Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic

Authored by: Wei Zhou, Dong Chen, Jun Yan, Zhaojian Li, Huilin Yin and Wanchen Ge

Presented by: Kelvin Mock

March 2024

Carleton University

Agenda

1. Background

- Motivation
- Problem Statement
- Research Questions

2. Overview

- High Level Algorithm
- Related Works

3. Methodologies

Preliminaries of RL
MARL Problem for Lane-Changing
Comparable Benchmarking Models

4. Experiments

Use of HDV Models
Technical Setup
Results & Analysis
Simulation Outcome

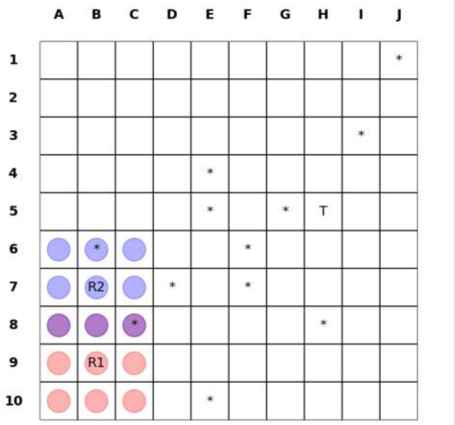
5. Conclusion

Background

How significant is this research? How is the problem formulated?

Motivation

Fatigue (e.g., long-haul driving)



Remember our Project?

Traffic Congestion

Lane Changing

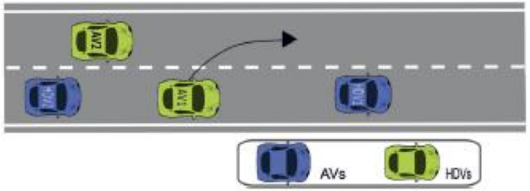


Figure 1 Illustration of the considered lane-changing scenario

Carleton University

Problem Statement

Everyday Challenge as a driver

- Lane Assist / Self-Driving (taking over your wheel) → Accurate? Safe?
- Mixed-Traffic Highway Environment

Challenges in the Research Domain

- Most studies focuses on **single agent** settings.
- Different driving behaviors
  - ❑ Another Autonomous Vehicle (AV)
  - ❑ Human – Aggressive vs Defensive

Research Questions

MARL Design

- How do the agents interact with each other and human-driven vehicles?
- How is the lane-changing problem formulated as a MARL task?
- How could this method be extended to real-world applications?

Algorithms

- How is an actor-critic mechanism applied in this problem?
- How does the parameter-sharing scheme improve MARL performance?
- How does the algorithm handle partial observability?

Overview

What is the goal of this research?

How are previous works supporting this research?

High Level Perspective of the Algorithm

- Multi-Agent Advantage Actor-Critic (MA2C) method
- Local Reward Design – safety, efficiency, passenger's comfort
- Parameter Sharing Scheme
- Mechanism:

1. Decentralized Cooperative in execution
2. Centralized Shared Critic in training

- With 3 Traffic Densities + 2 Driver's Behaviors

Related Works

Non-Data-Driven Methods

- Aim: To construct pre-defined a ruleset of virtual trajectory references:
- × Hard-coded rules are too naïve
  - ❑ Inter-vehicle traffic gaps 🤔
  - ❑ Time instances to perform maneuvers 🛞 ⏳
- × Dynamic models is a highly complicated algorithm
  - ❑ Optimization-based 👍
  - ❑ E.g., Quadratic Programming – specific traffic constraints
- × Unable to account for <u>stochastic driving behaviors</u> on the road

Related Works

Data-Driven Methods

- Model-Free RL
- Deep Deterministic Policy Gradient
- Safe RL framework – regret theory
- Temporal & Spatial Attention
- × Single-agent → corporation issue
- Priority-based Safety Supervisor
  - ❑ Hard-coded MARL Constraints
  - ❑ ↓ gradient estimation error

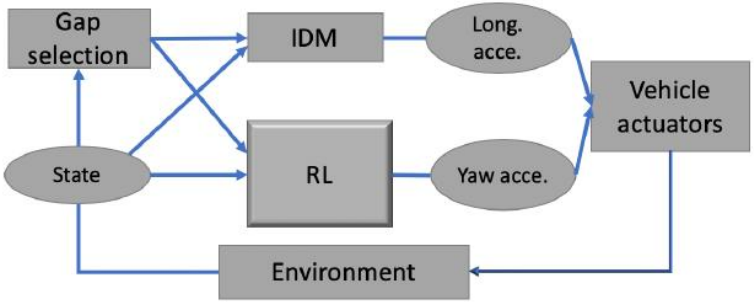


Figure 2: Vehicle system structure using DDPG [1]

DDPG

- Intelligent Driver's Model (IDM)
- Longitudinal Controller
- Acceleration
- Leader-following mechanism

----

[1] Pin Wang, Hanhan Li, Ching-Yao Chan (2019).

