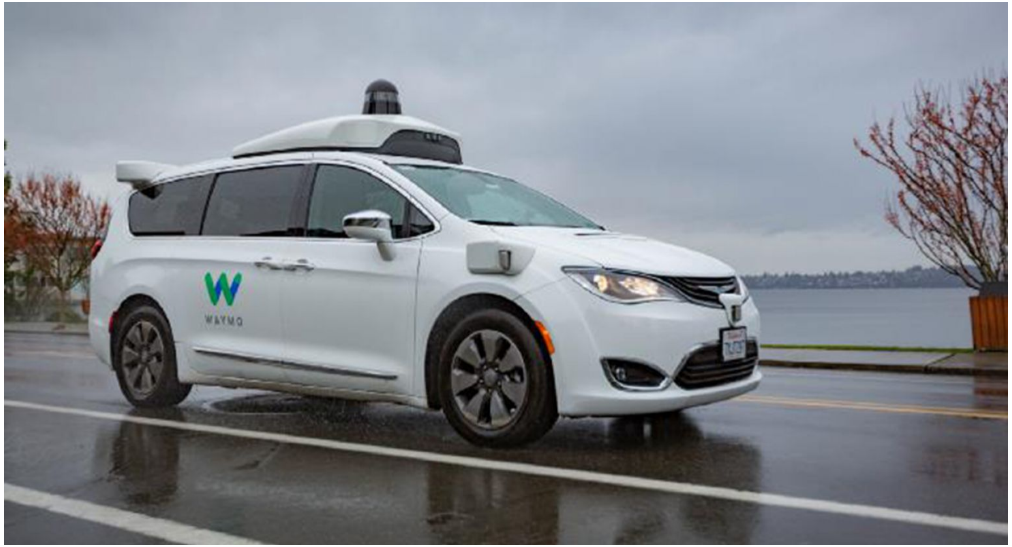
Safe Reinforcement Learning

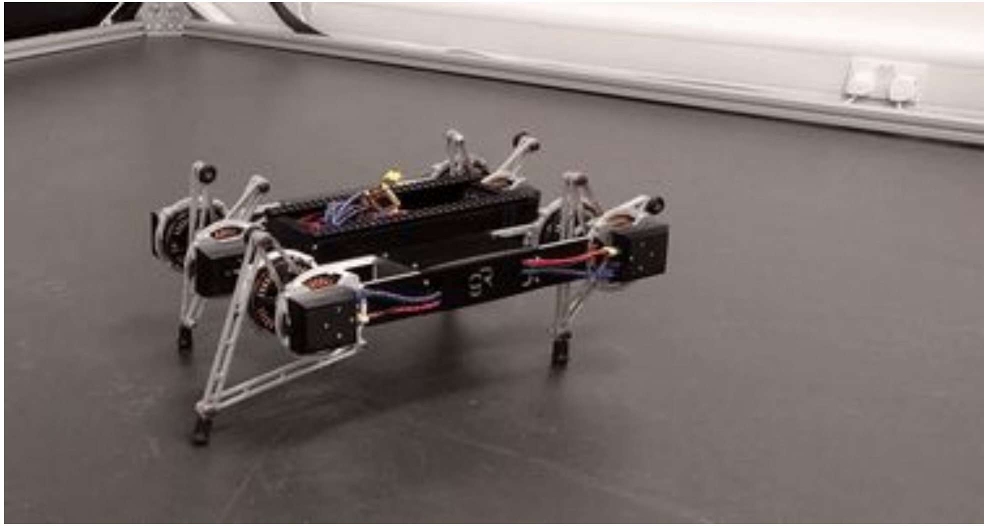
Jie Tan

CS 8803 Deep Reinforcement Learning for Intelligent Control

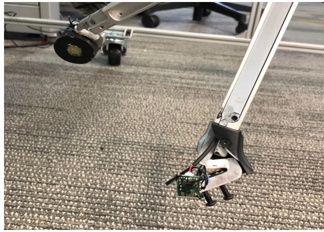
04/06/2022

Why Safety






Unitree
Unitree
宇树科技



Goals

- Definition of safety
- Two ways to improve safety during learning
  - Constrained Markov Decision Process
  - Safe learning via shielding

Brainstorming: What are Unsafe Behaviors?

- Autonomous cars: collision
- Legged robots: falling
- Robot manipulators: destroying the object in manipulation
- Investing: losing x% of values in the portfolio
- Data center cooling: overheating the servers
- Power grid: power supply shortage
- …

Safety as Constraints

- Autonomous cars: collision **$d < 0$**
- Legged robots: falling **$h < 0.5m$**
- Robot manipulators: destroying the object in manipulation **$f_{contact} > 7N$**
- Investing: losing x% of values in the portfolio **$\$ < 1M$**
- Data center cooling: overheating the servers **$t > 104^{o}F$**
- Power grid: power supply shortage **$E_{generator} - E_{consumer} < 0$**
- …

Three Levels of Constraints

Soft Constraints
Safety Level I

Probabilistic Constraints
Safety Level II

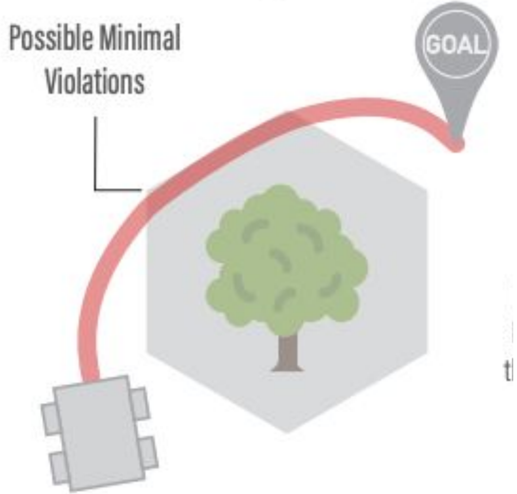
Hard Constraints
Safety Level III

Three Levels of Constraints

Soft Constraints
Safety Level I

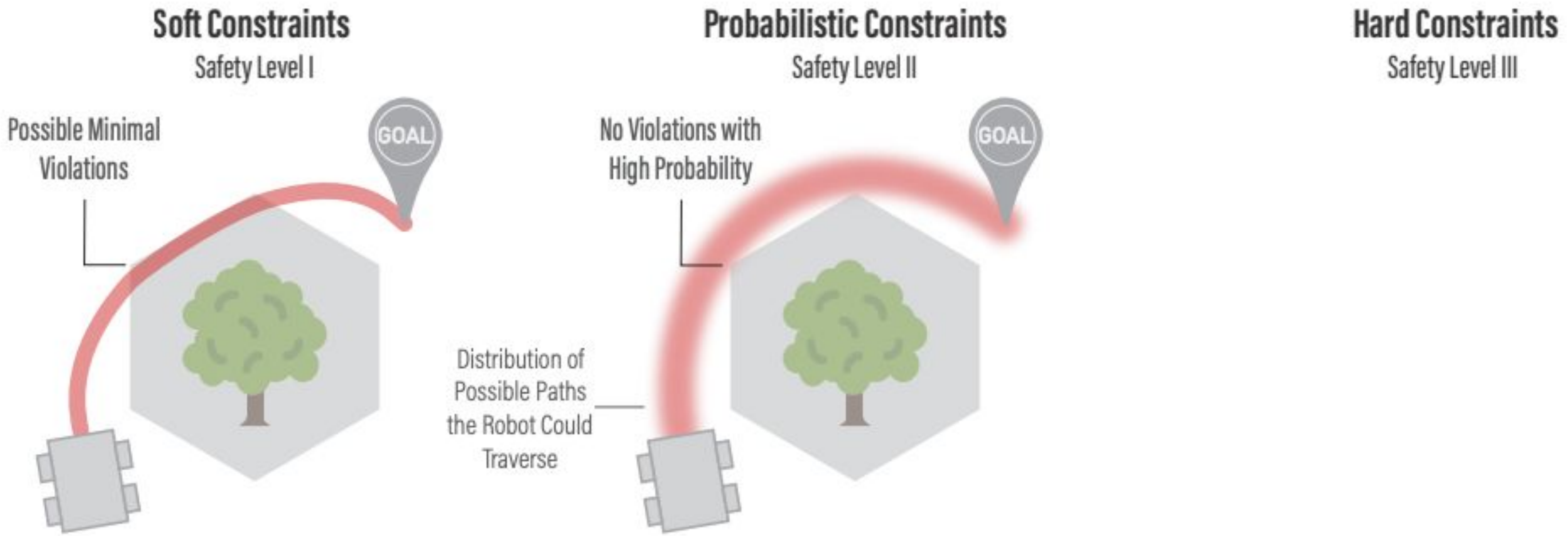
Probabilistic Constraints
Safety Level II

