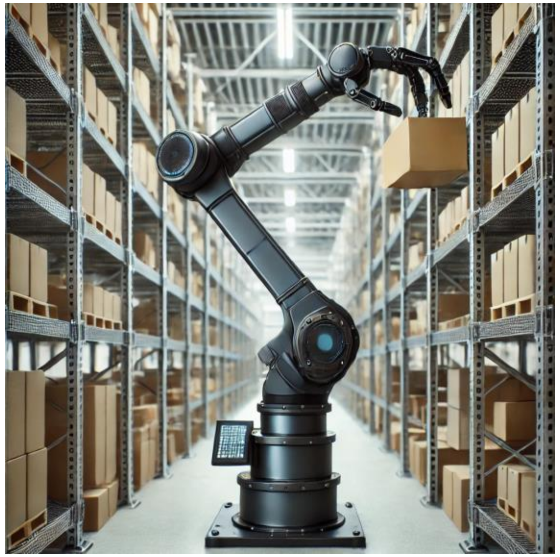
Safety through feedback in Constrained RL

Shashank Reddy[1], Pradeep Varakantham[1], Praveen Paruchuri[2]
[1]Singapore Management University
[2]International Institute of Information Technology, Hyderabad

Safety in Reinforcement Learning

- Safety is critical when deploying RL agents in real-world environments
- Agents must adhere to stringent safety constraints- such as speed limits, proximity to humans, operational boundaries, etc



Constrained Reinforcement Learning

Constrained Markov Decision Process A Constrained MDP [4] introduces a function $c(s, a) \in \mathbb{R}$ and a cost threshold $c_{max} \in \mathbb{R}$ that defines the maximum cost that can be accrued by a policy. The set of feasible policies is defined as $\Pi_c = \{\pi \in \Pi : \mathcal{J}^c(\pi) \leq c_{max}\}$. A policy is considered to be *safe* w.r.t c if it belongs to Π_c.

In this paper, we consider the constrained RL problem defined as,

$$\pi^* = \underset{\pi \in \Pi_c}{\operatorname{argmin}} \; \mathcal{J}^r(\pi)$$

We incorporate an additional constraint enforcing the cost function to be binary, i.e, $c(s, a) \in \{0, 1\}$. This ensures that each state-action pair is inherently categorized as either *safe* or *unsafe*. We opt for this approach because it is simpler for human evaluators to assign a binary safety value to state-actions when assessing policy safety, as emphasized in [31].

Challenges:

SMU SINGAPORE MANAGEMENT UNIVERSITY

- Cost function design
- Safety can depend on individual preferences
- Expensive to evaluate

Solution: Learn the cost function from feedback!

Learning from Feedback

- Collected from a **human** or an **expensive to evaluate system**
- Collected **offline** in between rounds
- Feedback must be **minimized**
- Feedback is **binary** *(safe / unsafe)*
- Cost is **inferred** from feedback

Feedback Collection

(Naïve Solution)

- Elicit feedback for every state of every trajectory collected by the agent
- Not feasible for Deep RL!

Efficient Feedback Collection (Proposed)

- Elicit Feedback for **longer horizons** (trajectory segments)
- **Selectively** sample trajectories that are shown to the evaluator (most informative)

Feedback over Longer Horizons

- Break the trajectory into segments of length ***k***
- Elicit feedback for each segment
- Segment is labeled *safe* if *all* states are *safe*, else it is marked *unsafe*

Inferring the Cost function

- Label all states in the *safe* segment 0
- Label all states in the *unsafe* segment 1
- Minimize the *cross-entropy loss*
- Why this works:
  - *True safe states receive both labels 0 and 1*
  - *True unsafe states receive label 1 only*

Assumption

- Each *safe* receives label 0 at least once
- Guaranteed only when segment length is 1
- In practice works for longer horizons as well

Efficient Subsampling of Trajectories

Motivation:

Sample the most informative trajectories to show to the evaluator

Idea:

