A HUMAN-CENTRIC METHOD FOR GENERATING CAUSAL EXPLANATIONS IN NATURAL LANGUAGE FOR AUTONOMOUS VEHICLE MOTION PLANNING

Balint Gyevnar, Massimiliano Tamborski, Cheng Wang,
Chris Lucas, Shay Cohen, Stefano Albrecht

In 2nd IJCAI Workshop on Artificial Intelligence for Autonomous Driving, 2022

OUR GOALS

A HUMAN-CENTRIC METHOD FOR GENERATING CAUSAL EXPLANATIONS IN NATURAL LANGUAGE FOR AUTONOMOUS VEHICLE MOTION PLANNING

OUR GOALS

A HUMAN-CENTRIC METHOD FOR GENERATING CAUSAL EXPLANATIONS IN NATURAL LANGUAGE FOR AUTONOMOUS VEHICLE MOTION PLANNING

Explore the role of explanations for passengers travelling with autonomous vehicles.

OUR GOALS

A HUMAN-CENTRIC METHOD FOR GENERATING CAUSAL EXPLANATIONS IN NATURAL LANGUAGE FOR AUTONOMOUS VEHICLE MOTION PLANNING

Explore the role of explanations for passengers travelling with autonomous vehicles.

Create a method to automatically generate explanations for autonomous vehicles which are beneficial for passengers.