Carleton University

Related Works

A MARL-based Implementation

- Graphic CNN [2]
- Deep Q Network (DDQN)

✓3-lane freeway with 2 off-ramps

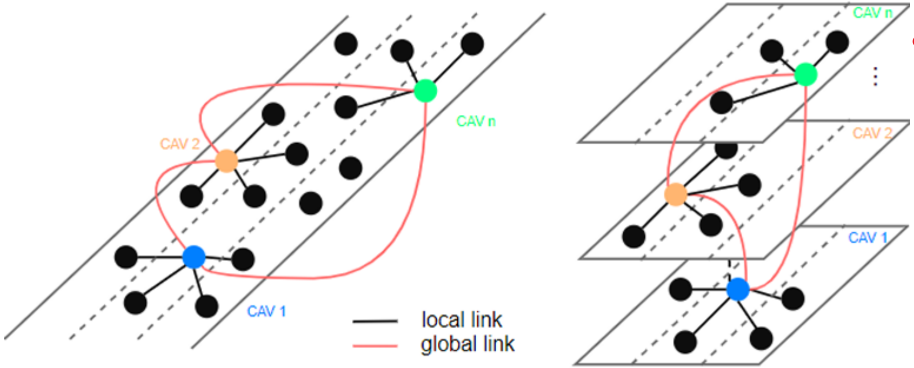


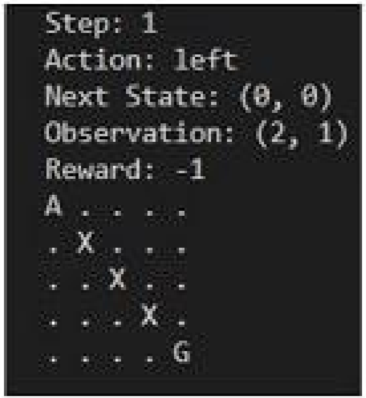
Figure 3: Graphic representation of Connected AV (CAV) network

----

[2] Jiqian Dong, Sikai Chen, Paul (Young Joun) Ha, Yujie Li, Samuel Labi (2020).

Applying to this Paper

- Multi-Objective Reward Function
- Parameter-Sharing Scheme
- Partial Observation
- Markov Decision Process (MDP)



Carleton University

Methodologies

How do we formulate the problem to achieve our goal?

Preliminaries of Reinforcement Learning

- Goal: Maximize Rewards with Partial Observability
- γ: Discount Factor ranging $(0,1]$
- T: Maximum number of steps per episode
- s_t: State at time t (= n-dimensional real-value **vector**)
- a_t: Action at time t (= m-dimensional vector, m = number of agents)
- r: Scalar Reward at time t

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k}$$

- Policy: a probability distribution over the Action Space (in a state)

Preliminaries of Reinforcement Learning

Model-Free RL methods

- Goal: Find an Optimal Q-function
- Action-Value function: $Q^{\pi}(s, a) = E[R_t \mid s = s_t, a]$
    - ❑ Choose an action and state → Evaluate an **expected return**
- State-Value function: $V^{\pi}(s_t) = E_{\pi}[R_t \mid s = s_t]$
    - ❑ Evaluate again an expected return w.r.t. <u>policy & state</u>
- Represented in a Neural Network $\pi_{\theta}(a_t|s_t)$
- Actor-Critic: diminishing gradient $\theta \leftarrow \theta + E_{\pi\theta}\left[\left(\nabla_{\theta}\log \pi\theta\ (a_t|s_t)\right)A_t\right]$
- Advantage function: to reduce sample variance
- Update State-Value function: minimize loss function

Formulating the MARL Problem of Lane-Changing

- Goal: Construct a **decentralized** approach with multi-agents

Discontinuous Evaluation

- State Space: $O_i: N_{N_i} \times F$ (in a matrix form, for agent i with F features)
  - ❑ Longitudinal Position (i.e., Distance between the vehicle ahead)
  - ❑ Lateral Speed
- Policy: $\pi_i: O_i \times S_i \rightarrow [0,1]$
- Action Space: {speed up, slow down, cruising, turn left, turn right}
- Reward Function:
  - ❑ Metrics: Safety, Headway evaluation, Speed Evaluation, Comfort
- Multi-Objective Reward: $r_{i,t} = w_s r_s + w_d r_d + \cdots$

