Development of explainable reinforcement learning approaches for safe autonomous driving

Shahin Atakishiyev

PhD student in Computing Science

Explainable AI (XAI) Lab

University of Alberta

August 23, 2022

Shahin Atakishiyev

Dr. Mohammad Salameh (Huawei)

Prof. Randy Goebel

UNIVERSITY OF ALBERTA

xAI

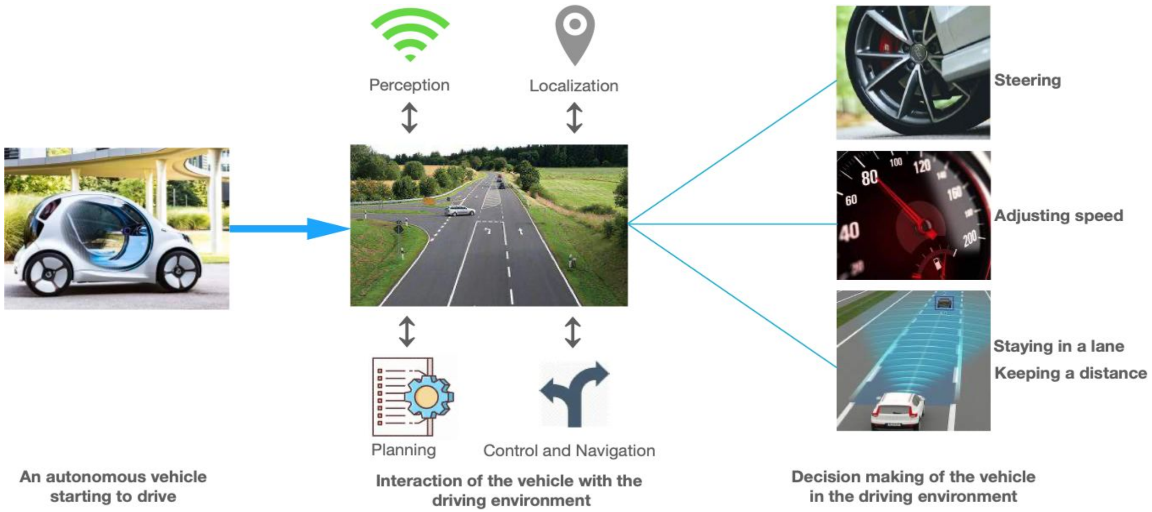
amii

Agenda

- Autonomous driving at a glance
- The need for explainability of AI in autonomous driving
- Reinforcement learning for autonomous driving
- Explainable reinforcement learning for autonomous driving
- Ongoing work
- Conclusions

Autonomous driving at a glance

- **Decision making in autonomous driving (AD):**



Atakishiyev et al, 2021. arXiv

Autonomous driving at a glance

- **Five levels of state-of-the-art AD as defined by the SAE International (SAE International, 2018):**

- **Level 0: No Driving Automation**

- **Level 1: Driver Assistance**
  - ✔ Adaptive cruise control

- **Level 2: Partial Driving Automation**
  - ✔ Advanced Driving Assistance Systems (ADAS)

- **Level 3: Conditional Driving Automation**
  - ✔ Object/Obstacle detection

- **Level 4: High Driving Automation**
  - ✔ Near-full automation, in a geofenced area (Alphabet's Waymo)

- **Level 5: Full Driving Automation**
  - ✔ Everything is automated, no human supervision is required

The need for explainability of AI in autonomous driving

- ***Three* primary reasons of the need for explainability of AI in AD (*Atakishiyev et al., 2021*):**

- **Psychological lens: Traffic accidents and the safety concerns with the presence of autonomous vehicles.**

Tesla sedan hits parked police vehicle in Laguna Beach, California, in 2018

CBC, 2020

The need for explainability of AI in autonomous driving

- **Sociotechnical lens:** **Design, development, and deployment of autonomous vehicles should be human-centered by**

  1. **reflecting the users' needs,**
  2. **take their prior opinions and expectations into account.**

Mercedes-Benz Group Media, 2017

UNIVERSITY OF ALBERTA

xAI

amii

The need for explainability of AI in autonomous driving

- **Philosophical lens:** **Explainable AI (XAI) decisions can provide descriptive information about the causality of real-time actions.**

- **Also, General Data Protection Regulation (GDPR): “Right of explanation ” for end users**

Atakishiyev et al., 2021

The need for explainability of AI in autonomous driving

Canonical example: "The Molly problem"

A young girl called Molly is crossing the road alone and is hit by an unoccupied self-driving vehicle. There are no eye-witnesses. What should happen next?