AN ISSUE OF TRUST AND KNOWLEDGE

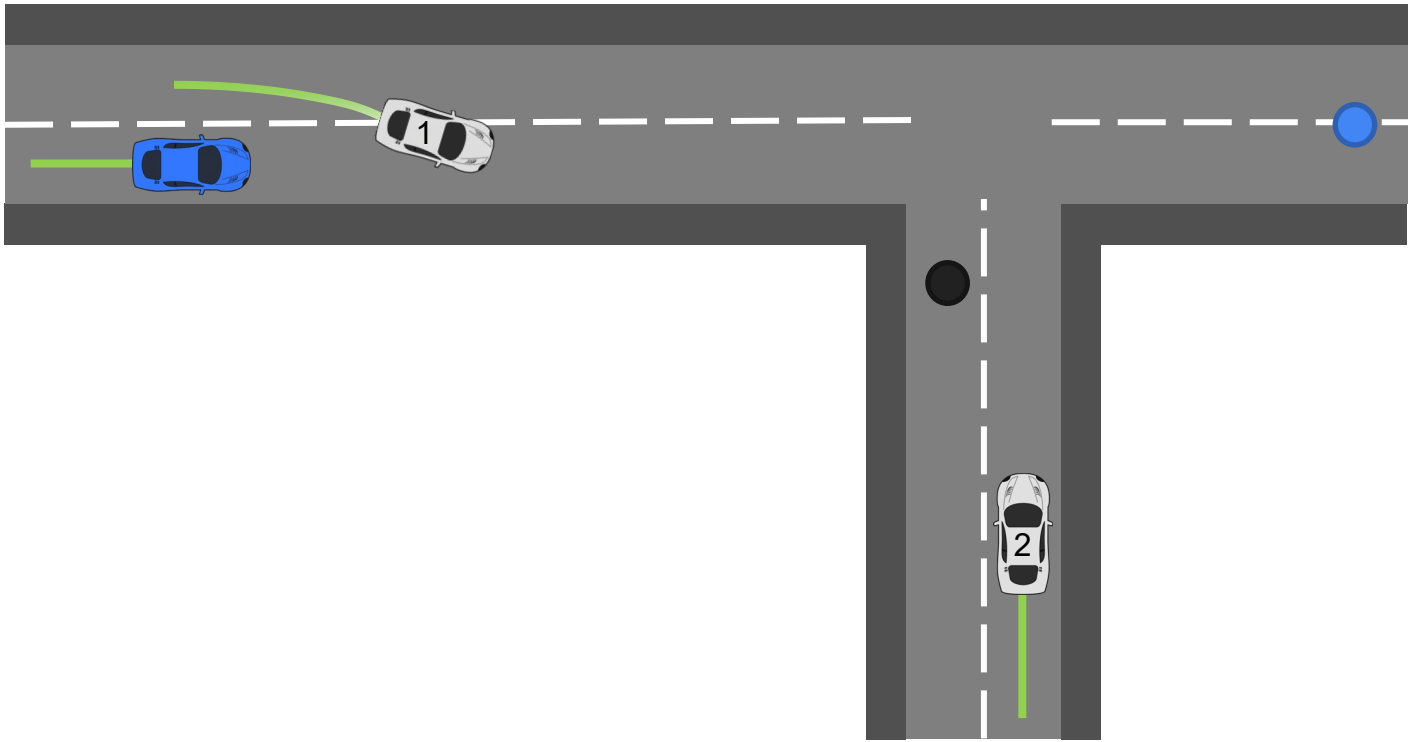
AN ISSUE OF TRUST AND KNOWLEDGE

AN ISSUE OF TRUST AND KNOWLEDGE

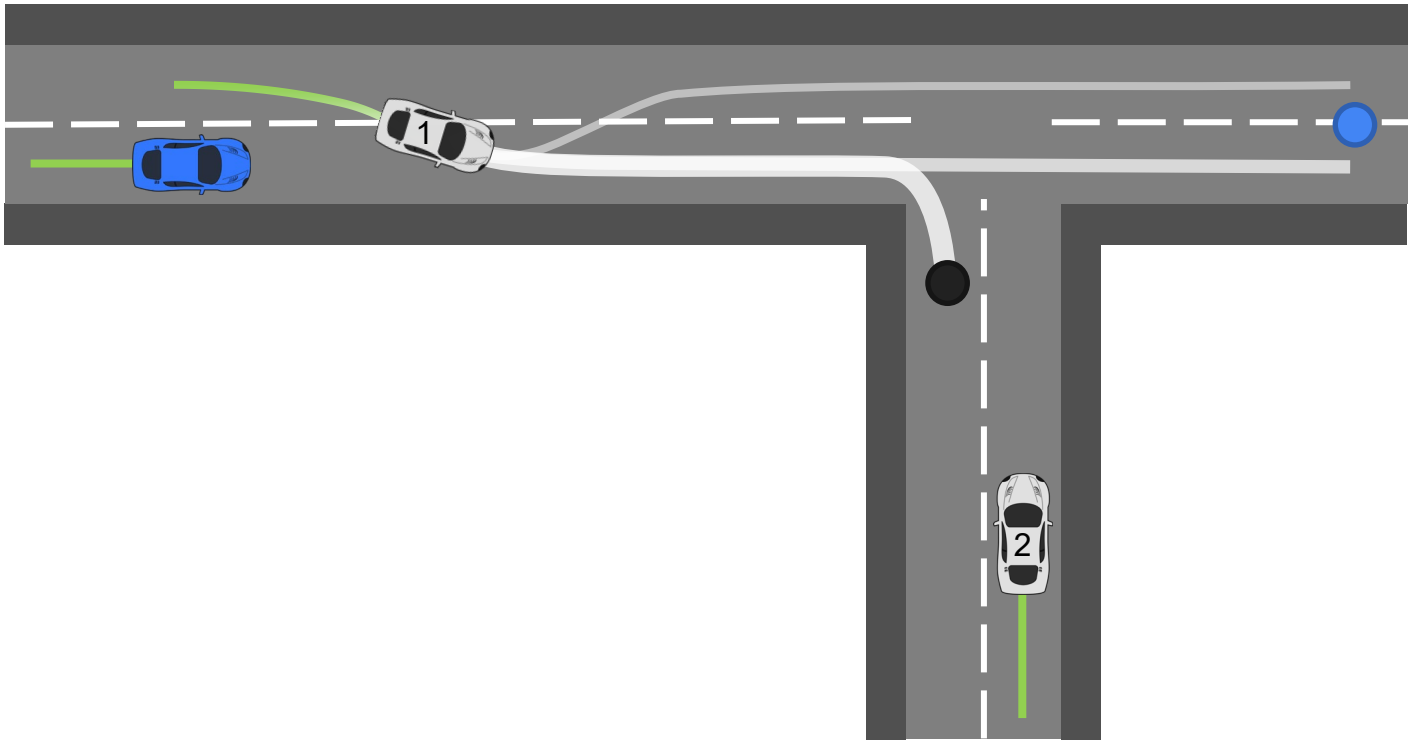
~~TRUST~~

~~KNOWLEDGE~~

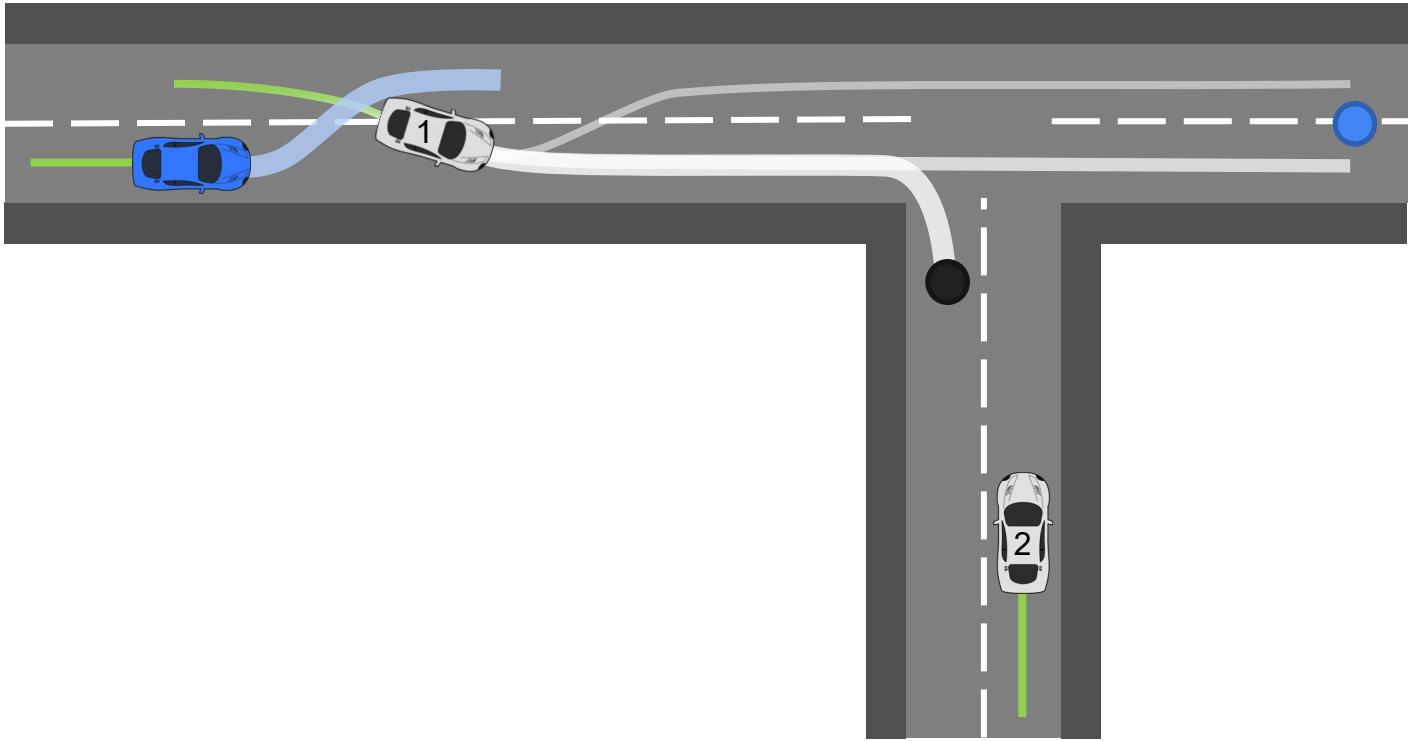
DEMONSTRATION



DEMONSTRATION

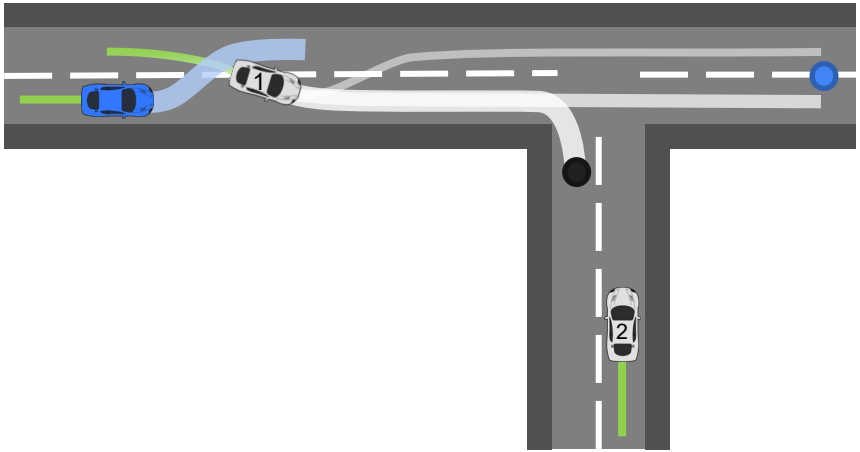


DEMONSTRATION