Hard Constraints
Safety Level III



$$\mathbb{E}\left[\sum_{t=0}^{T} f_s(s_t, a_t)\right] \geq 0$$

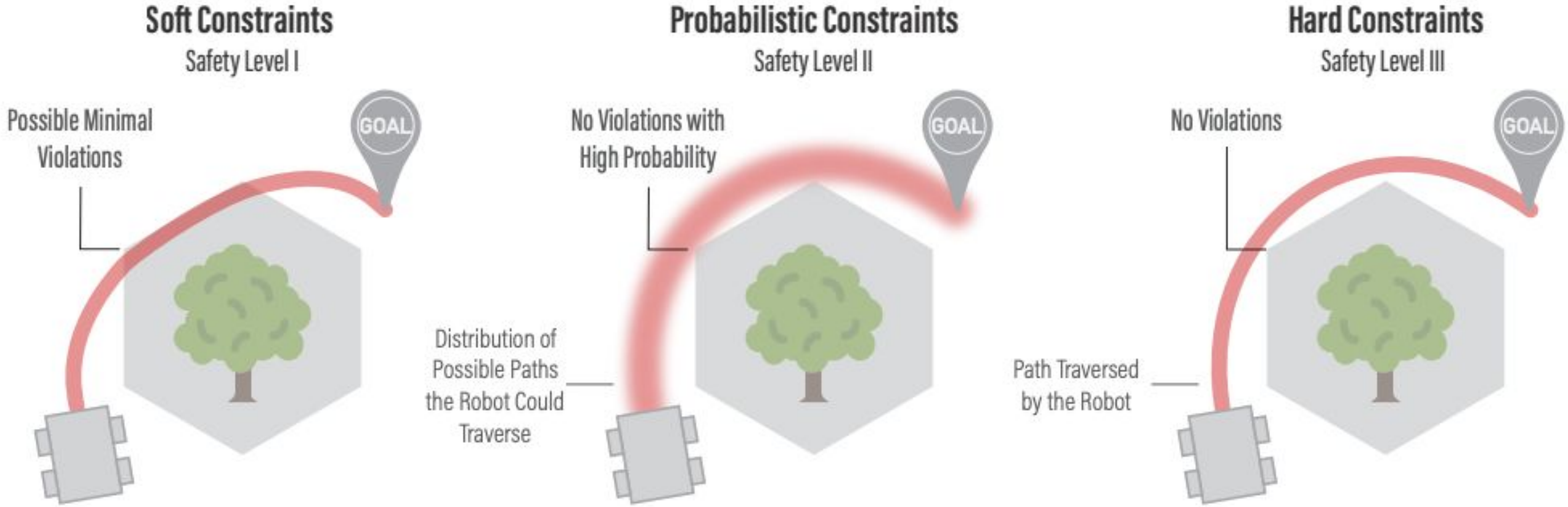
Three Levels of Constraints



$$\mathbb{E}\left[\sum_{t=0}^{T} f_s(s_t, a_t)\right] \geq 0$$

$$Pr(f_s(s_t, a_t) \geq 0) > 1 - \epsilon$$

Three Levels of Constraints

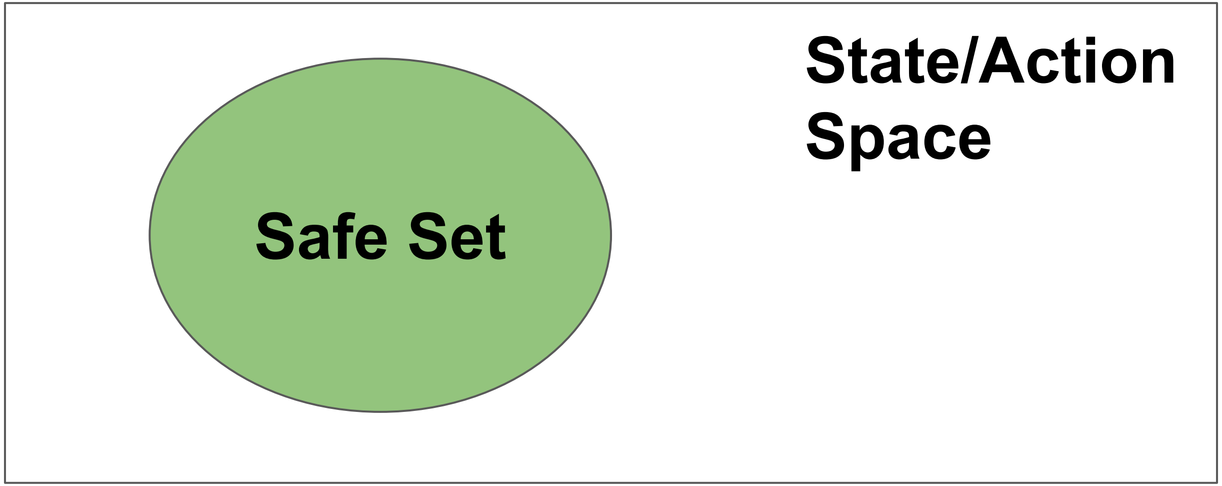


$$\mathbb{E}\left[\sum_{t=0}^{T} f_s(s_t, a_t)\right] \geq 0$$

$$Pr(f_s(s_t, a_t) \geq 0) > 1 - \epsilon$$

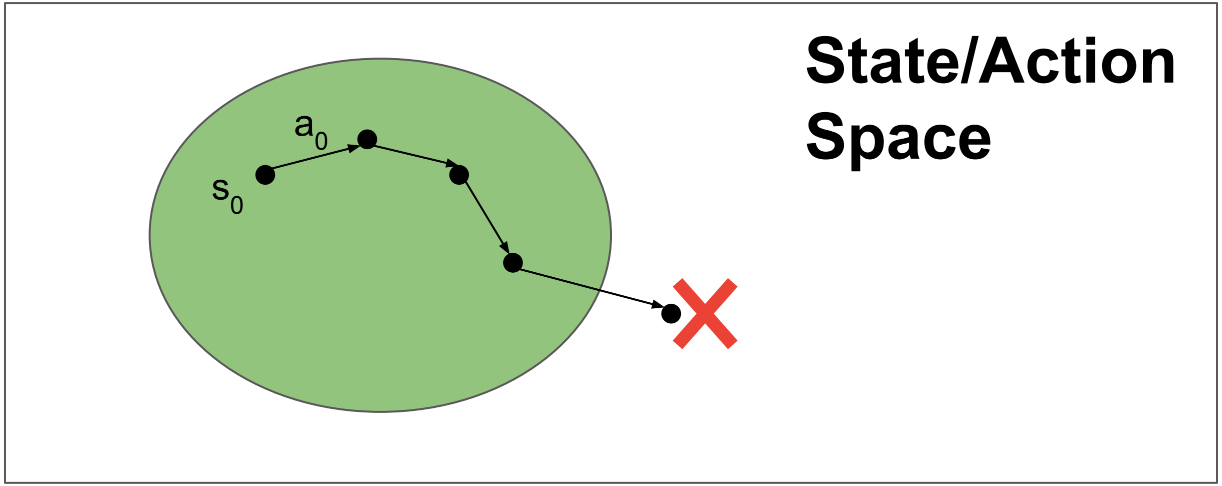
$$f_s(\mathbf{s}_t, \mathbf{a}_t) \ \geq 0$$

Constrained Markov Decision Process (CMDP)



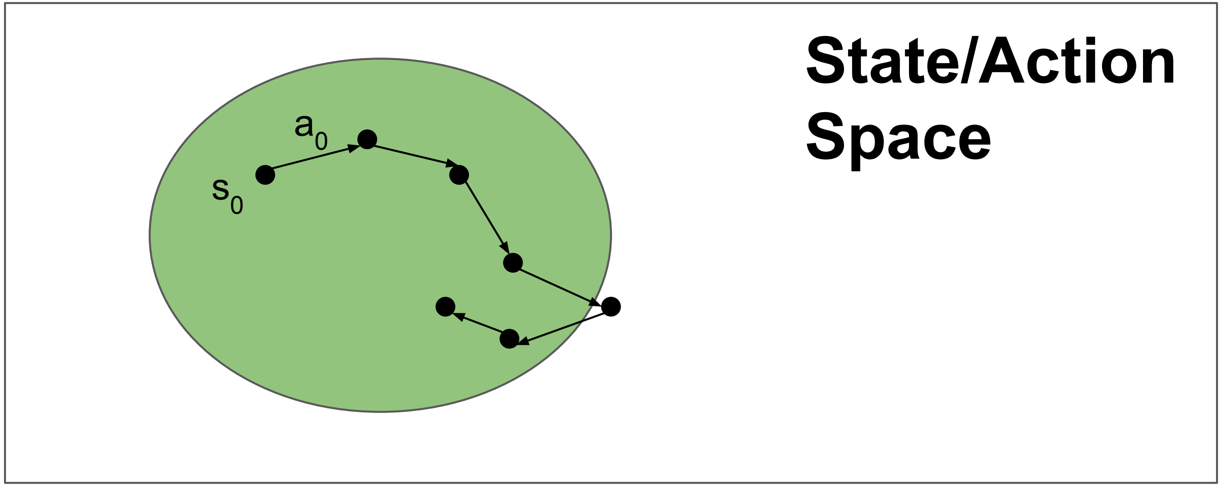
$$f_s(\mathbf{s}_t, \mathbf{a}_t) \geq 0$$

CMDP Problem Definition



$$\max_{\pi \in \Pi} \mathbb{E}_{\tau \sim \rho_\pi} \left[\sum_{t=0}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

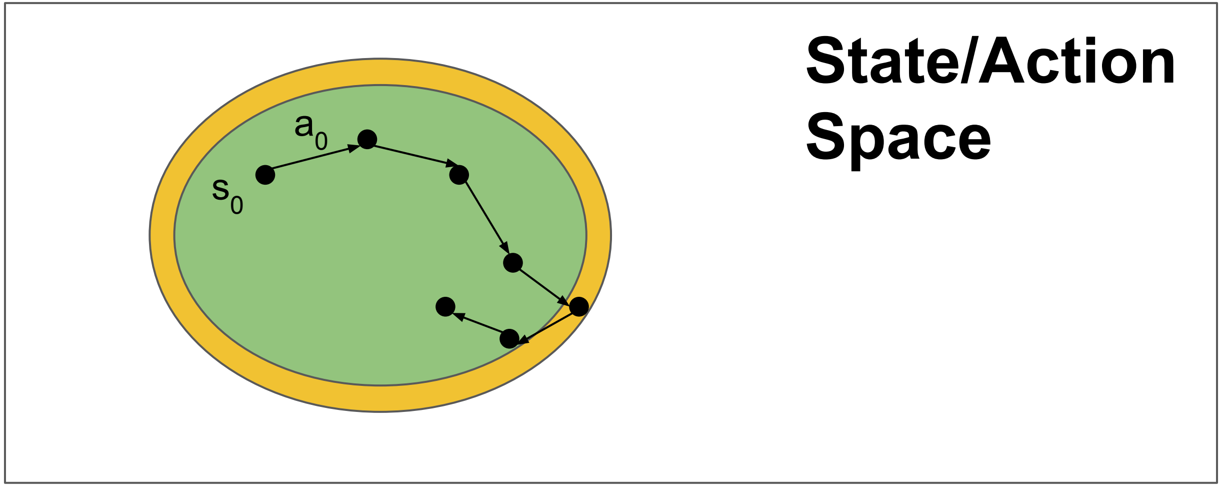
CMDP Problem Definition



$$\max_{\pi \in \Pi} \mathbb{E}_{\tau \sim \rho_\pi} \left[\sum_{t=0}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

$$\text{s.t. } \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[\sum f_s(\mathbf{s}_t, \mathbf{a}_t) \right] \geq 0$$

CMDP Problem Definition



$$\max_{\pi \in \Pi} \mathbb{E}_{\tau \sim \rho_\pi} \left[\sum_{t=0}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right]$$
$$\text{s.t. } \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[\sum f_s(\mathbf{s}_t, \mathbf{a}_t) \right] \geq 0$$

Solving CMDP: Lagrangian Method

$$\max_{\pi\in\Pi} \mathbb{E}_{\tau\sim\rho_\pi}\left[\sum_{t=0}^{T} r(\mathbf{s}_t,\mathbf{a}_t)\right]$$

$$\text{s.t.} \quad \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[\sum f_s(\mathbf{s}_t,\mathbf{a}_t)\right] \geq 0$$

$$= \max_{\pi} \min_{\lambda\geq 0} \mathbb{E}_{\pi\sim\rho_\pi}\left[\sum_{t=0}^{T} r(s_t,a_t) + \lambda f_s(s_t,a_t)\right]$$

Lagrangian

$$\mathcal{L}(\pi,\lambda) = \mathbb{E}_{\tau\sim\rho_\pi}\left[\sum_{t=0}^{T} r(\mathbf{s}_t,\mathbf{a}_t) + \lambda f_s(\mathbf{s}_t,\mathbf{a}_t)\right]$$

$$\max_{\pi} \min_{\lambda\geq 0} \mathcal{L}(\pi,\lambda)$$

$= 0 \quad \geq 0$

$$\max_{\pi} \boxed{\min_{\lambda\geq 0}} \mathbb{E}_{\pi\sim\rho_\pi}\left[\sum_{t=0}^{T} r(s_t,a_t) + \lambda f_s(s_t,a_t)\right]$$