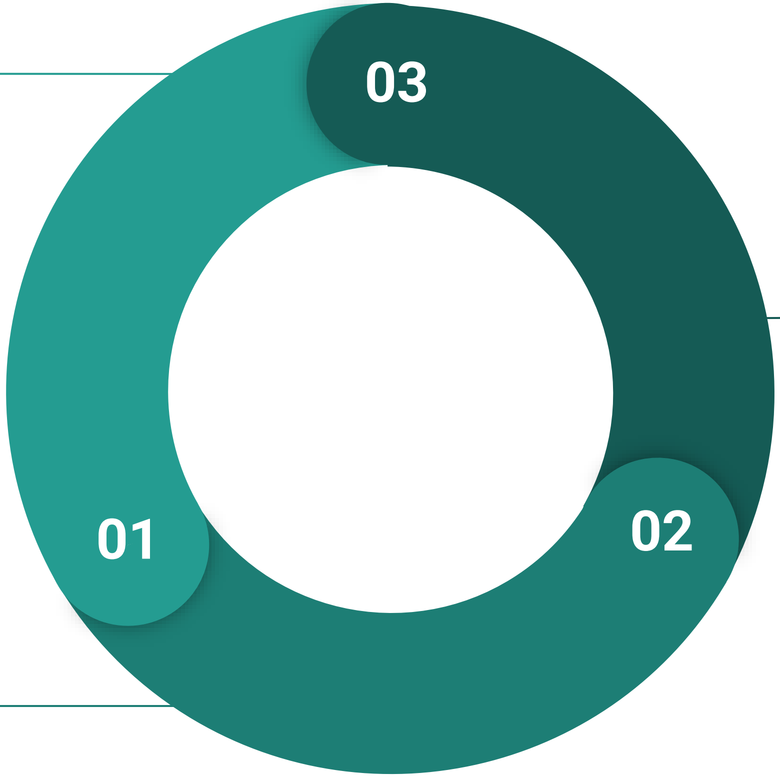
As the agent improves ⇨ explores new states ⇨ cost function prone to errors

Therefore, sample novel trajectories, defining a trajectory as novel if it includes at least ***e*** unseen states

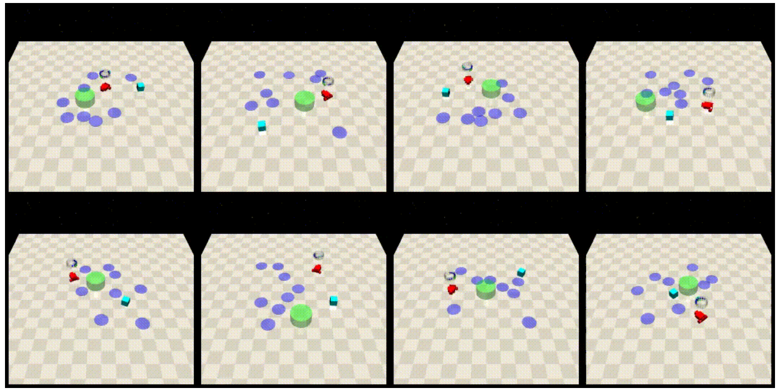
We call this method *novelty sampling*

Advantage: Automatically stops eliciting feedback once the trajectories are no longer novel

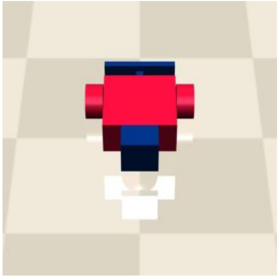
Our Approach: RL from Safety Feedback (RLSF)