DEMONSTRATION

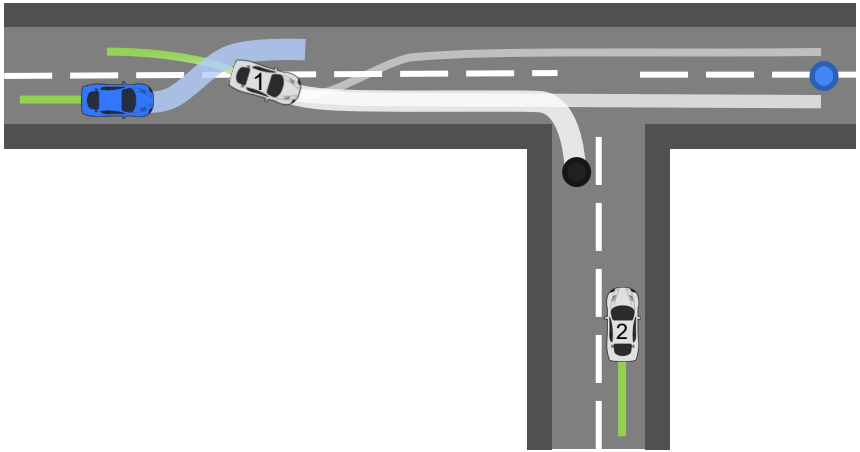
Q: Why did you change lanes instead of continuing straight?



DEMONSTRATION

Q: Why did you change lanes instead of continuing straight?

A: If I had continued straight then I would have reached the goal slower, because Vehicle 1 is slowing down, likely to turn right.