$\to \infty \quad \leq 0$

$$\max_{\pi} \boxed{\min_{\lambda\geq 0}} \mathbb{E}_{\pi\sim\rho_\pi}\left[\sum_{t=0}^{T} r(s_t,a_t) + \lambda f_s(s_t,a_t)\right]$$

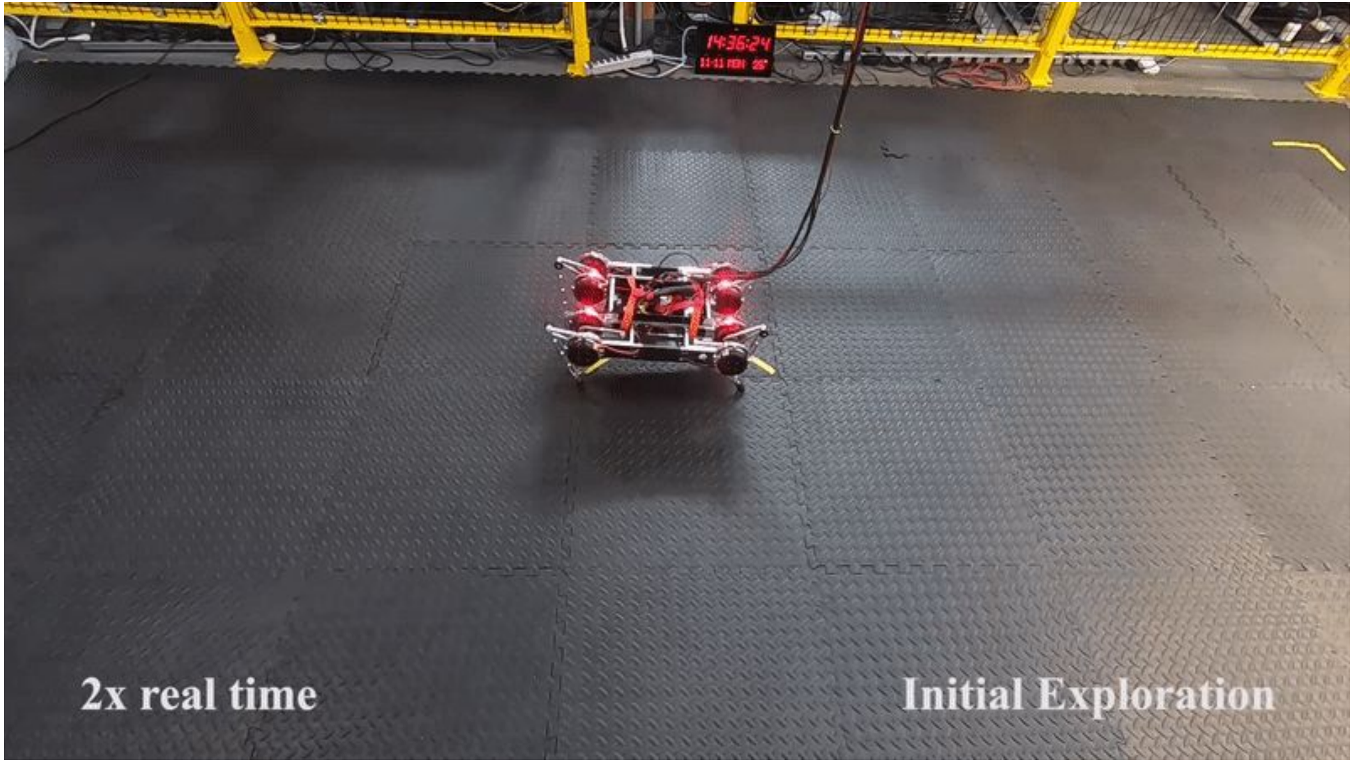
$$\max_{\pi} \cancel{\min_{\lambda \geq 0}} \, \mathbb{E}_{\pi \sim \rho_\pi} \left[\sum_{t=0}^{T} r(s_t, a_t) + \lambda f_s(s_t, a_t) \right]$$

$$\max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\pi \sim \rho_\pi} \left[\sum_{t=0}^{T} r(s_t, a_t) + \lambda \boxed{f_s(s_t, a_t)} \right]$$

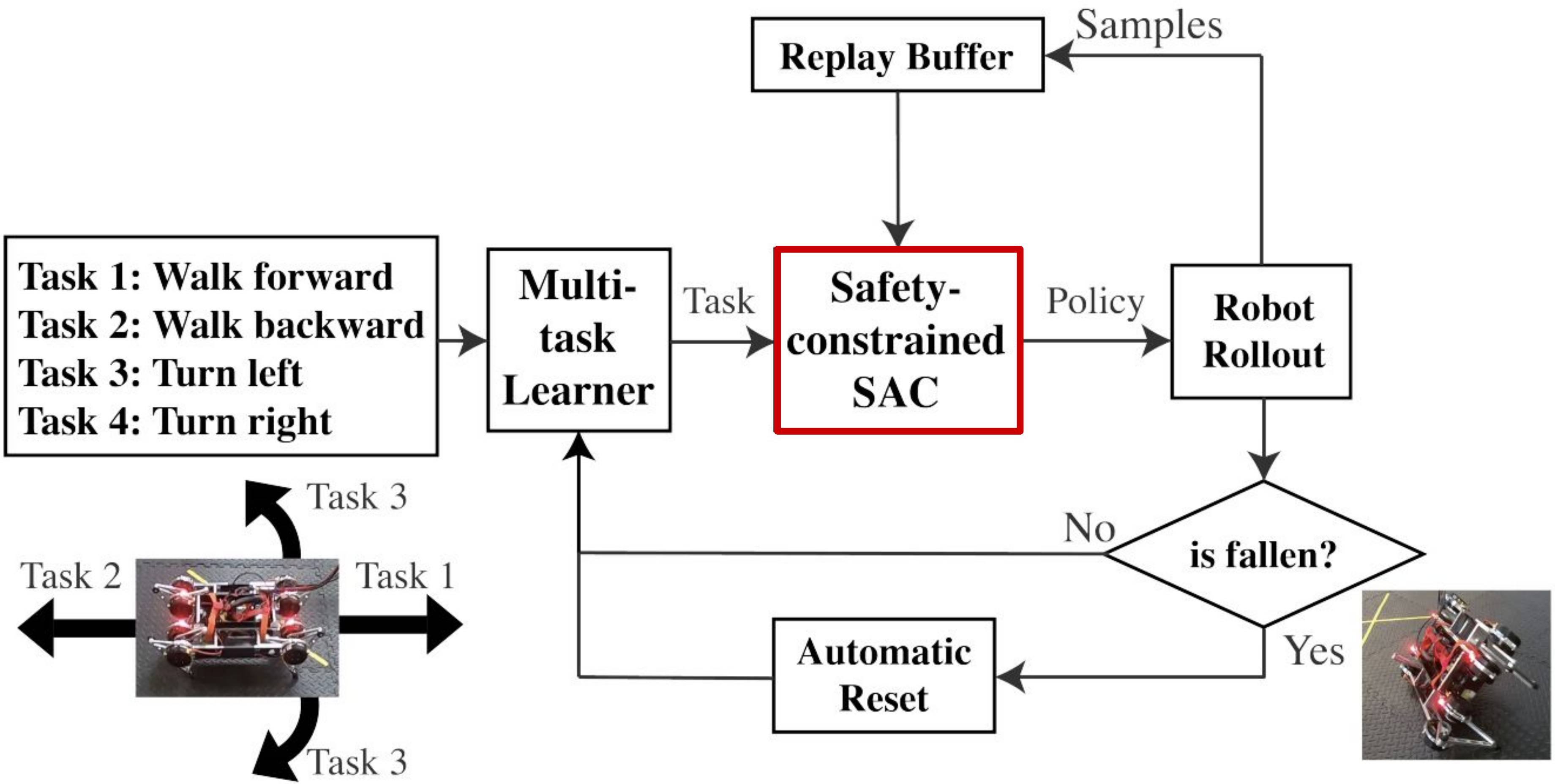
Pseudocode

1. Randomly initialize π, set $\lambda = 0$
2. Roll out policy π
3. Calculate policy gradient $\frac{\partial \mathcal{L}}{\partial \pi}$
4. $\pi = \pi + \alpha \frac{\partial \mathcal{L}}{\partial \pi}$
5. Calculate gradient $\frac{\partial \mathcal{L}}{\partial \lambda}$
6. $\lambda = \max(0, \lambda - \beta \boxed{\frac{\partial \mathcal{L}}{\partial \lambda}})$
7. Go to 2

Case Study: Learning Locomotion in Real World



[Learning to Walk in the Real World with Minimal Human Effort, Ha et al. CoRL 2020]


Replay Buffer
Samples
Task 1: Walk forward
Task 2: Walk backward
Task 3: Turn left
Task 4: Turn right
Multi-task Learner
Task
Safety-constrained SAC
Policy
Robot Rollout
is fallen?
No
Yes
Automatic Reset
Task 3
Task 2
Task 1
Task 3

Safety-Constrained SAC: Formulation

$$\max_{\pi \in \Pi} \mathbb{E}_{\tau \sim \rho_\pi} \left[\sum_{t=0}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

$$\text{s.t. } \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[f_s(\mathbf{s}_t, \mathbf{a}_t) \right] \geq 0, \ \forall t.$$

$$\boxed{\mathbb{E}_{\rho_\pi} \left[-\log \left(\pi_t(\cdot | \mathbf{s}_t) \right) \right] \geq \mathcal{H}}$$

Entropy
Constraints

Safety-Constrained SAC: Formulation

Safety Constraints

$$\max_{\pi \in \Pi} \mathbb{E}_{\tau \sim \rho_\pi} \left[\sum_{t=0}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

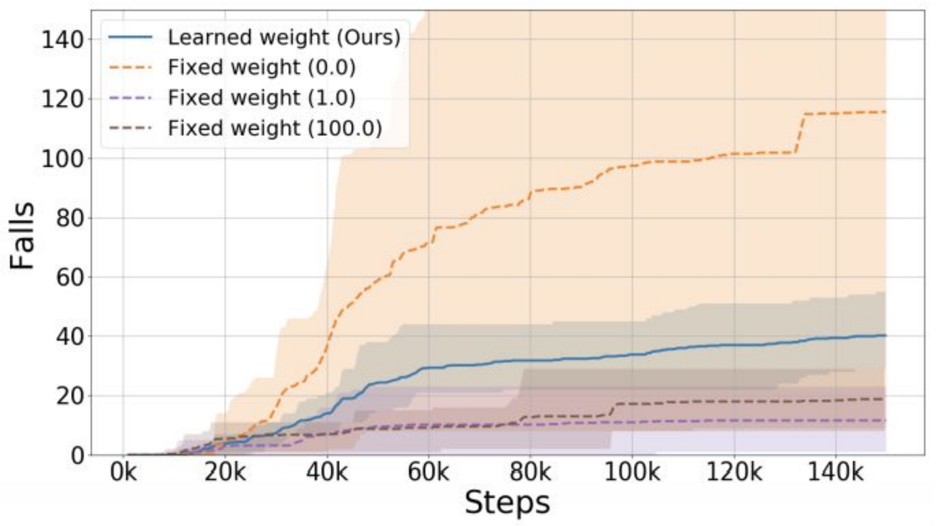
$$\text{s.t. } \boxed{\mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[f_s(\mathbf{s}_t, \mathbf{a}_t) \right] \geq 0, \ \forall t}$$