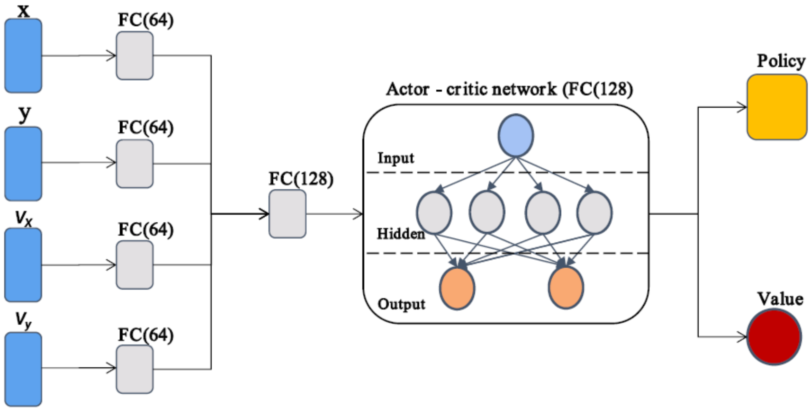
Carleton University

Formulating the MARL Problem of Lane-Changing

- Assumption: Weighted Total Reward
- **Safety** = $\begin{cases} 0 = safe \\ 1 = unsafe \end{cases}$
- **Headway Evaluation** = $\log\left(\frac{d_{headway}}{v_t t_d}\right)$
  - ❑ Thresholds: Velocity & Time
- **Speed Evaluation** = $\min\{\frac{v_t - v_{min}}{v_{max} - v_{min}}, 1\}$
  - ❑ Highway Situation: High Speed the better 😌
- **Driving Comfort** = $r_a + r_{lc}$ (Penalty Terms)
  - ❑ $r_a = \begin{cases} -1 \\ 0 \end{cases}$ = acceleration ; $r_{lc} = \begin{cases} -1 = change\ lane \\ 0 = keep\ in\ lane \end{cases}$

Formulating the MARL Problem of Lane-Changing

- Actor-Critic Network
  - ❑ Maximize global reward → **Scalability**
  - ❑ Local reward → **Credit Assignment**



Algorithm 1 MARL for AVs

Parameter: γ, η, p, T.
Output: θ.

```
Initialize o_0, t ← 0.
repeat
    for i ∈ V do
        Observe o_{i,t};
        Update a_{i,t} ~ π_{θ_{i,t}};
    end for
    Update t = t + 1;
    if DONE then
        for i ∈ V do
            Update θ_i ← θ_i + η∇_{θ_i} J(θ_i);
        end for
    end if
    if t = T then
        Initialize o_0, t ← 0;
    end if
until Stop condition is reached
```

Figures 4 & 5: The Architecture and the Pseudocode of the network

Carleton University

Comparable Benchmarking Models

	Multi-Agent Deep Q-Network	Multi-Agent Actor-Critic using Kronecker Factored Trust Region	Multi-Agent Proximal Optimal Optimization	MA2C (Ours)
Type	• Off-policy • Value-Based	• On-policy • Actor-Critic • Trust Region Optimization	• On-policy • Actor-Critic • Proximal Policy Optimization	• On-Policy • Actor-Critic ✅ Multi-Agent
Strength	• Sample Efficient ✅ Discrete Actions	• Stable Learning ✅ Efficient Updates	• Robust • Balances: (1) Stability (2) Exploration	• Local Reward • Param Sharing ✅ Scalable ✅ Co-operable
Weakness	❌ High variance ❌ Unstable for Multi-agent tasks	❌ Less sample efficient	❌ Slow (converge)	❌ Less sample efficient

Carleton University

Experiments

How does the series of concepts come into play?

Are those concepts proven valid and effective?