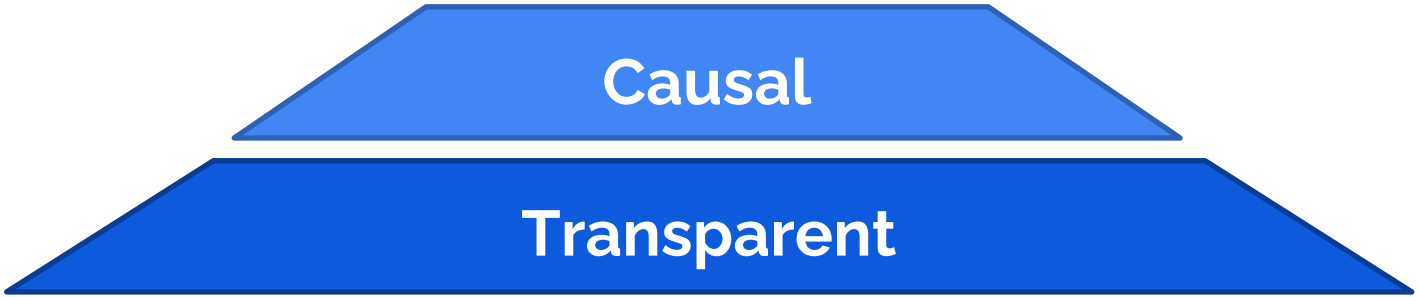
HUMAN-CENTRIC XAI

HUMAN-CENTRIC XAI

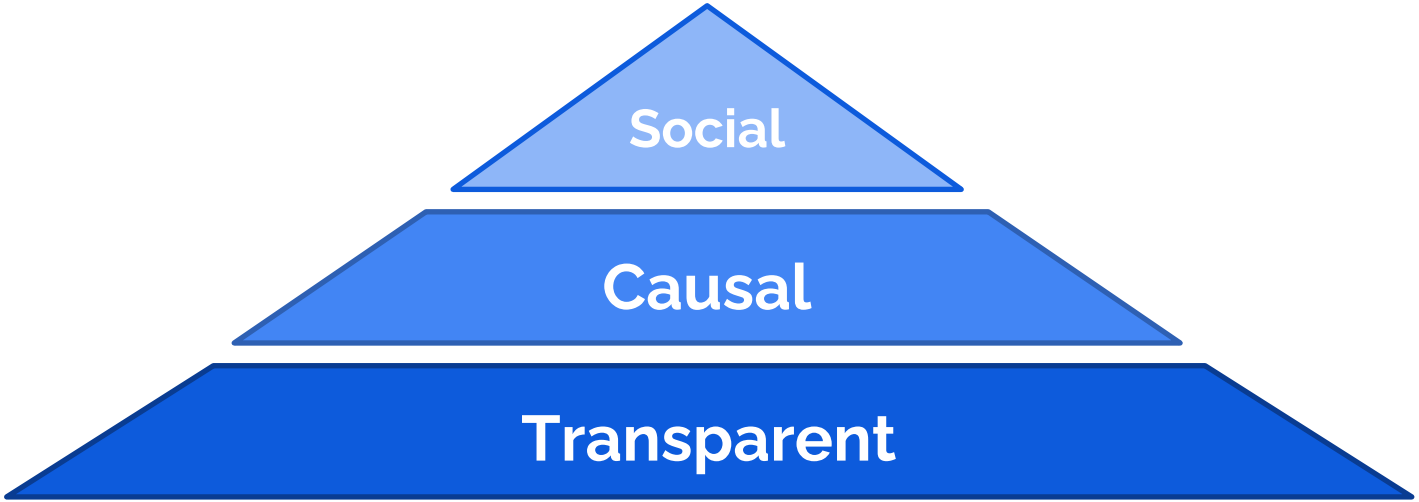
Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267:1-38, 2019.

HUMAN-CENTRIC XAI

HUMAN-CENTRIC XAI



HUMAN-CENTRIC XAI



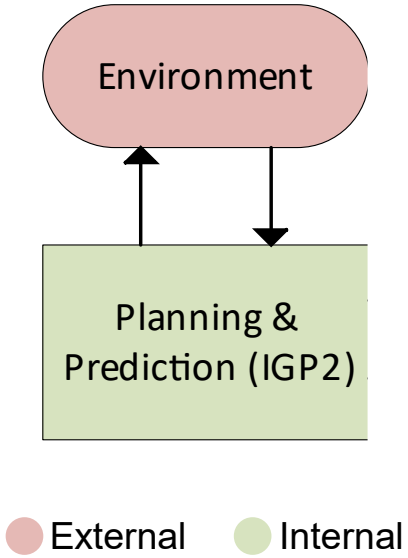
EXPLAINABLE AUTONOMOUS VEHICLE INTELLIGENCE (XAVI)

"Cars that Explain: Building Trust in Autonomous Vehicles through Explanations and Conversations"

Third Prize in "Shape the Future of ITS" Competition by IEEE Intelligent Transportation Systems Society, 2022

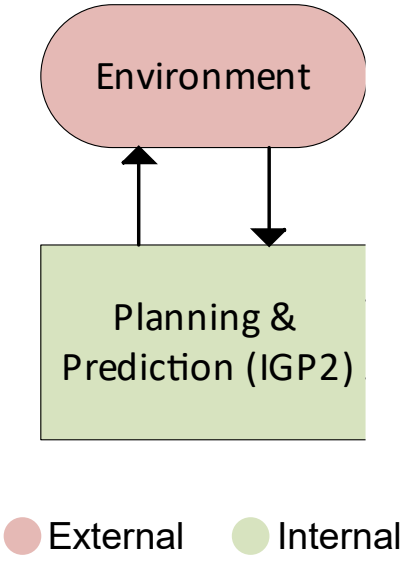
Blog post: https://agents.inf.ed.ac.uk/blog/explainable-autonomous-vehicle-intelligence/

EXPLAINABLE AUTONOMOUS VEHICLE INTELLIGENCE (XAVI)



IGP2: Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. *Interpretable Goal-based Prediction and Planning for Autonomous Driving.* IEEE International Conference on Robotics and Automation (ICRA), 2021.