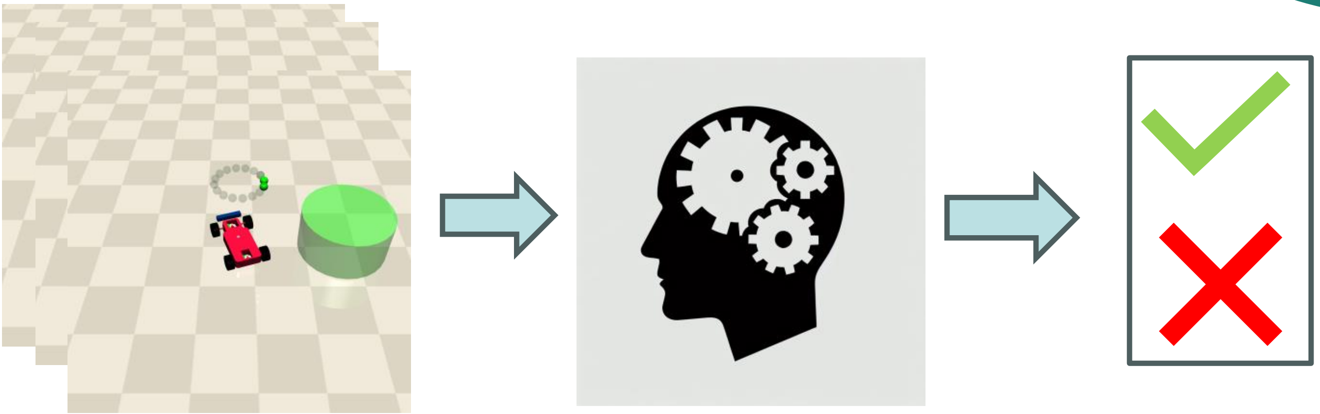
1. Collect Data from the Agent



3. Improve the Agent using the Feedback

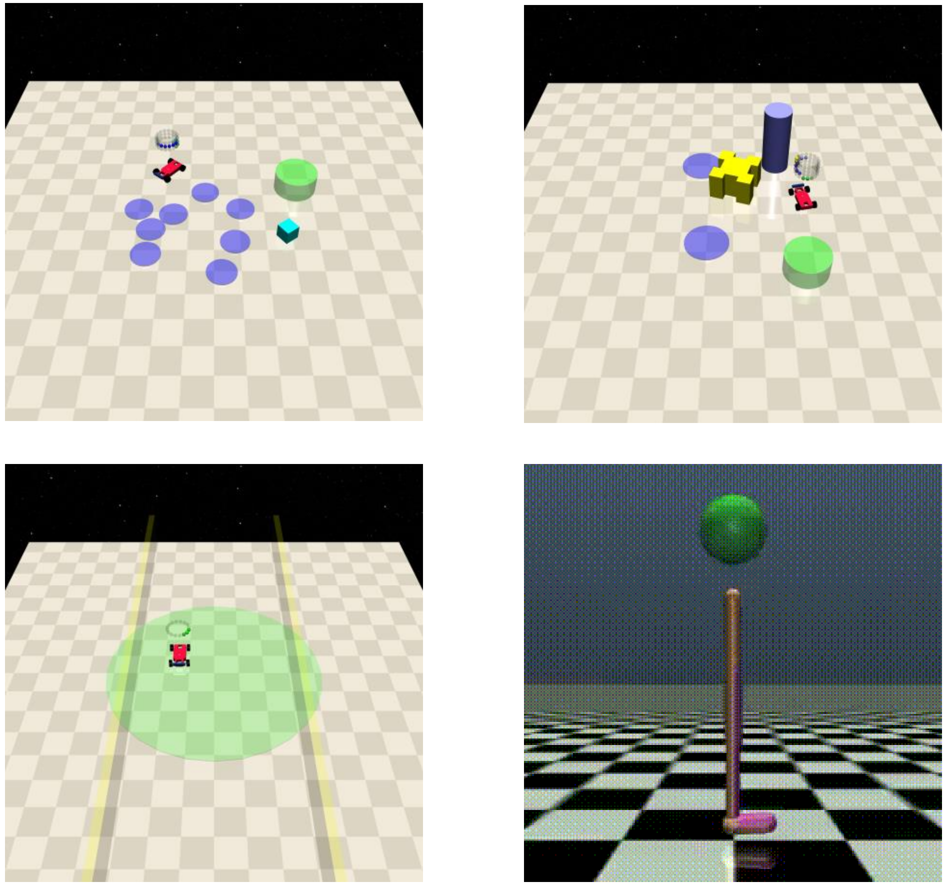


2. Elicit Safety Feedback

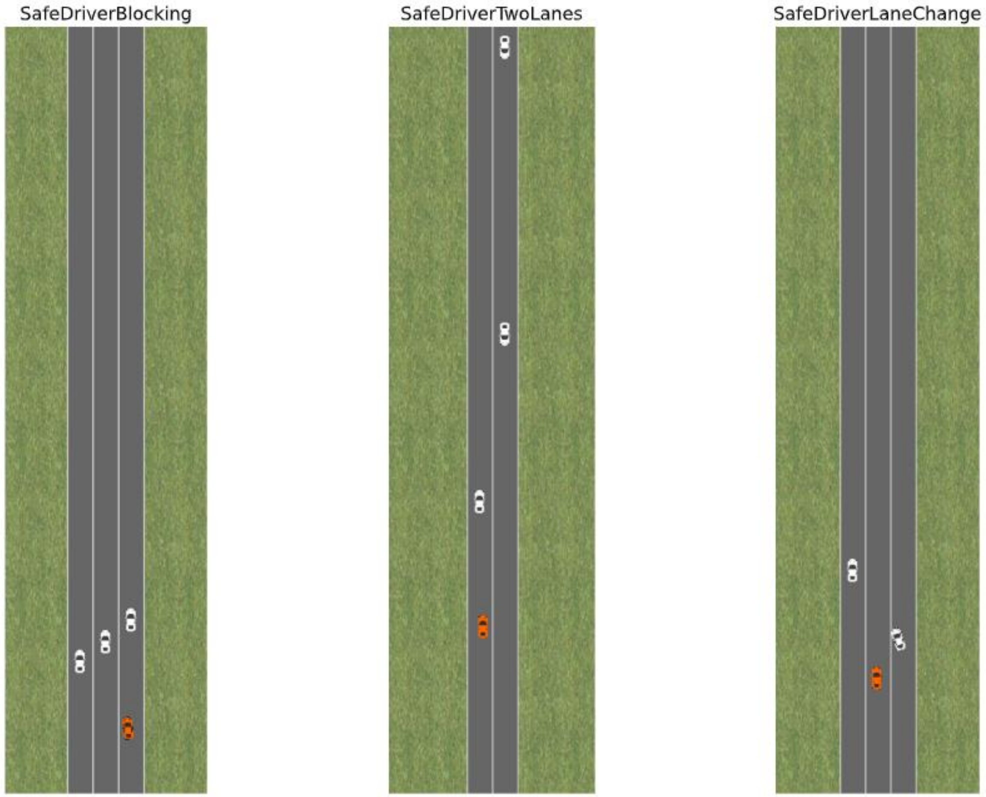


Experiments

Safety Gymnasium [Ji J., et al. 2023]



Driver [Lindner D, et al. 2022]



Quantitative Results

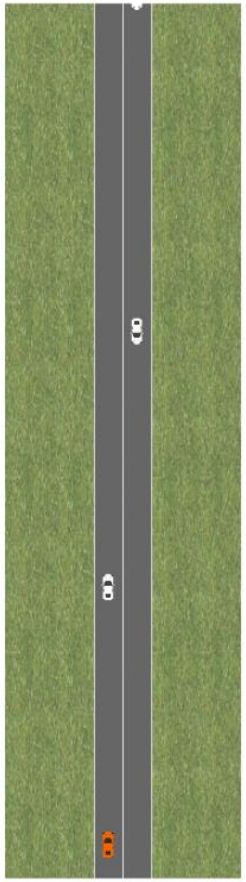
Table 1: Performance of different algorithms on the Safety Benchmarks. The first 7 environments represent the *hard* constraint case. The remaining environments illustrate the *soft* constraint case, with values in brackets indicating the cost threshold. Each algorithm is run for 6 independent seeds. (orange) and (blue) indicate the best performance in the known costs and inferred costs settings, respectively. Algorithms with a cost violation (C.V) rate below 1% are deemed to have equal performance in terms of safety.