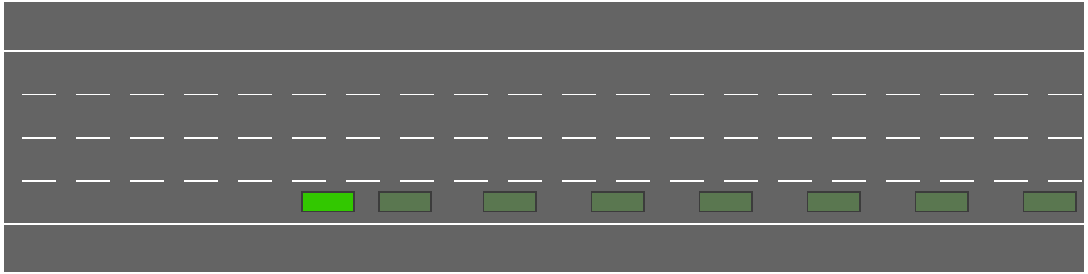
Carleton University

Use of HDV Models

- HDV = Human-Driven Vehicle Model
- Follows an **Intelligent Driver Model** – deterministic, continuous in time
- Car-following Model
    - ❑ Position
    - ❑ Speed & Acceleration
    - ❑ Distance between the vehicle ahead
    - ❑ Driving Habits
- Acceleration Problem: Minimize Overall Braking Induced by Lane Change

Experiment Setup

- OpenAI Gymnasium-based Simulation [3]



- Highway Road Length = 520 meters
- Randomly Spawn Vehicles on the Highway
  - Different Initial Speeds: 25 – 30 m/s (i.e., 56 mph – 67 mph)
- Vehicle Control Sampling Frequency (default 5Hz, ~ 0.2 seconds)

[2 https://github.com/Farama-Foundation/HighwayEnv

Experiment Setup

Training

	Training Parameters
Iterations	1 million steps / epochs
Random Seeds	x2 Random Seeds (Sharing the same seed among agents)
Discount Factor γ	0.99
Learning Rate η	5×10^{-4}
Weights	Safety = 200 Heading Distance = 4 Speed = 1

Evaluation

3 Traffic Density Modes

Traffic density modes	AVs	HDVs	Explanation
1	1-3	1-3	low level
2	2-4	2-4	middle level
3	4-6	4-6	high level

Computational Resources

- macOS server
- 2.7 GHz Intel Core I5 processor
- 8 GB Memory

Results & Analysis

- Local Reward designs outperforms Global Reward designs

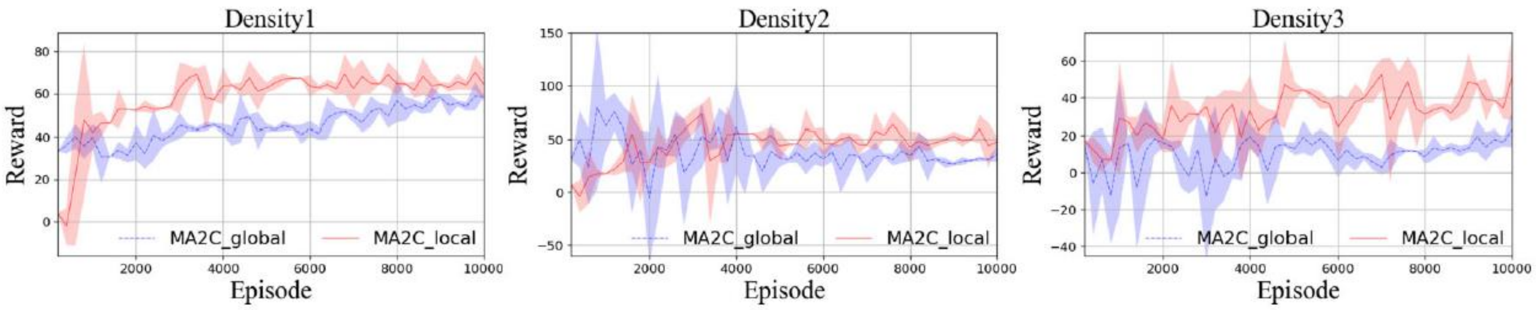


Figure 6: Performance comparisons between local and global reward designs

- ❑ Variance
- ❑ Credit Assignment issues

- That's why we need **decentralized execution** with **centralized training** benefits