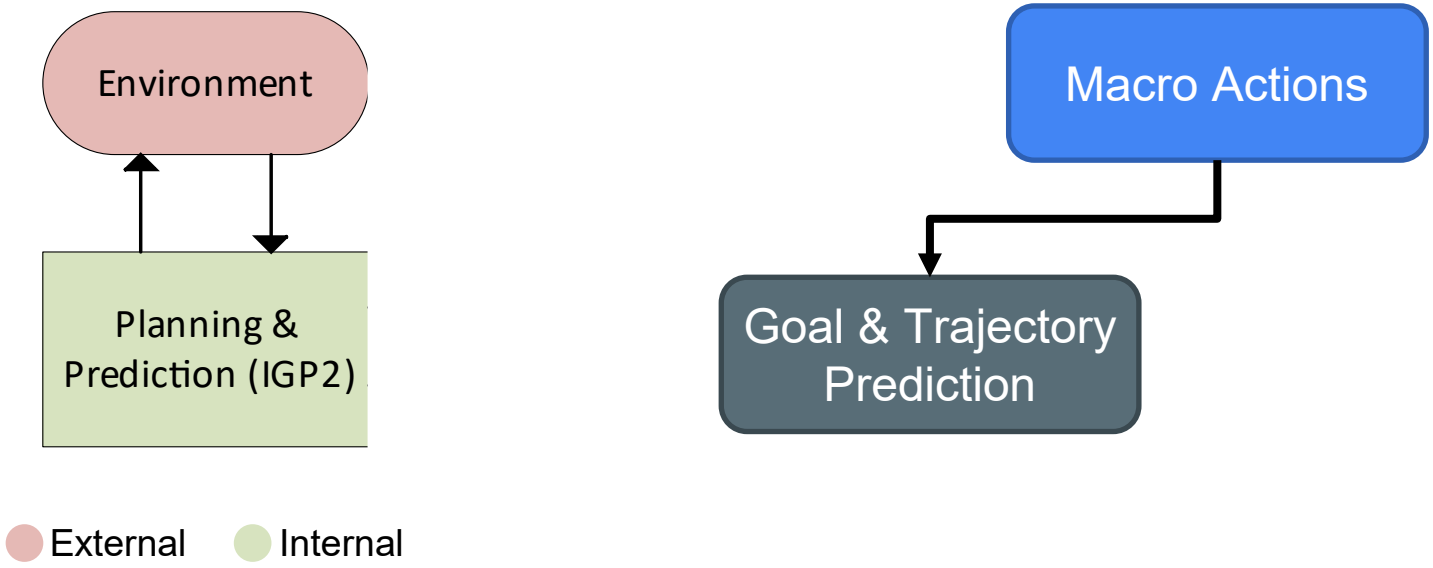
EXPLAINABLE AUTONOMOUS VEHICLE INTELLIGENCE (XAVI)



Macro Actions

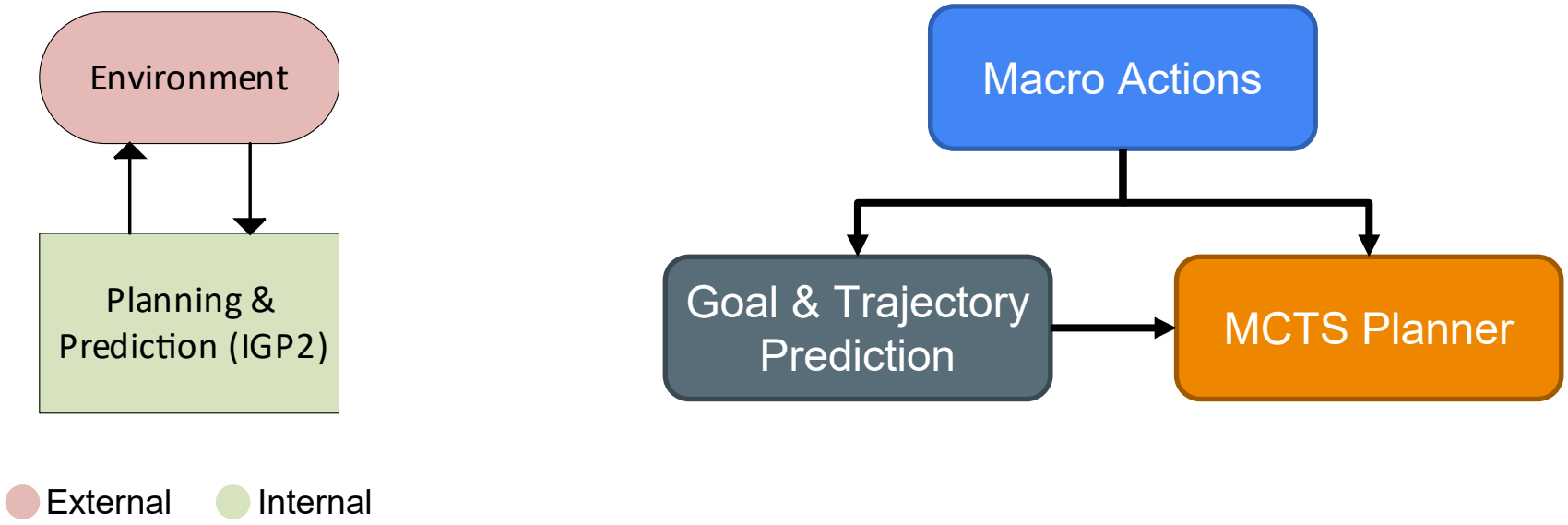
IGP2: Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. *Interpretable Goal-based Prediction and Planning for Autonomous Driving.* IEEE International Conference on Robotics and Automation (ICRA), 2021.

