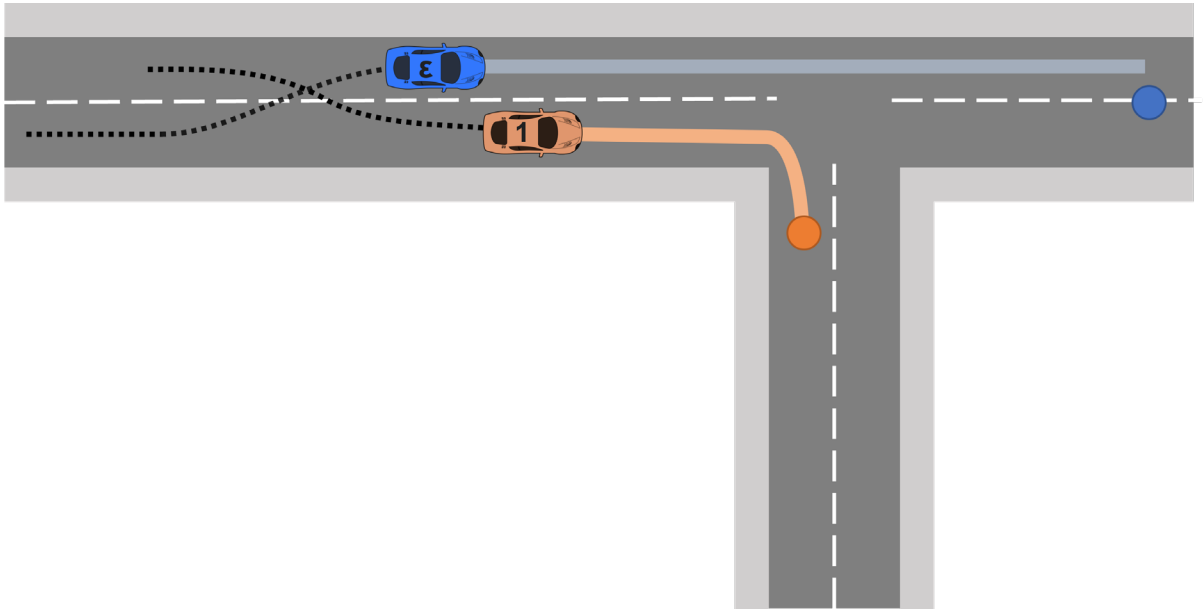
Causal Explanations for Stochastic Multi-Agent Decision-Making

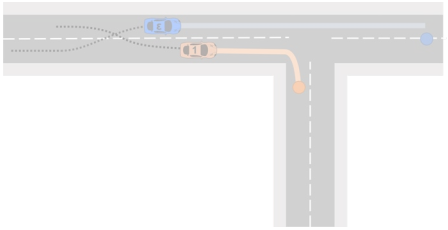
Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, Stefano V. Albrecht

In 5th International Workshop on Explainable and Transparent AI and Multi-Agent Systems

- Example domain: autonomous driving;
- Many agents, coupled interactions, difficult to explain;
- Safety-critical environment: explanations to build trust and understandability.

EXAMPLE
TIME


3
1



User **Agent**

Why did you change lanes?

It decreases the time to reach the goal. } Teleological

Why does it decrease the time to the goal?

Because vehicle 1 was slower than us.

Why was it slower?

It was decelerating and turning right.

What if it hadn't changed lanes before?