https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/MollyProblem.aspx

The need for explainability of AI in autonomous driving

Thus, XAI for autonomous driving is a compendium of AI-driven approaches

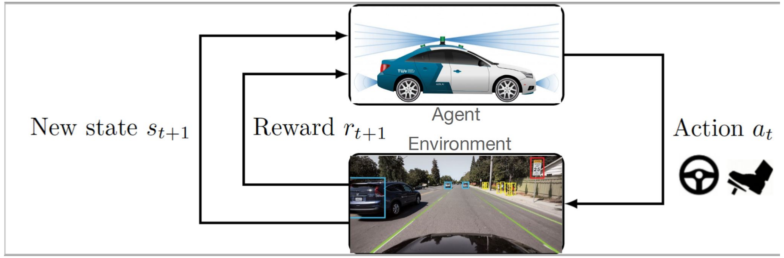
1) ensuring an acceptable level of safety in a vehicle's real-time decisions,

2) providing explanations and transparency on an automated car's decisions in critical traffic scenarios, and

3) obeying all traffic rules established by the regulators."

Reinforcement learning for autonomous driving

- **Supervised learning methods are not effective in AD, except scene understanding.**
- **Decisions are temporal and sequential.**
- **Thus, the goal becomes to solve sequential decision-making problems: We need reinforcement learning (RL).**



Atakishiyev et al., 2021

Explainable reinforcement learning for autonomous driving

Given that we necessitate

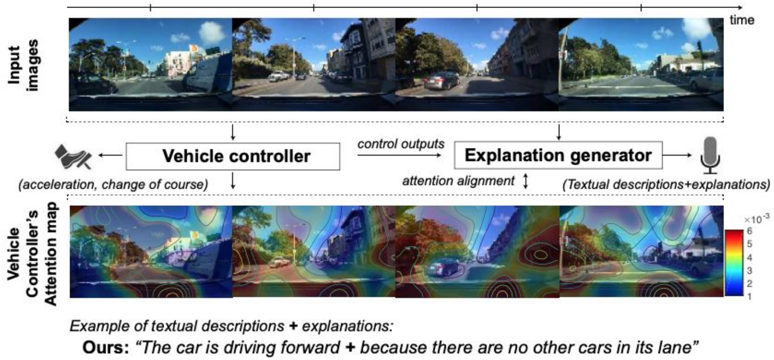
- **Sequential decision making**
- **Explainability**

We need to develop explainable RL (XRL) methods for autonomous driving.

Explainable reinforcement learning for autonomous driving

How? Three potential approaches:

- **Textual explanations:** Describing an agent's decisions linguistically in a natural language

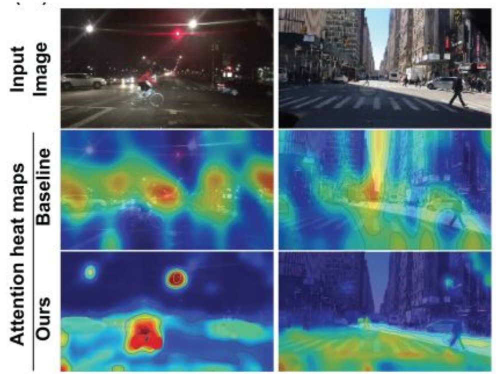


Kim et al., 2018. *ECCV*

Explainable reinforcement learning for autonomous driving

How?

- **Visual explanations:** Providing saliency maps (heatmaps) that topographically highlight areas which are crucial for an agent's decision.



Kim et al., 2021.

Explainable reinforcement learning for autonomous driving

How?

- Policy-level explanations: Developing interpretable RL policies, such as by summarizing transitions.

Ongoing work

Developing an end-to-end explainable RL framework on the CARLA simulator that powers safe actions for a car and provides justifications on these actions

CARLA: Open-source simulator for autonomous driving research

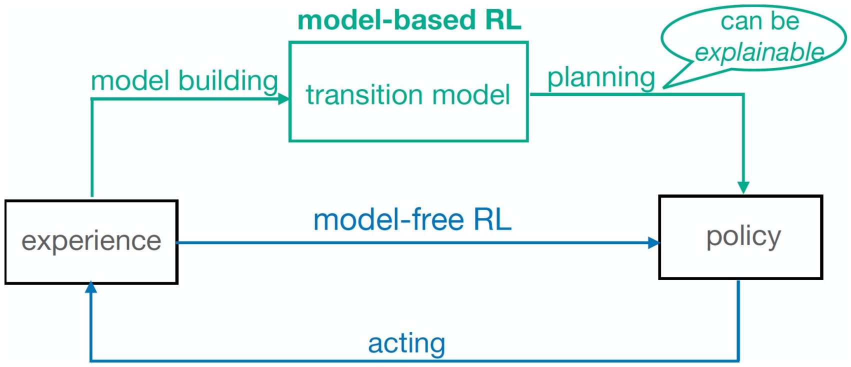
https://carla.org/

Video clip from: Zhang, 2021

Ongoing work: Problems being explored

1. Investigating model-free vs model-based approaches in terms of explainability

- **Model-free RL lacks explainability of learning**
- **Model-based RL has a planning component, which can be explainable**