EXPLAINABLE AUTONOMOUS VEHICLE INTELLIGENCE (XAVI)



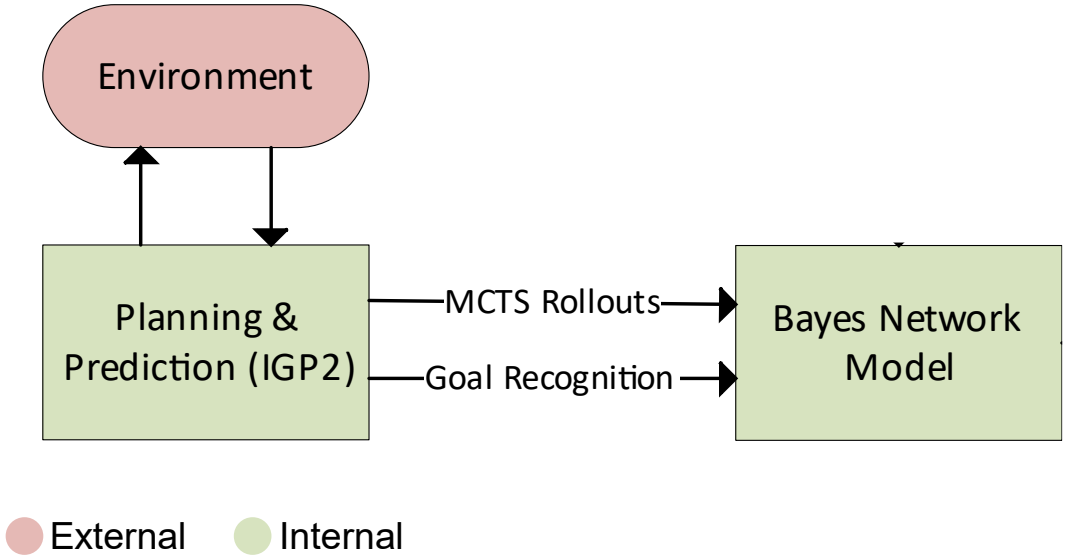
IGP2: Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. *Interpretable Goal-based Prediction and Planning for Autonomous Driving.* IEEE International Conference on Robotics and Automation (ICRA), 2021.

EXPLAINABLE AUTONOMOUS VEHICLE INTELLIGENCE (XAVI)



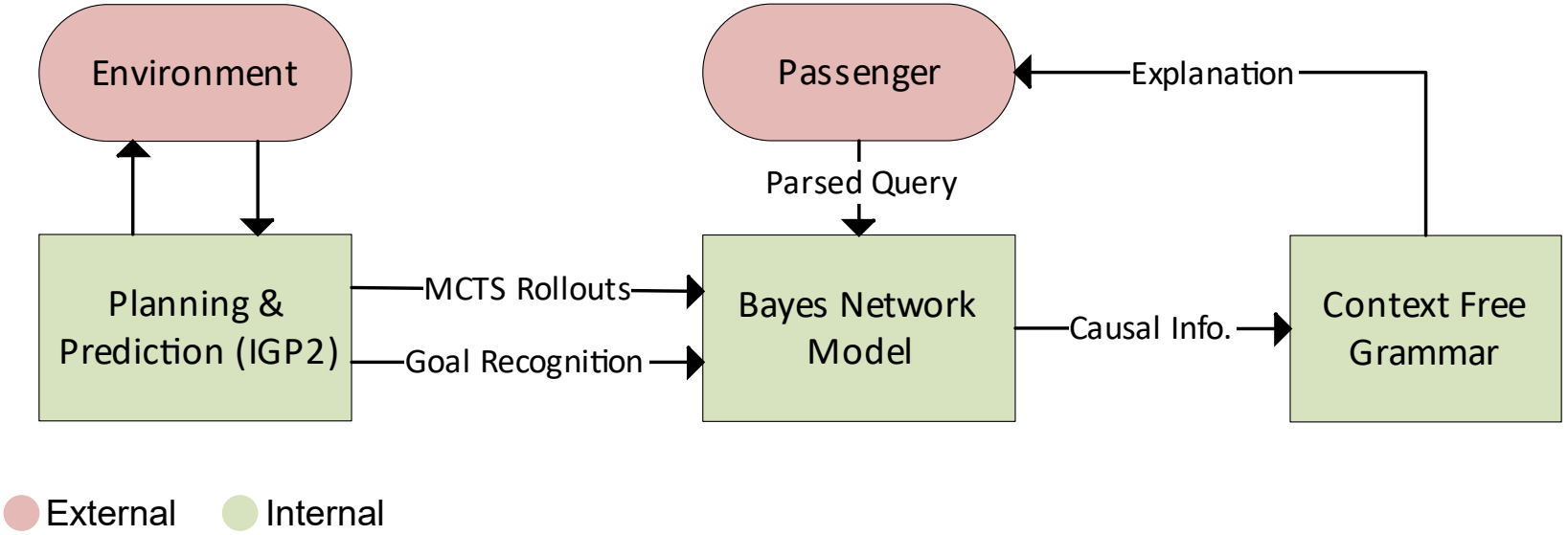
IGP2: Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. *Interpretable Goal-based Prediction and Planning for Autonomous Driving.* IEEE International Conference on Robotics and Automation (ICRA), 2021.

EXPLAINABLE AUTONOMOUS VEHICLE INTELLIGENCE (XAVI)



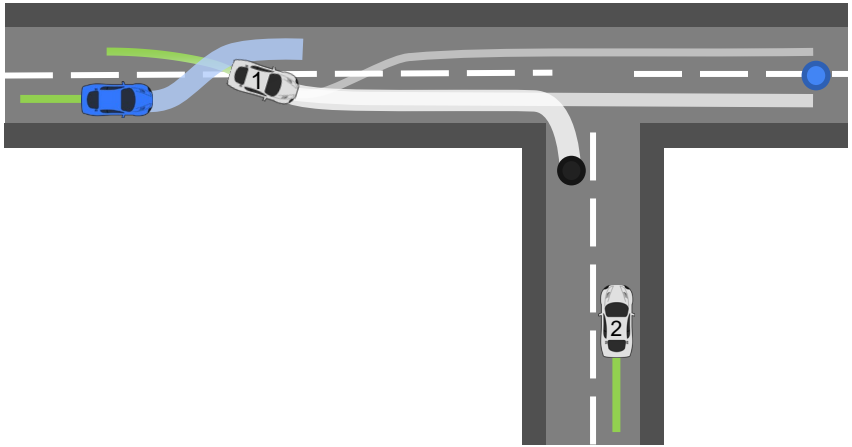
IGP2: Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. *Interpretable Goal-based Prediction and Planning for Autonomous Driving.* IEEE International Conference on Robotics and Automation (ICRA), 2021.

