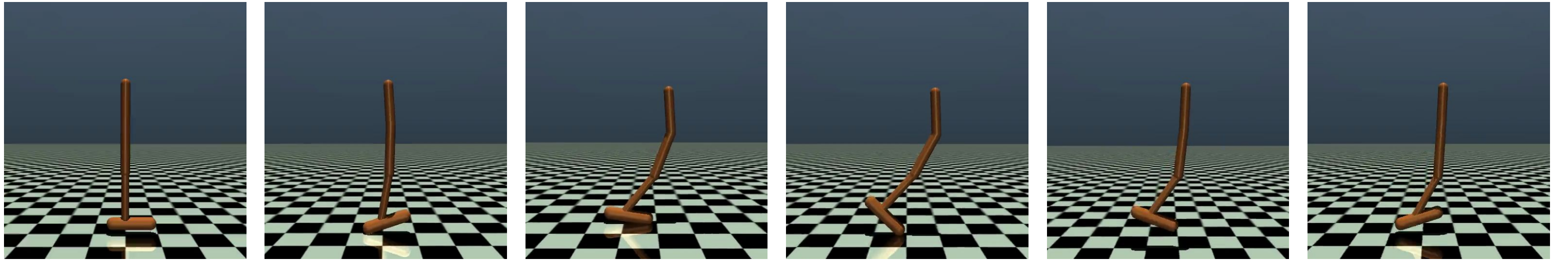
Summary: CAROL

Thank you!

CAROL: **Certifiable** Robust Reinforcement Learning with **Long-Horizon** Reward Bound

Key Idea: Abstract Interpretation for **Verification** in the Learning Loop

White-Box Environment Representation Learning



Future: More **Accurate** and **Scalable** Certified RL

Code: https://github.com/chenxi-yang/carol