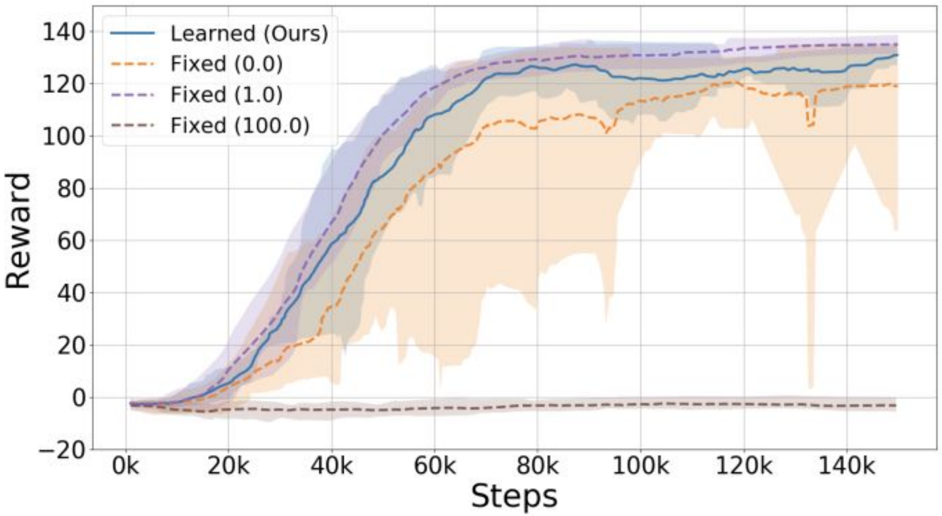
$$\mathbb{E}_{\rho_\pi} \left[-\log(\pi_t(\cdot | \mathbf{s}_t)) \right] \geq \mathcal{H}$$

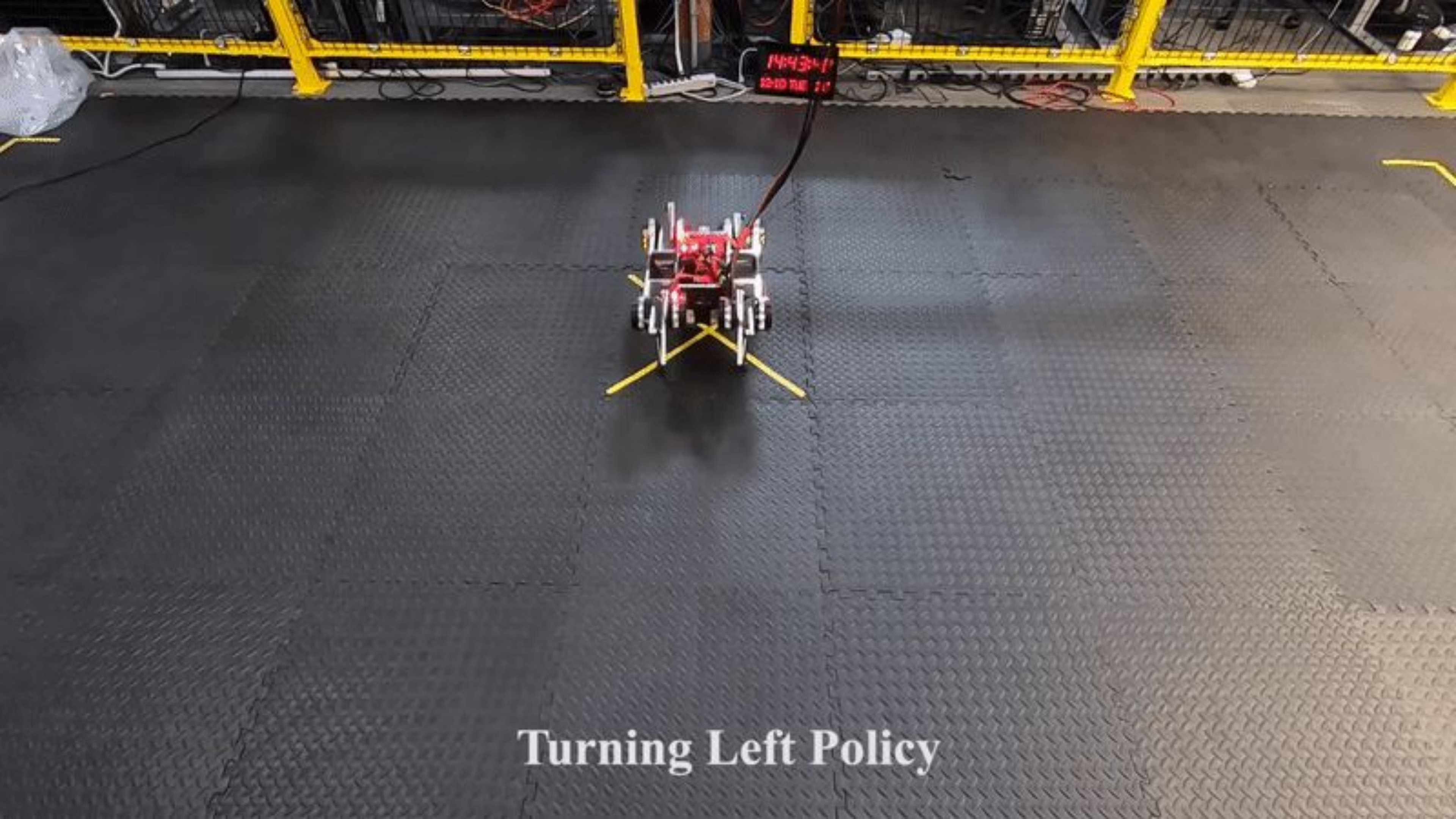
where

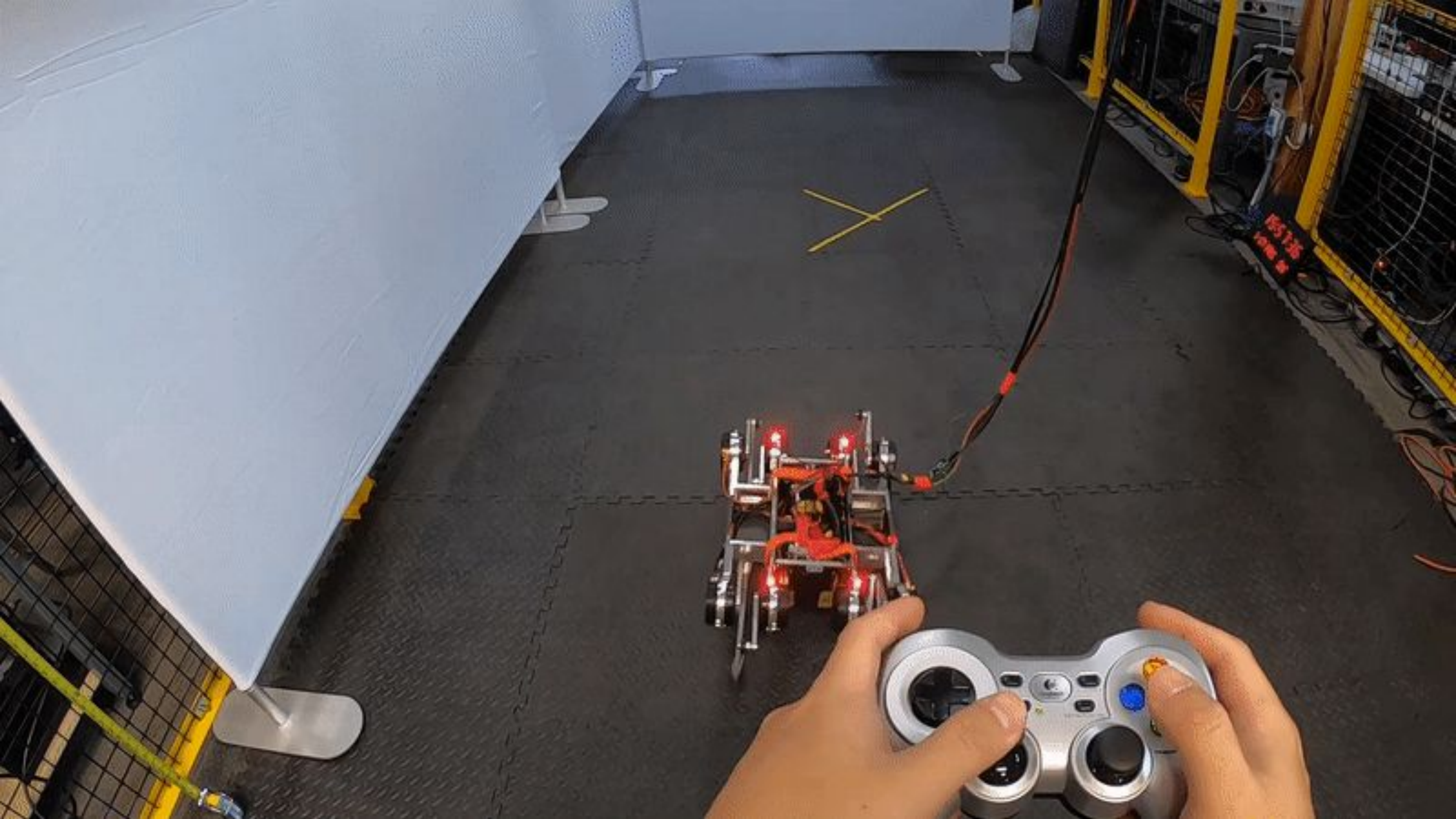
$$f_s(\mathbf{s}_t, \mathbf{a}_t) = \min(\hat{p} - |p_t|, \hat{r} - |r_t|)$$