EXPLAINABLE AUTONOMOUS VEHICLE INTELLIGENCE (XAVI)



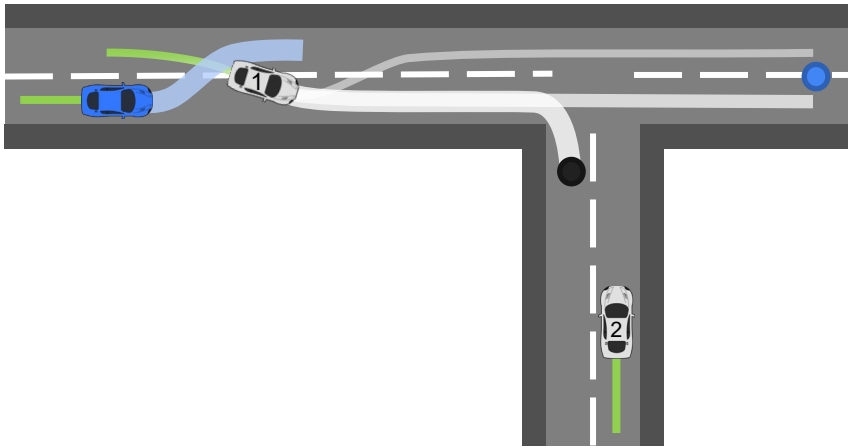
IGP2: Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. *Interpretable Goal-based Prediction and Planning for Autonomous Driving.* IEEE International Conference on Robotics and Automation (ICRA), 2021.

PRELIMINARY EVALUATION



SCENARIO 1

PRELIMINARY EVALUATION

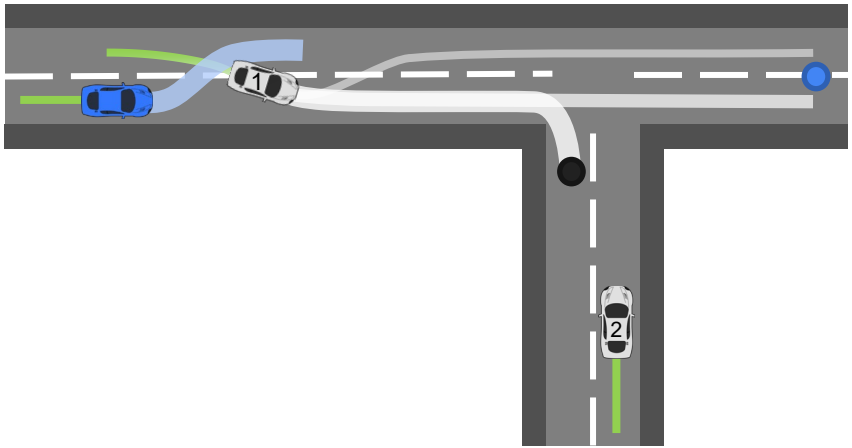


SCENARIO 1

Factual: Change lane left
Counterfactual: Continue straight

"If we had continued straight, we would have…

PRELIMINARY EVALUATION



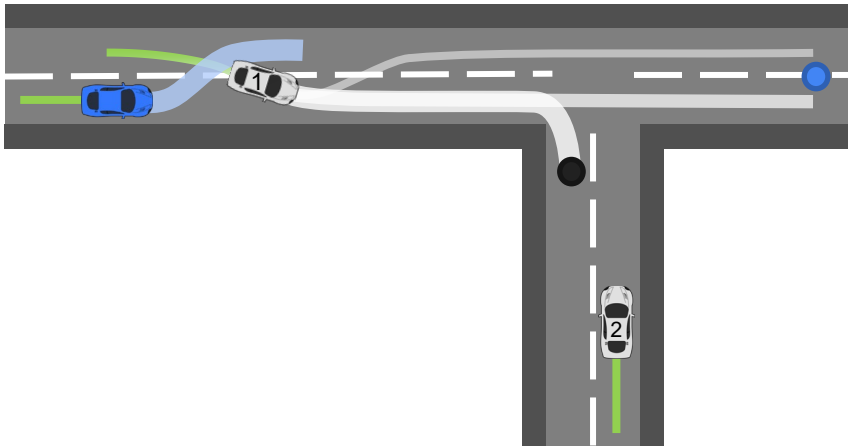
SCENARIO 1

Factual: Change lane left
Counterfactual: Continue straight

"If we had continued straight, we would have…

*reached the goal **slower**, because vehicle 1 likely changes right then exits right.*

