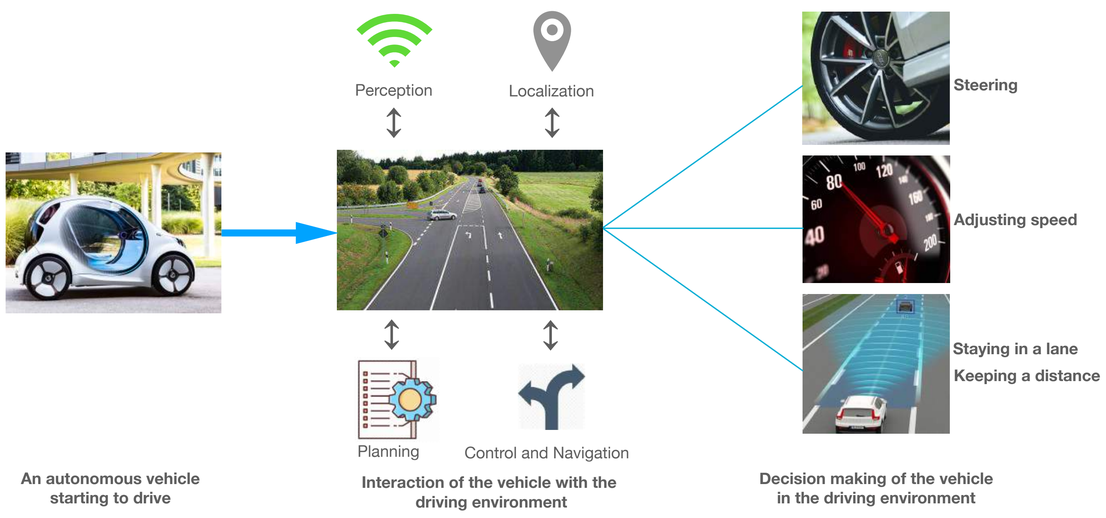
I drive, I explain, you trust: Development of explainable reinforcement learning approaches for safe and interpretable autonomous driving

Shahin Atakishiyev[1], Mohammad Salameh[2], Randy Goebel[1]

[1]Department of Computing Science, University of Alberta, Edmonton, Canada, [2]Huawei Technologies Canada Co., Ltd., Edmonton, Canada

Autonomous driving at a glance

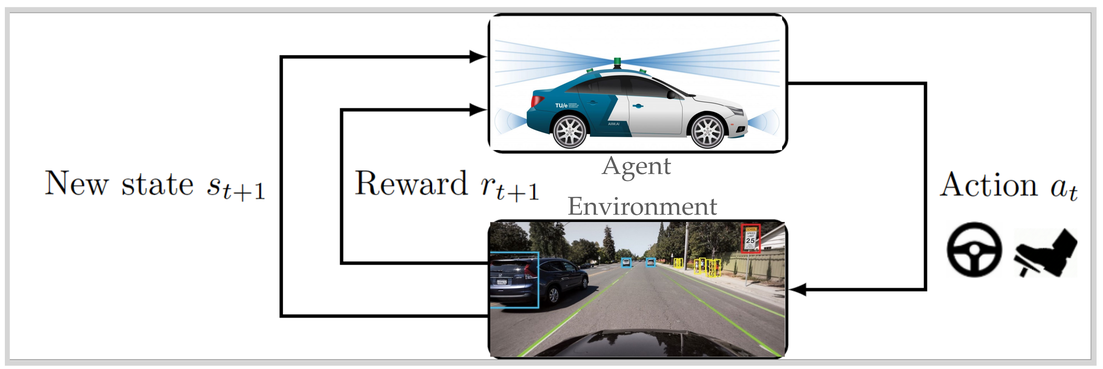


The need for explainability of AI in autonomous driving

- Traffic accidents and safety concerns with autonomous vehicles
- The need to understand causality of critical driving decisions
- European Union GDPR - Emphasis on ''Right of explanation"

Reinforcement learning for autonomous driving

- Classical supervised learning: Not effective in autonomous driving, except environmental perception
- Decisions are temporal and sequential
- The goal then becomes to solve a sequential decision-making problem: We need reinforcement learning (RL)



Explainable RL for autonomous driving

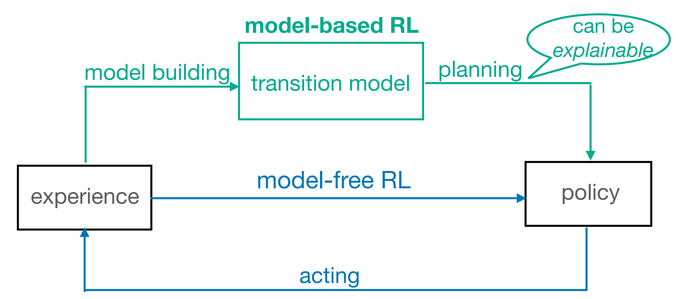
- **Sequential decision making**
- **The need for explainability**

We need to develop explainable RL for autonomous driving.

Ongoing work: How to develop explainable RL for autonomous driving?

Approach 1: Investigating model-free vs. model-based RL from an explainability perspective

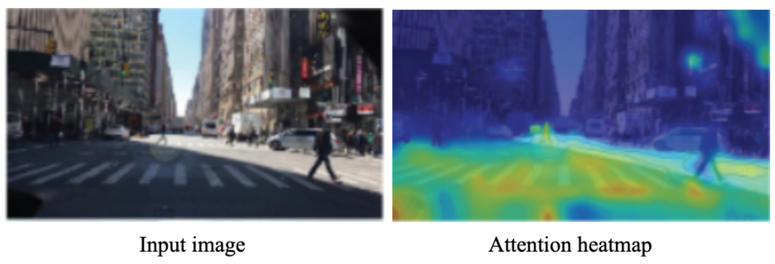
- **Model-free RL: Not explainable by its nature**
- **Model-based RL: Potentially explainable because of planning**



Approach 2: Developing an explanatory functionality using *NLP* and *vision* methods on the empirically safest RL algorithms

An example of a textual explanation:

An example of a visual explanation:



Input image | Attention heatmap

Conclusion

In our ongoing study, we aim to develop explainable RL approaches for safe and interpretable autonomous driving.

References

[1] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel. Towards Safe, Explainable, and Regulated Autonomous Driving. *AAAI-TRASE 2022*

[2] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel. Explainable artificial intelligence for autonomous driving: An overview and guide for future research directions. *Under review*