Carleton University

Results & Analysis

- Sharing an Actor-Critic network is better than separating

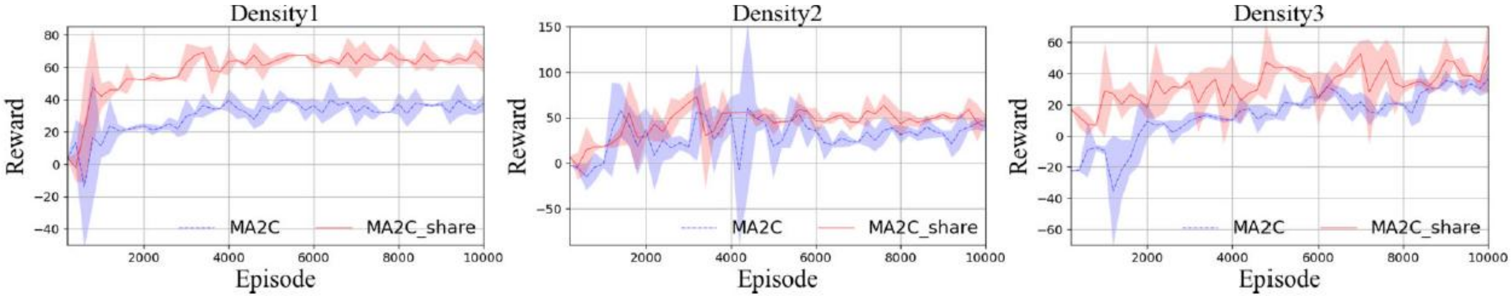


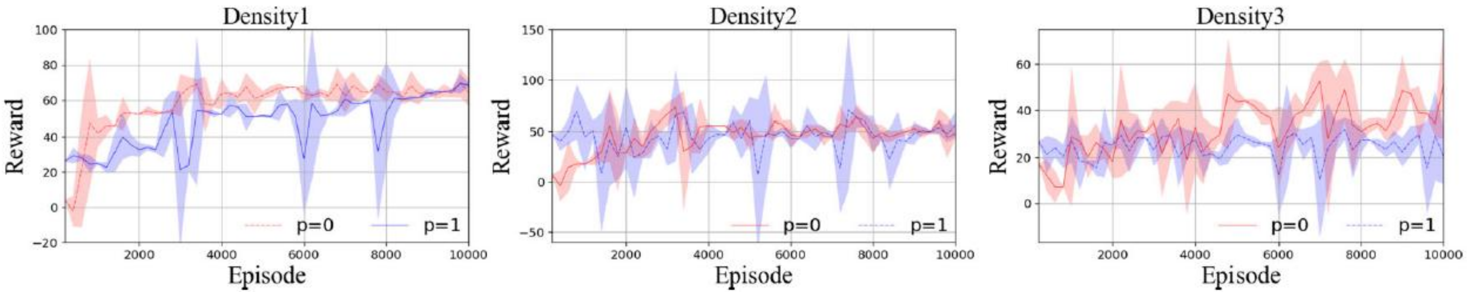
Figure 7: Performance comparisons between with and without actor-critic network sharing

  - ❑ Higher Rewards
  - ❑ Lower Variance
- Separating the network takes longer time to converge 🤔
- That's why we need a **centralized** network

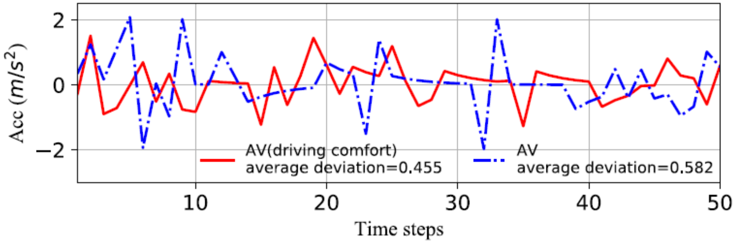
Carleton University

Results & Analysis

- Verified Effectiveness via Driving Comfort



- ✓ Low Variance
- ✓ Smoother (avg. deviation ~0.455 m/s^2)
- ✓ Scalable & Stable (regardless of HDV)



Figures 8 & 9: Performance comparisons of acceleration; Performance comparisons on different politeness coefficients p

Carleton University

Results & Analysis

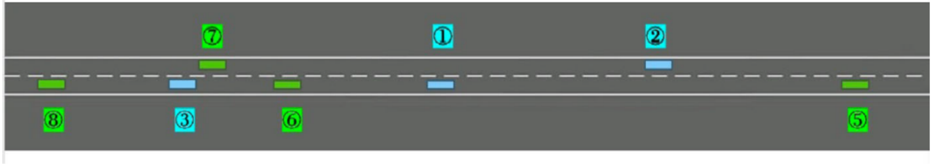
- Multi-Agent Deep-Q-Network (MADQN)
- Multi-Agent Actor-Critic using Kronecker-Factored Trust Region (MAACKTR)
- Multi-Agent Proximal Policy Optimization (MAPPO)
- MA2C – Our Approach