PRELIMINARY EVALUATION



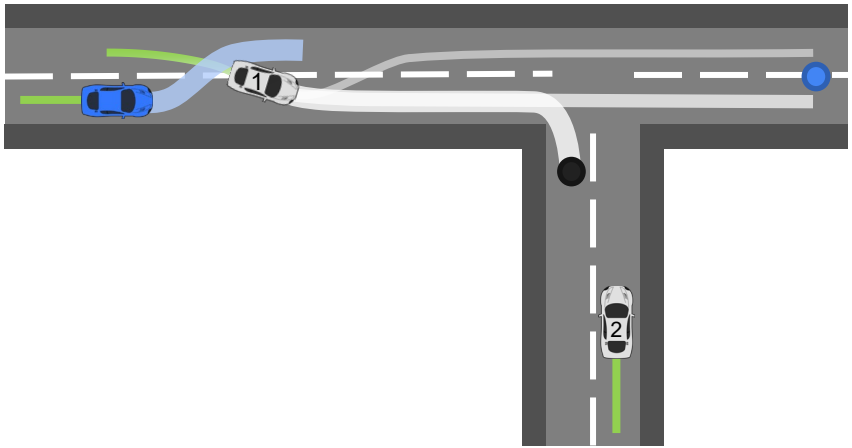
SCENARIO 1

Factual: Change lane left
Counterfactual: Continue straight

"If we had continued straight, we would have…

*reached the goal **slower** and with **more jerk** and with **less angular** velocity, because vehicle 1 probably changes right then exits right.*

PRELIMINARY EVALUATION



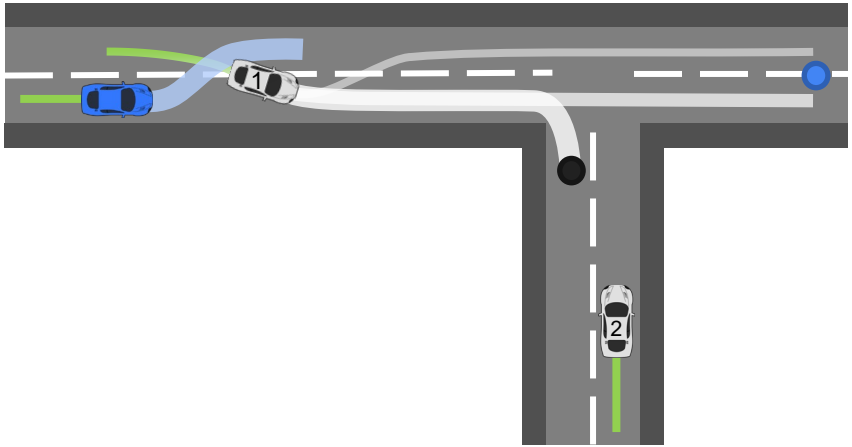
SCENARIO 1

Factual: Change lane left
Counterfactual: Continue straight

"If we had continued straight, we would have…

*likely **collided** with vehicle 1 because vehicle 1 probably changes right then exits right.*

PRELIMINARY EVALUATION

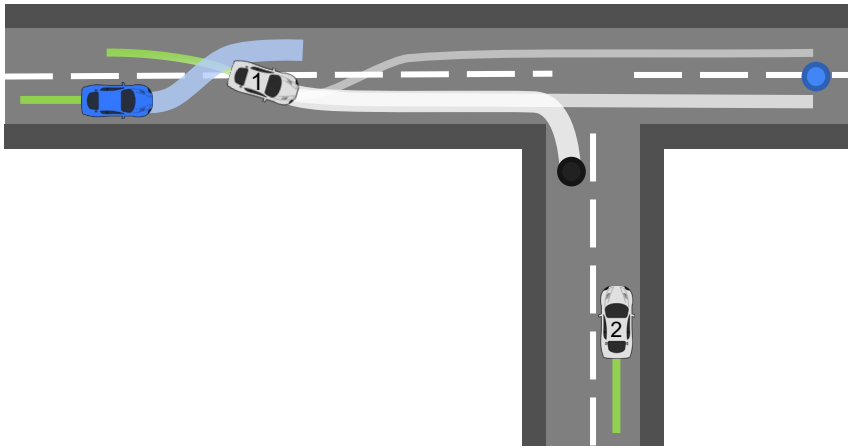


SCENARIO 1

Factual: Change lane left
Counterfactual: Exit right

"If we had turned right, we would have…

PRELIMINARY EVALUATION



SCENARIO 1

Factual: Change lane left
Counterfactual: Exit right

"If we had turned right, we would have…

Not reached the goal.

DISCUSSION

DISCUSSION

- What if macro actions are too abstract?

DISCUSSION

- What if macro actions are too abstract?
- How to support other types of questions?

DISCUSSION

- What if macro actions are too abstract?
- How to support other types of questions?
- How do we guarantee soundness of the generated explanations?

DISCUSSION

- What if macro actions are too abstract?
- How to support other types of questions?
- How do we guarantee soundness of the generated explanations?
- Are our generated explanations social enough?

DISCUSSION

- What if macro actions are too abstract?
- How to support other types of questions?
- How do we guarantee soundness of the generated explanations?
- Are our generated explanations social enough?
- How could we involve human participants into our evaluation to fully understand and tackle their concerns?

A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning

arXiv: https://arxiv.org/abs/2206.08783
Code: https://github.com/uoe-agents/xavi-ai4ad

Contributions:

1. We propose a framework for creating human-centric explanations for autonomous vehicles.
2. We propose a method to automatically generate human-centric explanations for the high-level driving behaviour of autonomous vehicles

In 2nd IJCAI Workshop on Artificial Intelligence for Autonomous Driving, 2022

THE UNIVERSITY of EDINBURGH informatics
Autonomous Agents Research Group
NLP UKRI CENTRE FOR DOCTORAL TRAINING
UKRI UK Research and Innovation

THE BAYES NETWORK MODEL: