# Safe Controller for Reinforcement Learning: Safety and Optimization

**Samar Rahmouni, Prof. Giselle Reis**
*Carnegie Mellon University, Computer Science Department*

## Introduction

Implementing autonomous agent controllers that can robustly and efficiently adapt to different dynamic and complex scenarios is still an open challenge in robotics and AI. In the case of Reinforcement Learning (RL), an autonomous car trained by trial-and-error is bound to learn how to drive. This makes RL nearly impossible to deploy in the real world.
We investigate both the **security** and the **interpretability** aspect of reinforcement learning in a cooperative adaptive cruise control inspired from [1], in the aim of finding how formal security frameworks can guide the representation, robustness and extrapolation of knowledge in Reinforcement Learning agents.

## Car Platooning : Case Study

Multiple vehicles following each other, how do we ensure that we minimize the gap while guarantying no crashes?

**Why is it a problem we should care about?**
Taking less space on the roads and allowing more vehicles to occupy highways => Less traffic and faster commute, especially when we consider the deployment of autonomous vehicles.

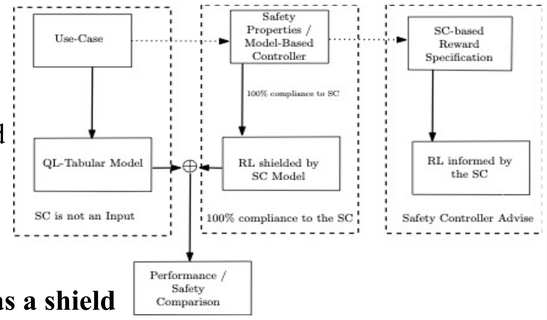
**Why is it a difficult problem?**
The difficulty stems from the uncertainty of roads, rather than the physics of minimizing distances. Consider (1) stop signs and traffic lights, (2) faulty communication between vehicles, (3) faulty sensors or even (4) malfunctioning breaks.

**It is a key problem that requires a tradeoff between optimization and safety.**

**What has been done to tackle it?**
- Safe controller (SC) that utilizes soft constraints to compute safe velocities [1]. **Safety.**
- Deep reinforcement learning on an actor/critic policy to optimize gap control [2]. **Optimality.**

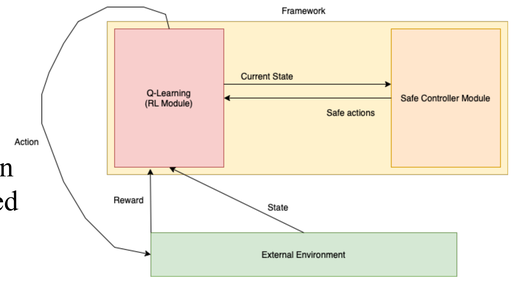
## Methodology

**I. Approach**
We consider the three following approaches to compare optimization and safety, basic RL, shielded RL using SC and Logic-based inference RL.




**II. Architecture for SC as a shield**

In the following work, we focused on (1) formalizing the car platooning scenario, (2) implement it and (3) combine our implementation with a simplified SC inspired from [1].




## Scenario

**Model in Reinforcement Learning**
- S - A state vector to denote *time*, *position*, *velocities*.
- A - An action vector for possible accelerations.
- R - A reward function (i.e., given the distance between the vehicles.
- $\phi$ - A transition function s.t. $\phi(s_t,a_t) = s_{t+1}$
- $\gamma$ - A discounting factor in [0,1]

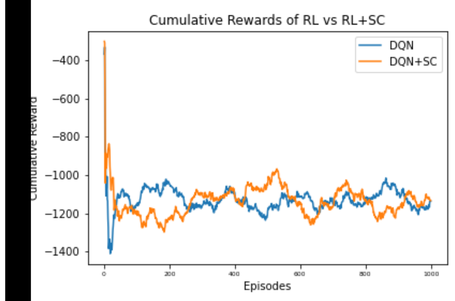
Assumes a joint state representation of the agents

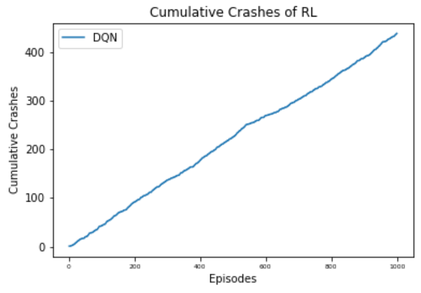
**Model in the Safe Controller**
- Local Knowledge Base
  - Set of grounded facts p@t
- Events = ev@t
  - Equivalent to task@0
- Executable semantics
- System configuration search to ensure that no path given those accelerations result in a crash.

## Results

**Optimization**
Assessed through the cumulative rewards.

**Safety**
Assessed through the cumulative crashes.




Note that rewards are negative since we want to minimize the gap.

## Discussion

**Comparing Safety Guarantees**
Linear cumulative crashes indicate that almost every episode ends with a crash, thus the RL never learns not to bump into other vehicles. Using a SC, safety is always guaranteed.

**Comparing Optimization i.e., minimal gap**
RL+SC achieves the max cumulative compared to basic RL. It also never reaches the max-min of the basic RL. This can be explained by never taking the negative reward from crashing. Overall, they both achieve a similar average of cumulative reward, indicating that both converge to an optimal solution, i.e., stay constant which is evident in the resulting policy.

**Comparing training time**
For the same number of episodes and max-steps per episode, the RL+SC takes on average longer as the computation of the SC gets more complex, although it takes on average less episodes to train.

## Current work

Following those results on a simplified car platooning scenario, the next step is to add a layer of uncertainty by **considering miscommunication**. Second, another aspect of the project is the third approach: **Logic-based inference RL**. Rather than enforcing the decision of the SC, the RL is informed by communication of a reward system towards improving its policy.
A key question is the interpretability of those results which comes with considering **inductive reasoning** in a data-driven algorithm towards balancing between optimality and strict safety.

## References

[1] Yuri Gil Dantas, Vivek Nigam, and Carolyn Talcott. A formal security assessment framework for cooperative adaptive cruise control. In *2020 IEEE Vehicular Networking Conference (VNC)*, pages 1–8, 2020.
[2] Ramadan, Amr & Abdulaaty, Omar & Hussein, Ahmed & Shehata, Omar. (2020). Reinforcement Learning Based Approach for Multi-Vehicle Platooning Problem with Nonlinear Dynamic Behavior.

## Acknowledgments

We would like to thank Vivek Nigam for useful discussions and continuous valuable feedback on the project. The following work is done as part of a CS Senior Honors Thesis.