Method	Density 1	Density 2	Density 3
MADQN	47.451 (±27.948)	51.568 (±32.943)	48.509 (±24.078)
MA2C	58.000 (±9.308)	44.744 (±10.895)	32.579 (±8.160)
MAACKTR	8.812 (±6.217)	3.759 (±10.858)	4.892 (±10.986)
MAPPO	31.988 (±6.567)	19.300 (±16.097)	5.073 (±19.762)

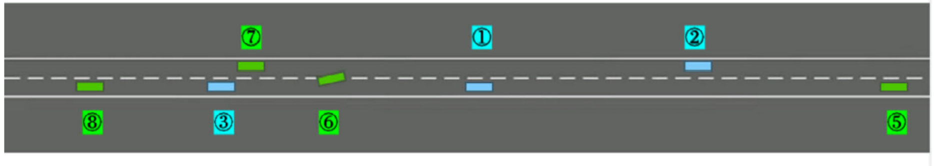
Figures 10: Mean episode reward in different traffic flow scenario

Carleton University

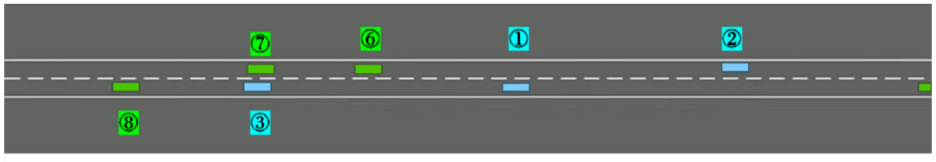
Simulation Outcome



(a) initial state



(b) changing lanes

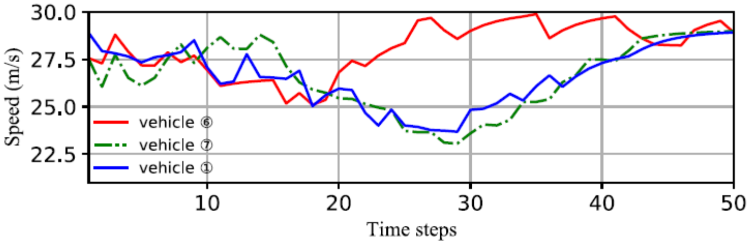


(c) lane change completed

Figures 11: Lane change in simulation environnent

Cooperative Reasonably

- Slow down to make space
- Speed up while merging
- Picks up speed after merging



Figures 12: Speeds of the AVs

Carleton University

Conclusion

- Developed an on-policy RL framework in a mixed-traffic environment
- Extended Actor-Critic into Multi-Agent settings
- Proven Efficiency of a local reward design + parameter sharing
- Compared with a compromising set of benchmarking models
- Compared with a convincing set of metrics
  - ❑ Driving Efficiency
  - ❑ Driving Comfort
  - ❑ Safety – ensuring no collisions
- The proposed MA2C method outperforms!!!
- Extension: Our Project 😍

References

- Wei Zhou, Dong Chen, Jun Yan, Zhaojian Li, Huilin Yin and Wanchen Ge (2022). Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. Zhou et al. Autonomous Intelligent Systems (2:5). Available at: https://link.springer.com/article/10.1007/s43684-022-00023-5.
- Pin Wang, Hanhan Li, Ching-Yao Chan (2019). Continuous Control for Automated Lane Change Behavior Based on Deep Deterministic Policy Gradient Algorithm. 30th IEEE Intelligent Vehicles Symposium (IV).
- Partially Observable Markov Decision Process (POMDP) in AI (2024). Geeksforgeeks. Available at: https://www.geeksforgeeks.org/partially-observable-markov-decision-process-pomdp-in-ai/.

Carleton University

References

- Jack Sackman (n.d.). 10 Tips For Avoiding Driver Fatigue. Available at https://auto.howstuffworks.com/10-tips-for-avoiding-driver-fatigue.htm.
- Consulting.ca (2018). CAA recommends eight ways to reduce traffic congestion in Canada. Available at https://www.consulting.ca/news/128/caa-recommends-eight-ways-to-reduce-traffic-congestion-in-canada.
- Leurent, Edouard (2018). An Environment for Autonomous Driving Decision-Making. GitHub. Available at https://github.com/Farama-Foundation/HighwayEnv.

Thank You

Any Questions?

Carleton University