Safety-Constrained SAC: Evaluation

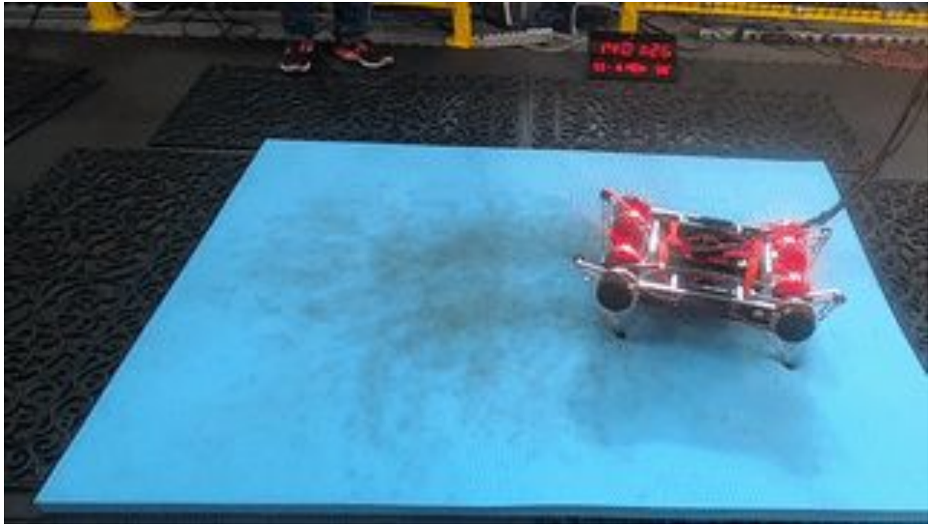




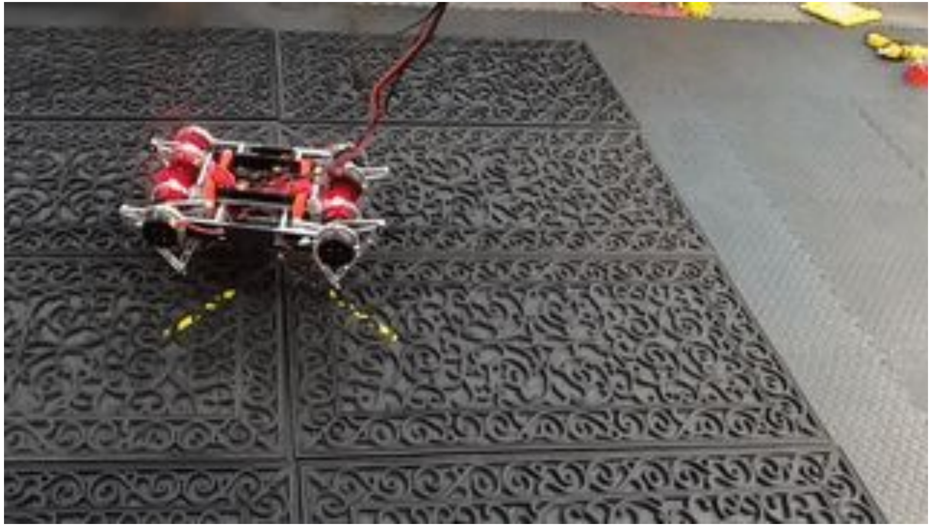

Turning Left Policy



Learning on challenging terrains



Memory foam

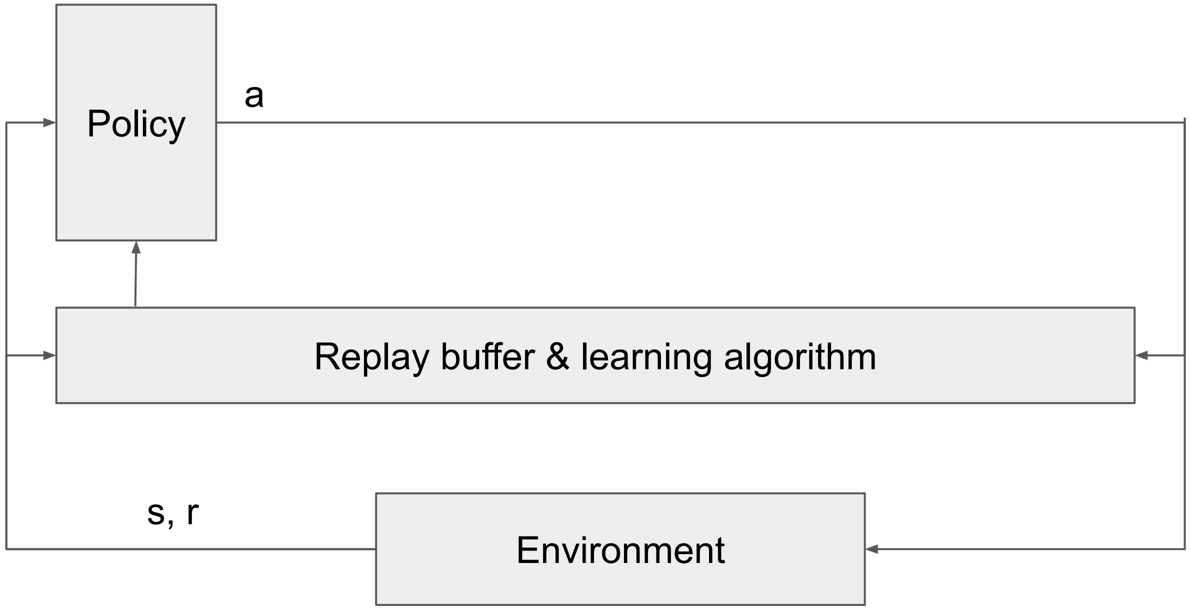


Rubber mat with crevices

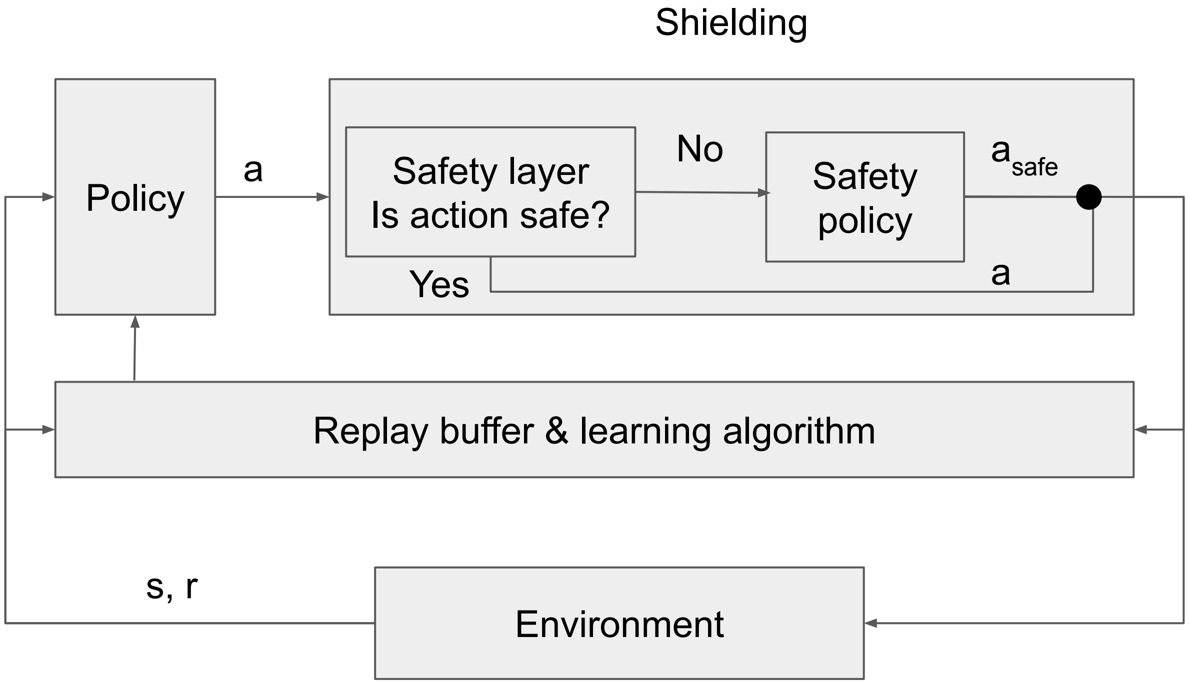
Limitations

- Unsafe events can still happen, though less frequently
- Hard to specify safety constraints in many applications
  - Can we learn safety constraints? [Recovery RL: Safe Reinforcement Learning with Learned Recovery Zones, Thananjeyan et al. RA-L, 2021]