Environment		Cost Known (Best Run)		Cost Inferred (Mean ± Standard error)		
		PPOLag	SIMKC	SDM	SIM	RLSF (Ours)
Point Circle	Return	45.26	46.09	36.20 ± 3.95	22.26 ± 9.59	36.42 ± 1.78
	C.V Rate (%)	0.4	0.43	11.43 ± 0.69	35.21 ± 10.09	1.9 ± 0.09
Car Circle	Return	14.34	15.21	5.18 ± 2.48	6.34 ± 2.87	9.37 ± 0.97
	C.V Rate (%)	0.84	5.4	6.2 ± 6.18	4.53 ± 4.00	0.54 ± 0.30
Biased Pendulum	Return	717.43	983.27	495.58 ± 160.84	577.15 ± 184.31	721.48 ± 111.49
	C.V Rate (%)	0.0	0.1	39.91 ± 17.05	48.58 ± 21.67	0 ± 0
Blocked Swimmer	Return	22.62	21.05	86.96 ± 10.69	2.15 ± 8.58	16.09 ± 1.44
	C.V Rate (%)	3.91	0.01	92.8 ± 1.65	13.33 ± 12.11	0.01 ± 0.01
HalfCheetah	Return	2786.71	2497.82	3031.7 ± 336.48	257.34 ± 147.35	2112.63 ± 161.26
	C.V Rate (%)	0.42	0.06	59.4 ± 8.28	0.0 ± 0.0	0.06 ± 0.01
Hopper	Return	1705.00	1555.25	1097.57 ± 56.35	990.08 ± 8.66	1408.71 ± 27.3
	C.V Rate (%)	0.19	0.02	0.0 ± 0.0	0.0 ± 0.0	0.29 ± 0.02
Walker2d	Return	2947.25	2925.23	2195.94 ± 134.21	993.38 ± 17.69	2783.29 ± 57.51
	C.V Rate (%)	0.16	0.0	1.58 ± 1.53	0.0 ± 0.0	0.05 ± 0.01
Point Goal	Return	26.16	26.1	1.61 ± 1.8149	10.86 ± 4.1	24.65 ± 0.59
	Cost (40.0)	34.19	31.83	30.57 ± 13.29	52.76 ± 12.85	35.08 ± 1.08
Car Goal	Return	27.37	26.44	1.05 ± 2.83	10.88 ± 7.1	24.28 ± 2.1
	Cost (40.0)	41.67	35.41	34.71 ± 9.87	33.33 ± 11.26	41.25 ± 2.27
Point Push	Return	6.00	10.84	0.16 ± 0.14	3.63 ± 1.77	2.68 ± 1.03
	Cost (35.0)	26.08	26.96	22.89 ± 5.95	45.43 ± 3.86	30.51 ± 3.4
Car Push	Return	3.07	2.68	-3.04 ± 3.3	1.56 ± 0.46	1.54 ± 0.51
	Cost (35.0)	20.53	20.95	23.25 ± 7.78	36.55 ± 1.48	27.69 ± 1.19

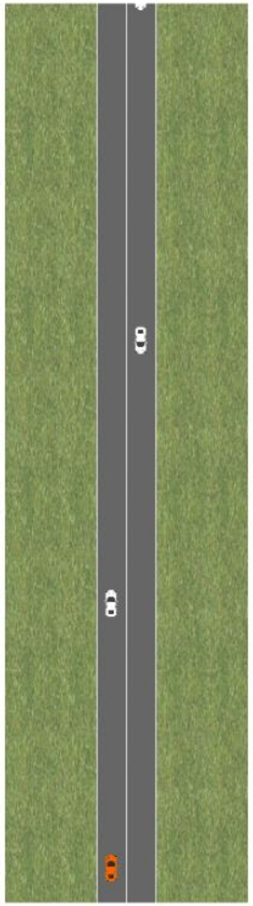
SMU SINGAPORE MANAGEMENT UNIVERSITY

Qualitative Results (Safe vs Unsafe Agents)