Atakishiyev et al., 2021

Ongoing work: Problems being explored

2. Investigating safety-explainability dilemma: Does explainability affect the safety performance of an RL algorithm negatively?

- **Implementing model-free approaches, such as DQN, DDPG, PPO, and SAC**
- **Implementing model-based RL approaches such as Dyna**
- **Make a comparative analysis between model-free and model-based approaches in terms of safety**

Ongoing work: Problems being explored

3. Once having safest decision-making RL algorithm(s) for a car, developing explanations on top of them using

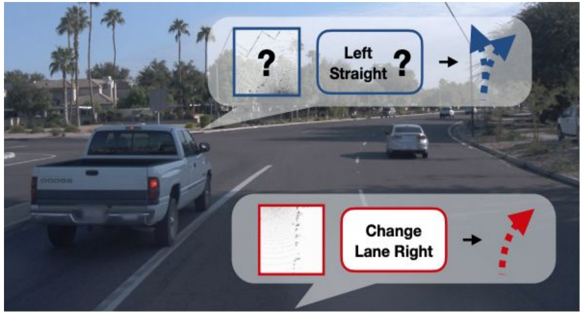
- **Visual techniques:** Developing interpretable saliency maps that provide visual rationales behind temporal actions
- **Natural language:** We propose to use visual question-answering (VQA) for this purpose
- **Interpretable policies:** Possibly by summarizing transitions

Ongoing work: Problems being explored

VQA: How can it be used for explanatory purposes in autonomous driving?

1. **Why (Causal) questions:**

   Q: <u>Why</u> was the lane changed to the right?
   A: <u>Because</u> the current lane narrows down ahead.



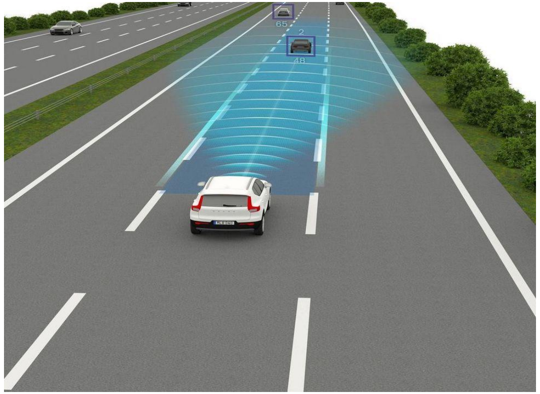
Chen 2022, *CVPR*

Ongoing work: Problems being explored

VQA: How can it be used for explanatory purposes in autonomous driving?

2. Descriptive questions:

Q: How many cars are around?
A: Two.



https://www.autoweek.com/news/technology/a35492454/levels-of-autonomous-driving-explained/

UNIVERSITY OF ALBERTA

xAI

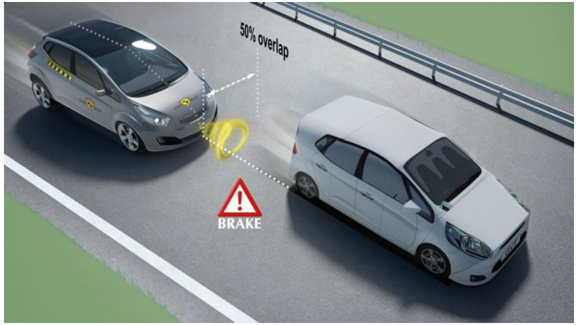
amii

Ongoing work: Problems being explored

VQA: How can it be used for explanatory purposes in autonomous driving?

3. **Temporal questions:**

   Q: What was the speed <u>before</u> the emergency brake?
   A: 50 km/h.



https://www.euroncap.com/en/vehicle-safety/the-ratings-explained/safety-assist/aeb-car-to-car/

Conclusions

- We aim to develop a safe and end-to-end explainable reinforcement learning framework for autonomous vehicles.
- The proposed framework combines reinforcement learning, computer vision, and natural language processing methods for explainable decisions.
- With safe and transparent AI, we can move a step closer to publicly approved and environmentally friendly intelligent vehicles in the near future.

Thank you very much for your attention!