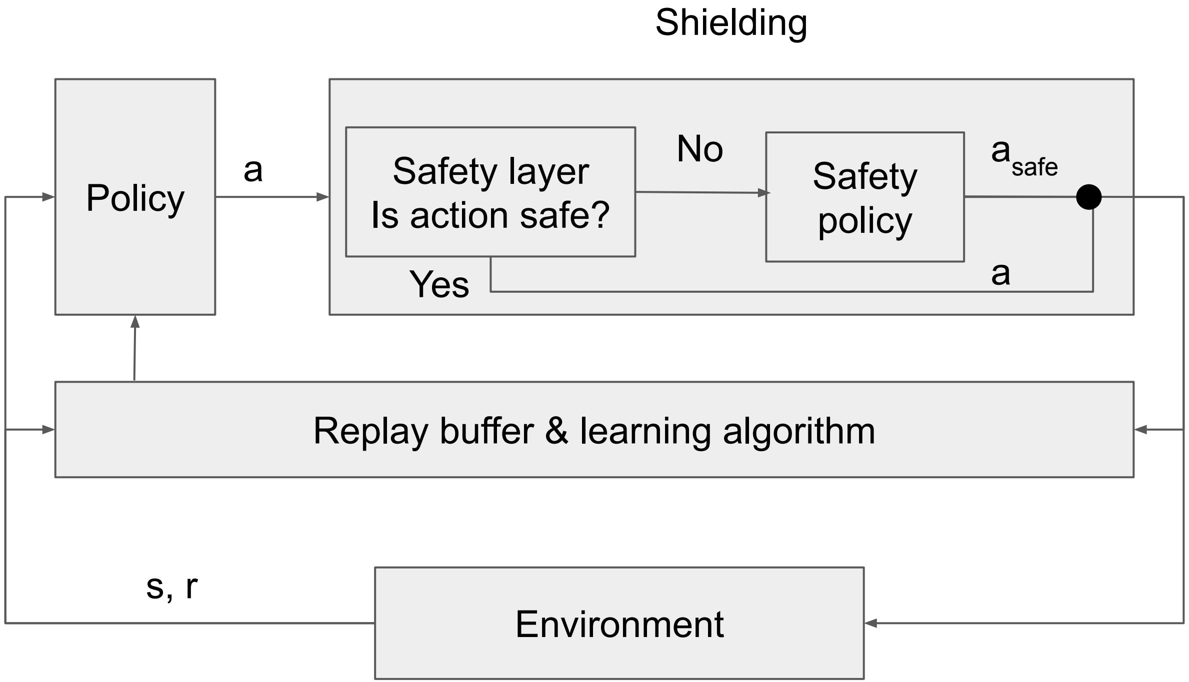
Safe Learning via Shielding



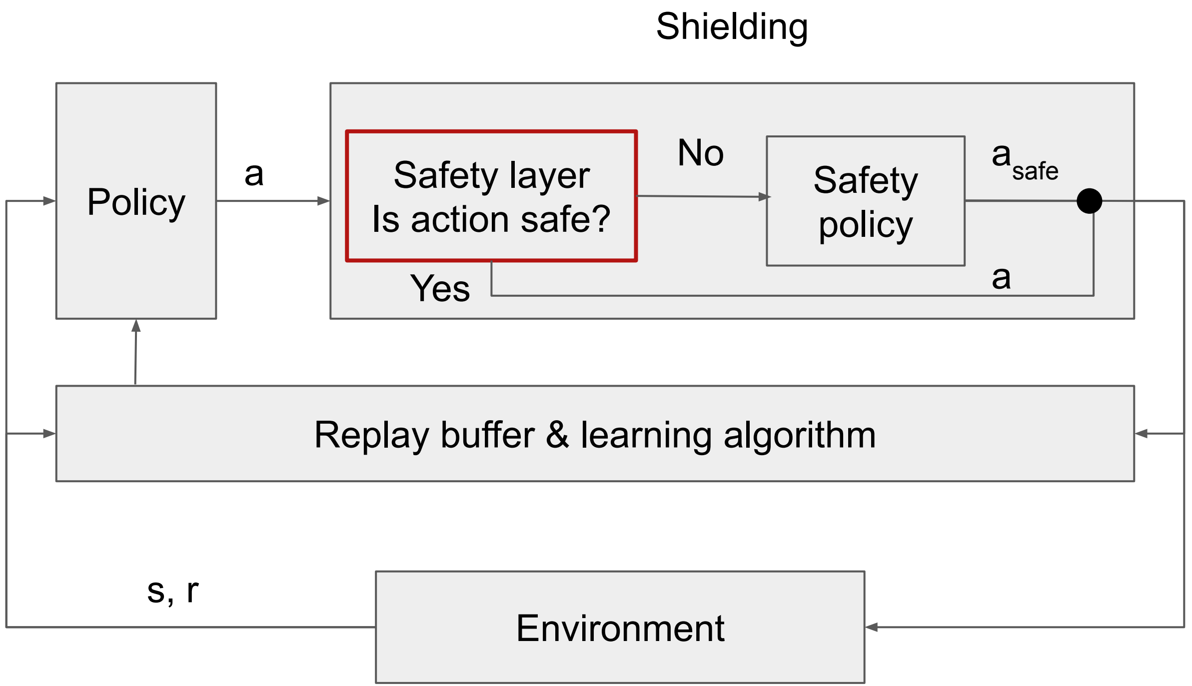
Safe Learning via Shielding



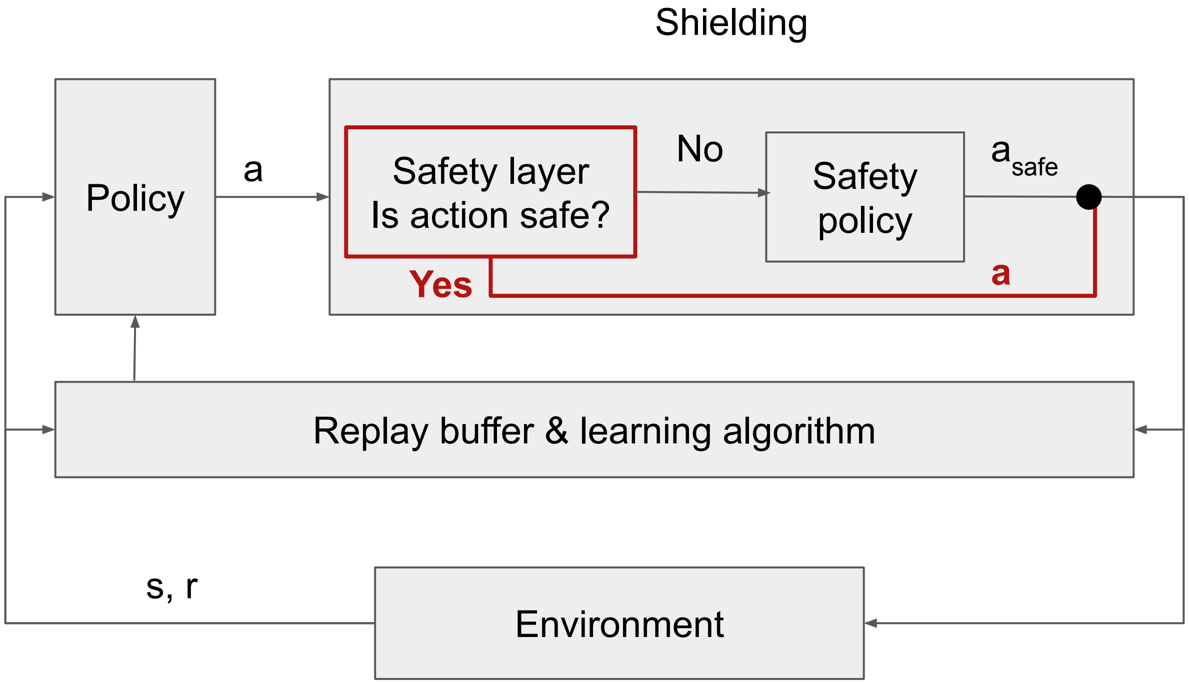
Safe Learning via Shielding



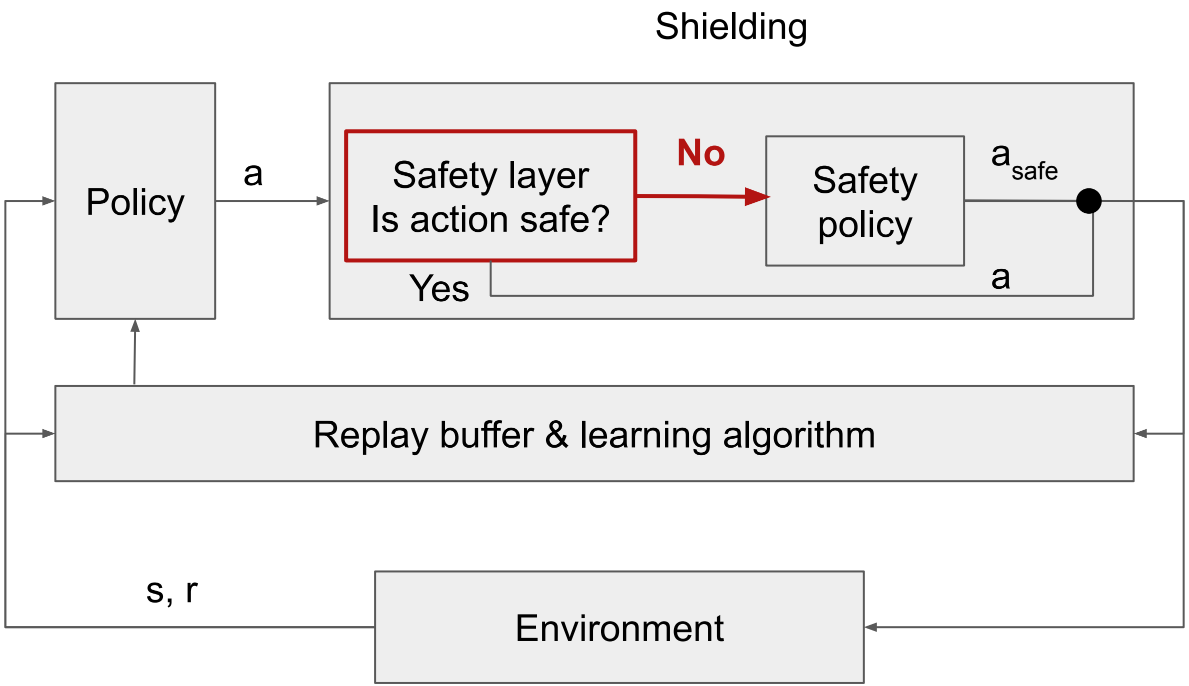
Safe Learning via Shielding



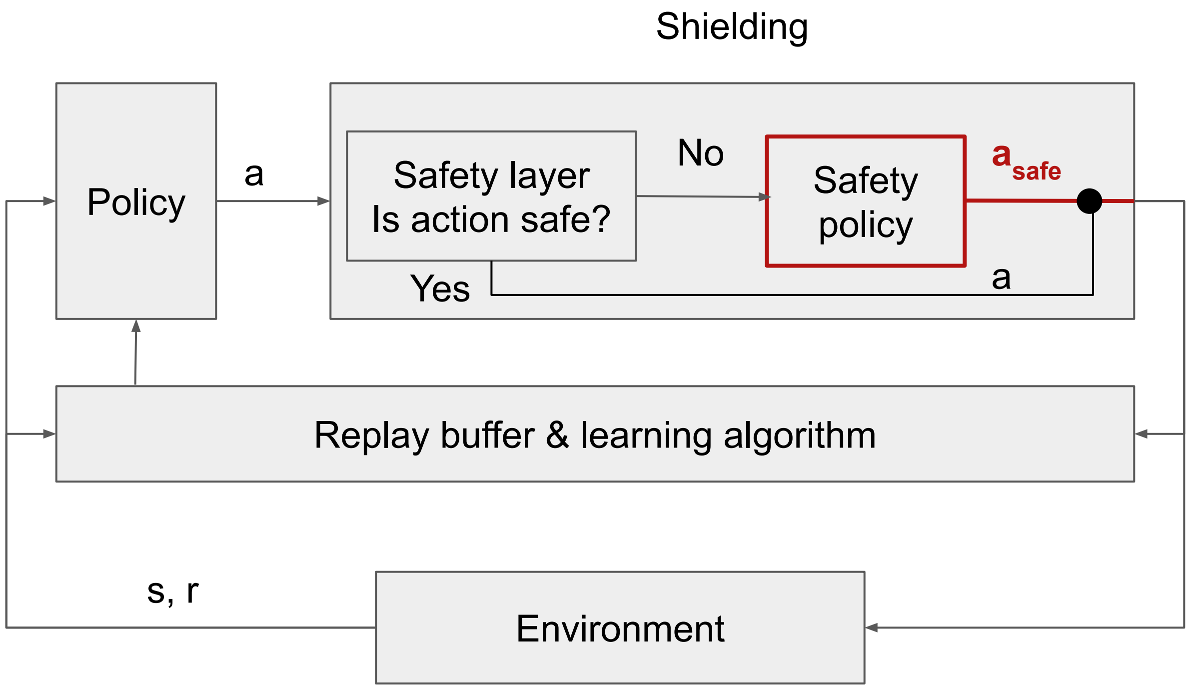
Safe Learning via Shielding



Safe Learning via Shielding

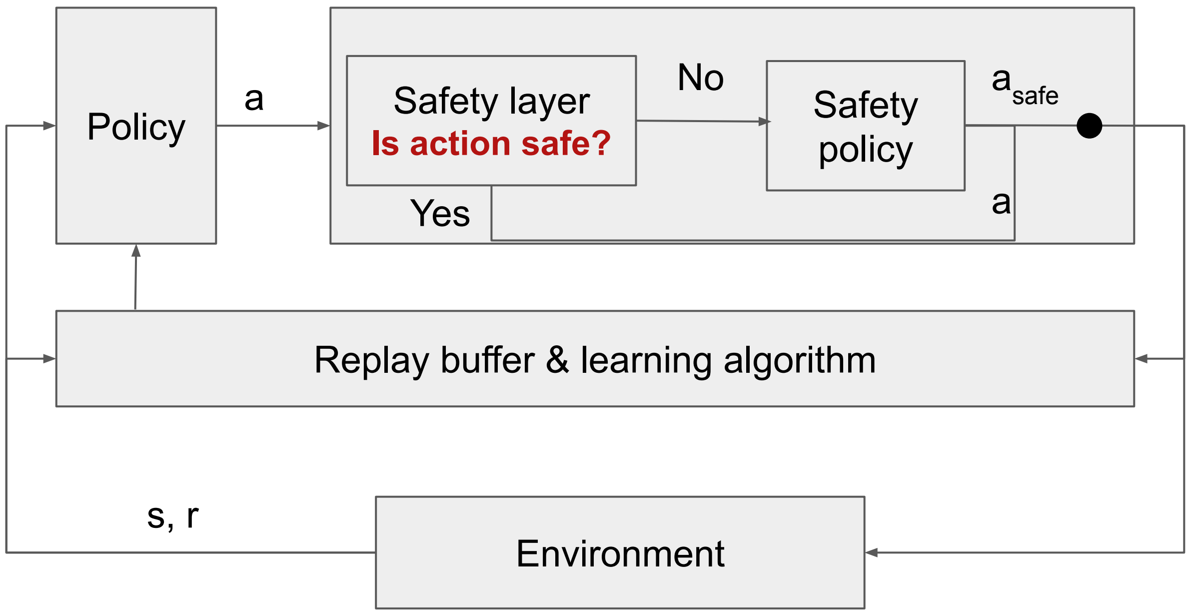


Safe Learning via Shielding



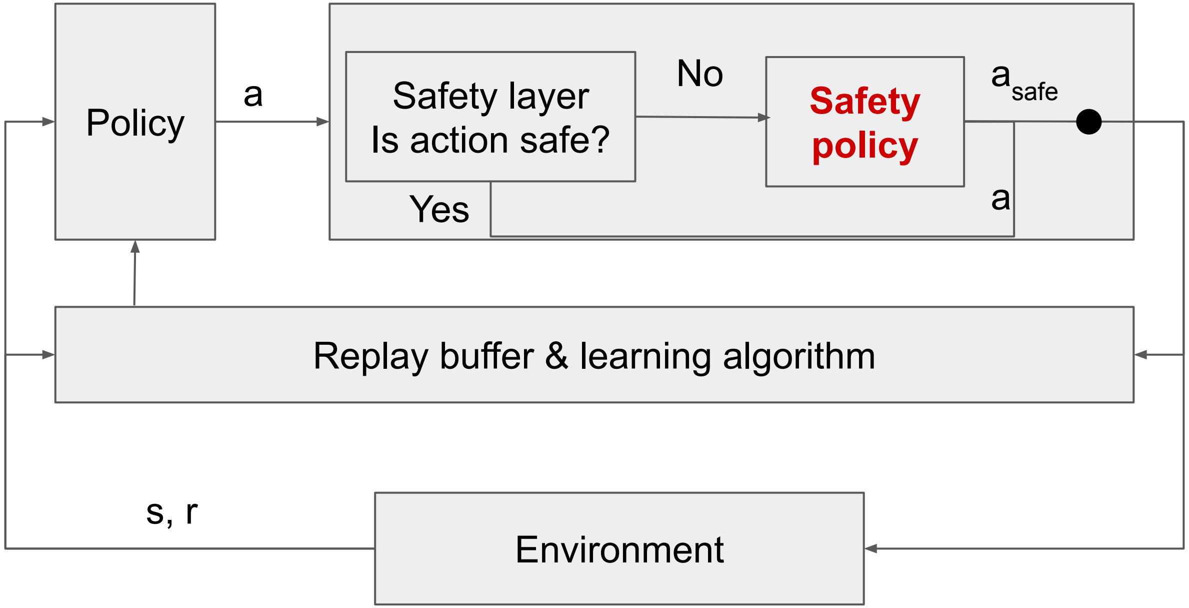
Two Questions

- How to decide whether an action is safe?



Two Questions

- How to decide whether an action is safe?
- Where does the safety policy come from?