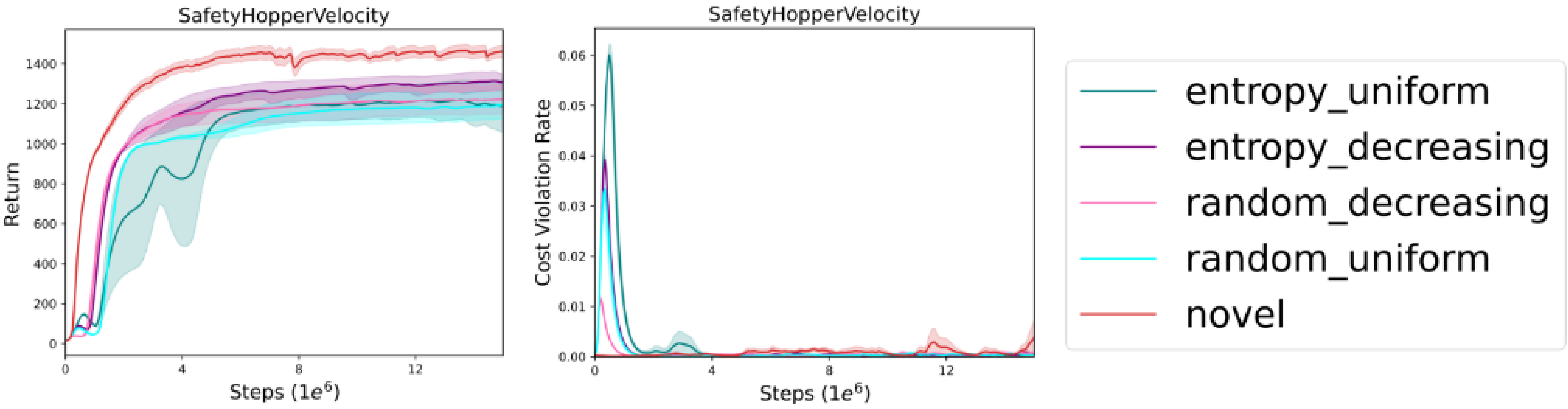
PPO | Reward: -0.00 | Cost: 0.00

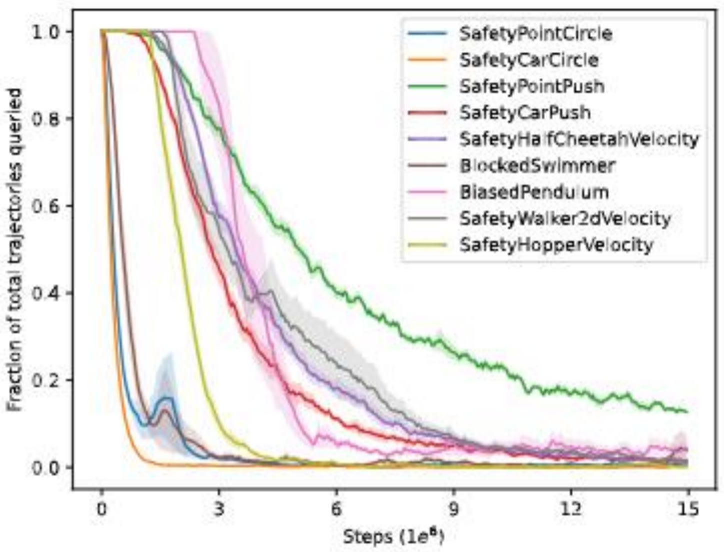


RLSF | Reward: -0.00 | Cost: 0.00



Ablation on sampling methods





Novelty reduces over time

Future Work

- Some of the more complex environments (*Goal, Push, Driver*) require *state level* feedback; rest use *trajectory level* feedback
- Synthetic feedback was used in the experiments, real world feedback is more *noisy* ⇨ human subject experiments

Thanks

Image Credits: AI-generated images (DALL-E 3) were used to illustrate key points. The images in Slide 8 (Overall Architecture) are sourced from [Ji J. et al., 2023], licensed under the Apache License 2.0.