We would have gone straight. } Mechanistic

Actual responses generated by our system

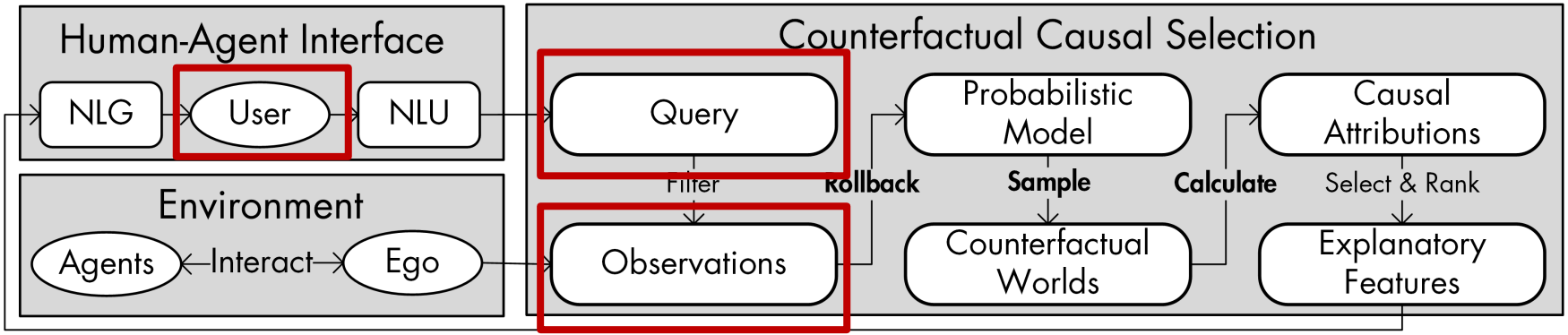
CEMA:

Causal Explanations for Multi-Agent decision-making

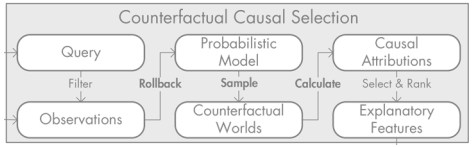
CEMA

- Applicable if you have:
  - Probabilistic model to predict future states;
  - No need to assume causal structure.
- Contrastive, causal, selected explanations.
- Designed for interaction.

CEMA


Human-Agent Interface
NLG
User
NLU
Environment
Agents
Interact
Ego
Counterfactual Causal Selection
Query
Filter
Observations
Rollback
Probabilistic Model
Sample
Counterfactual Worlds
Calculate
Causal Attributions
Select & Rank
Explanatory Features

CEMA



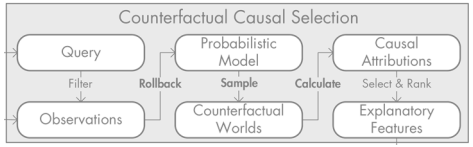
Rollback → Sample → Calculate

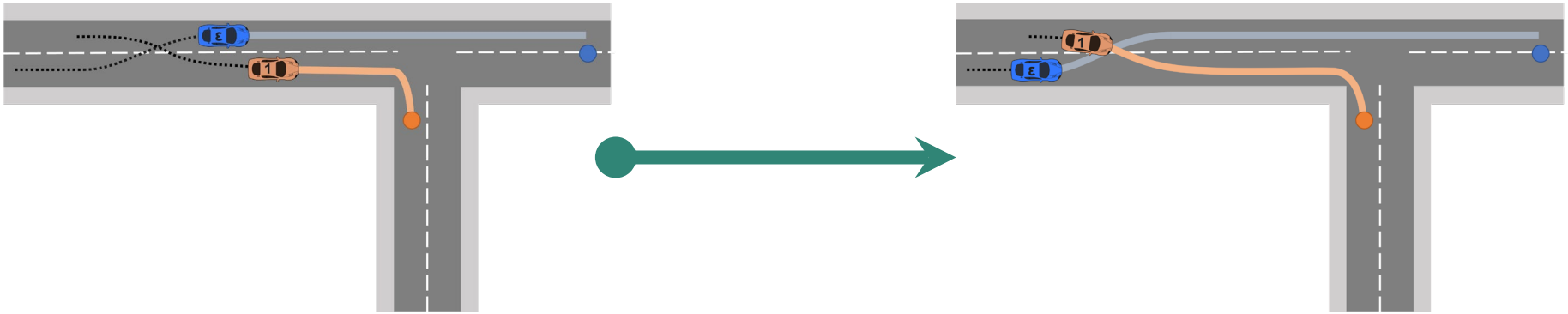
Counterfactual Effect Size Model

Tadeg Quillien and Christopher G Lucas, "Counterfactuals and the logic of causal selection"; *Psychological Review,* 130, 2023.