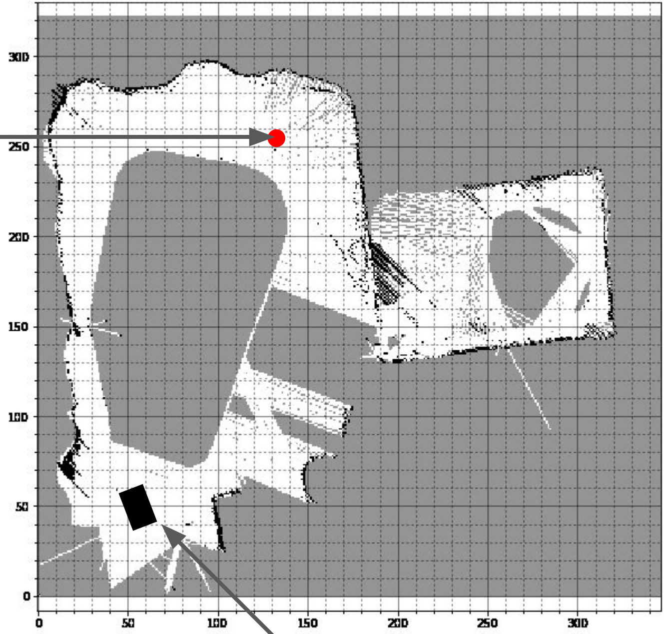
Two Questions

- How to decide whether an action is safe?
  - Manually specified threshold on actions (e.g. torque limit, joint limit)
  - Future state in safe set by model rollout
  - Query pretrained safety critic: [Learning to be Safe: Deep RL with a Safety Critic, Srinivasan et al. 2020]
- Where does the safety policy come from?
  - Simple engineered solution (e.g. stop)
  - Traditional model-based control (e.g. model-predictive control)
  - Learned safety policy in simulation
    - domain randomization
    - adversarial training: [Robust Adversarial Reinforcement Learning, Pinto et al. ICML, 2017]

Case Study: Navigation

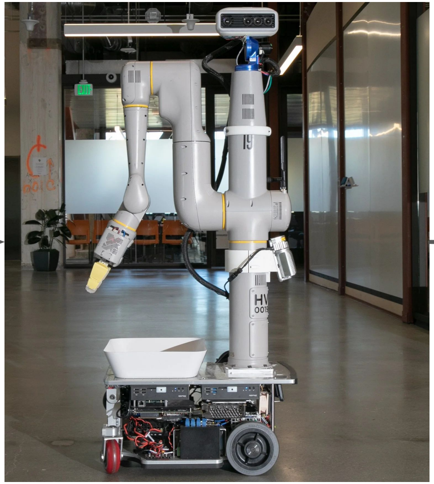
Observations

Goal



Robot pose

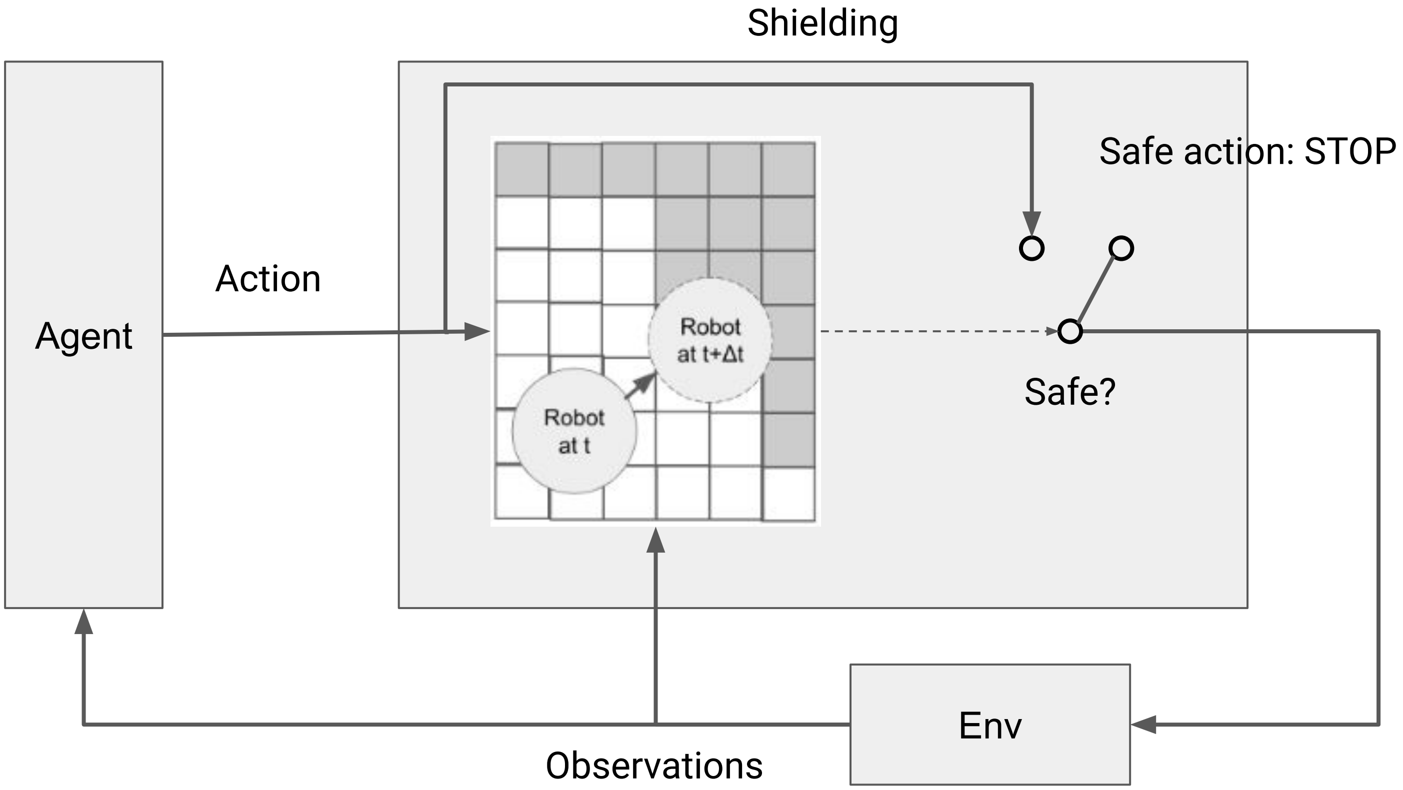
Agent



Actions

Desired turning rate
Desired speed

System Overview

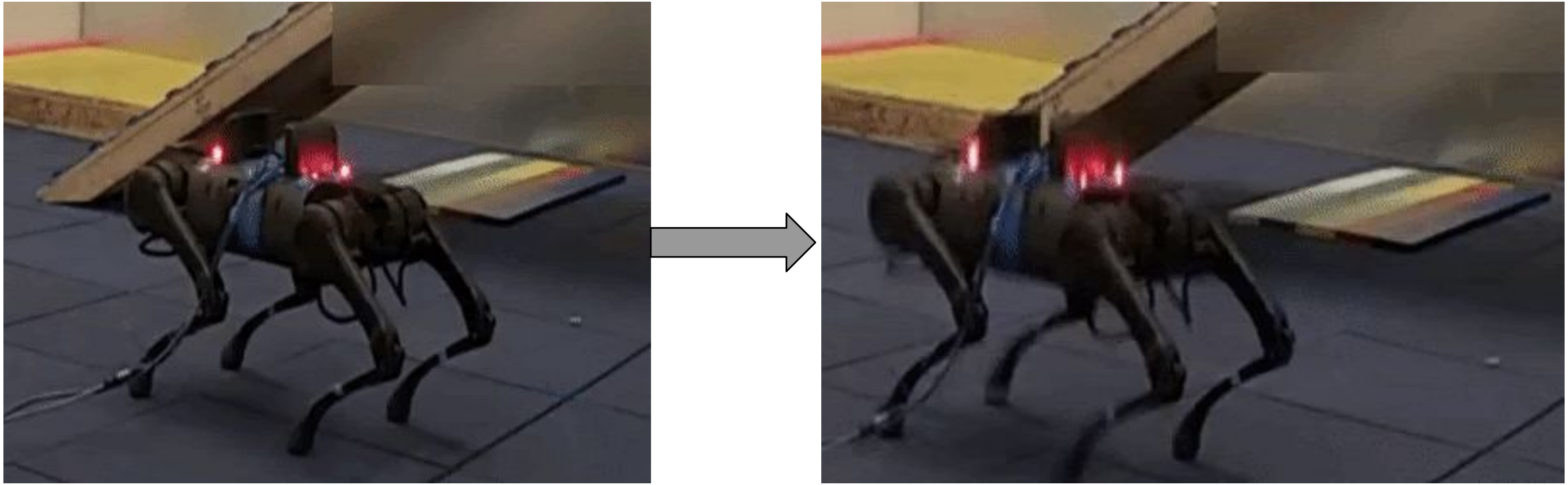


Result of Shielding

Without shielding

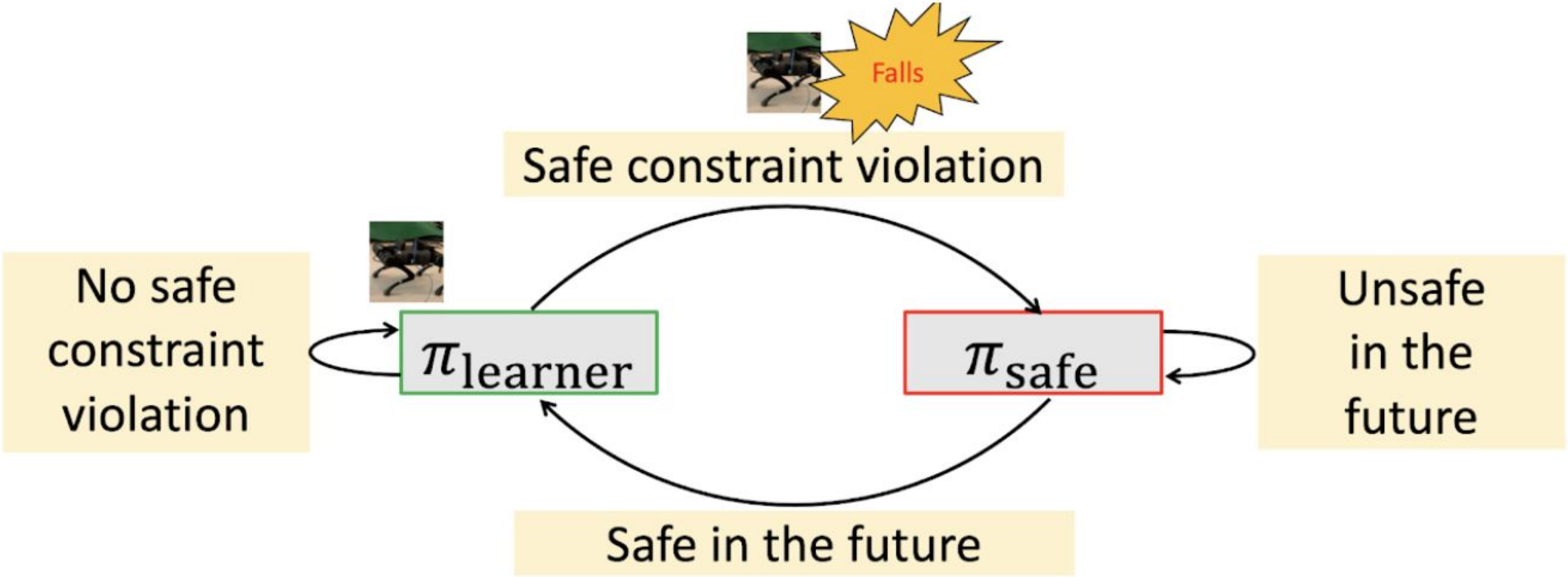
With shielding

Case Study: Locomotion



[Safe Reinforcement Learning for Legged Locomotion, Yang et al. 2022, Under Review]

System Overview



[Safe Reinforcement Learning for Legged Locomotion, Yang et al. 2022, Under Review]

System Details

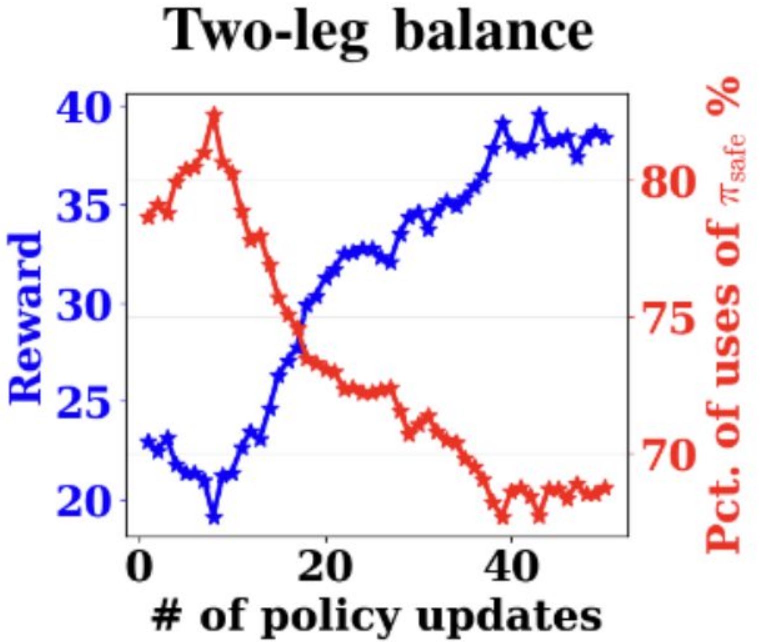
- Safe constraint
    - Thresholds on roll and pitch of the base
- Safe policy
    - Model-predictive control based on simplified dynamics
    - RL to modulate MPC parameters (stepping frequency, swing location, etc.)
    - Trained in simulation with domain randomization

[Safe Reinforcement Learning for Legged Locomotion, Yang et al. 2022, Under Review]

Training Process (Timelapse)

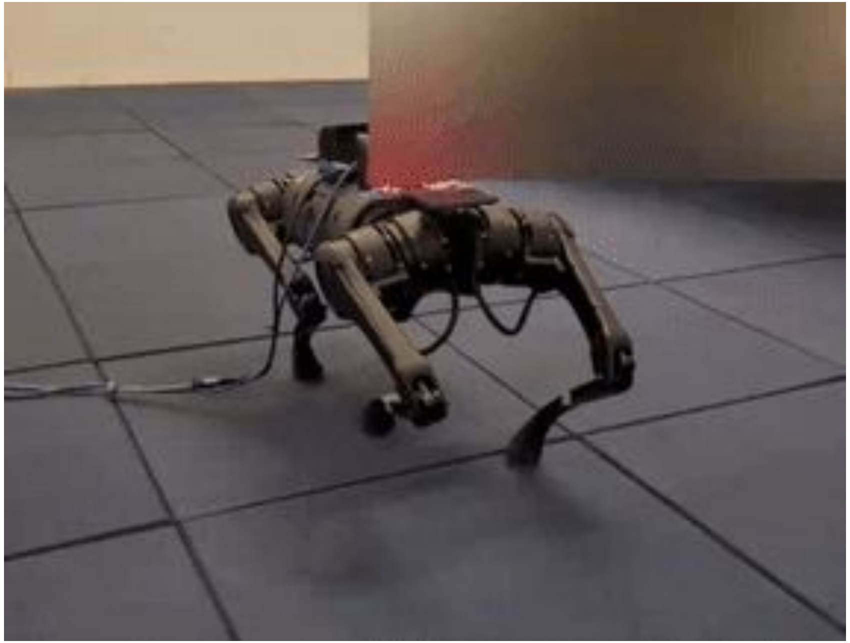
[Safe Reinforcement Learning for Legged Locomotion, Yang et al. 2022, Under Review]

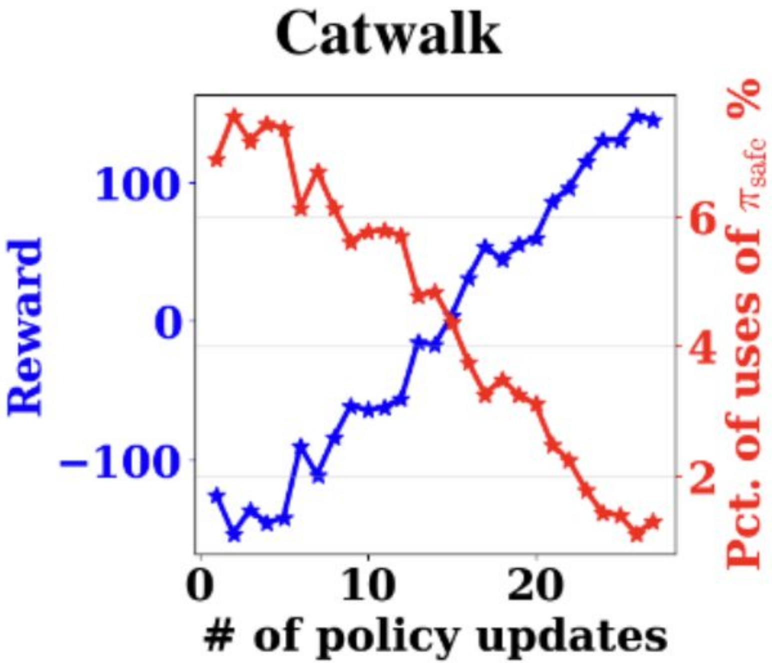
Results: Two-Leg Balance



[Safe Reinforcement Learning for Legged Locomotion, Yang et al. 2022, Under Review]

Results: Catwalk





[Safe Reinforcement Learning for Legged Locomotion, Yang et al. 2022, Under Review]

Limitations

- Switching between two policies can lead to unsafe jerky motions
- Hard to balance between safety and learning efficiency
- Difficult to design or learn safety policies for complex tasks

Summary

- Formulate safety as constraints
- Two ways to improve safety during learning
  - Constrained Markov Decision Process
  - Safe learning via shielding