CEMA

Rollback!

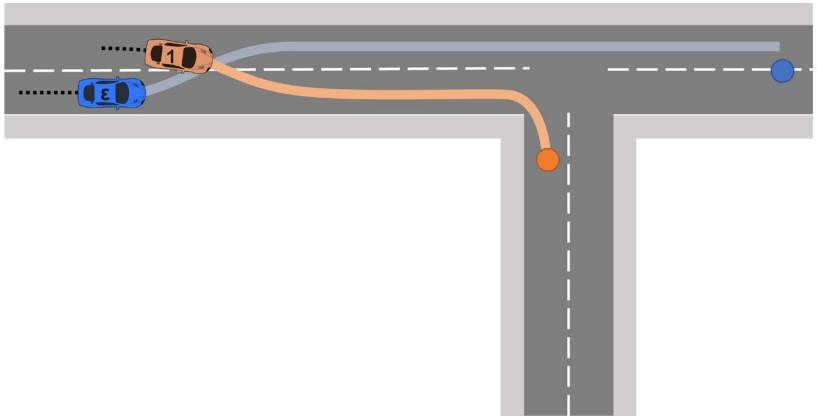


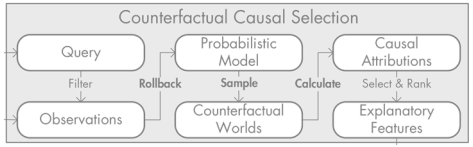


Rollback → Sample → Calculate

CEMA

Sample!





Action presence:
Lane change (1)

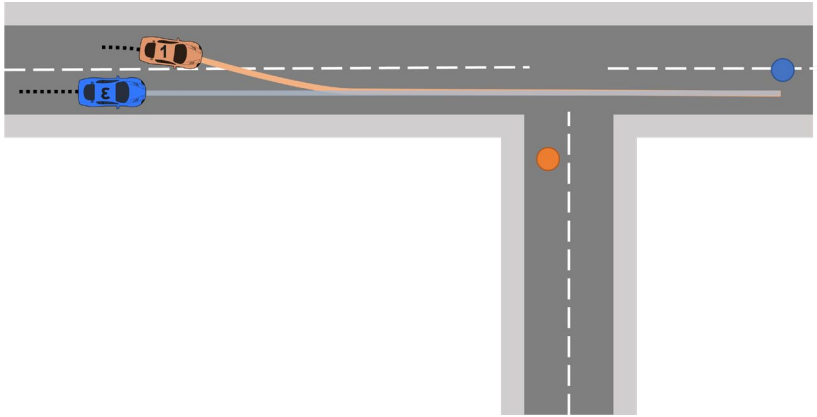
Intrinsic rewards:
Time-to-goal: 5 s
Jerk: 0.2 m/s^3
Collision: No

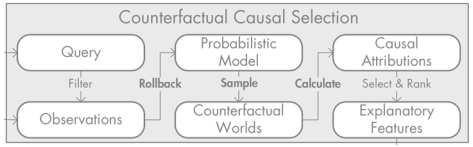
Features (binary):
{Decelerate, Turn, Slower, etc…}

Rollback → Sample → Calculate

CEMA

Sample!





Action presence:
No lane change (0)

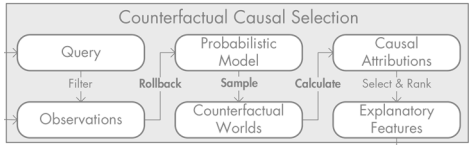
Intrinsic rewards:
Time-to-goal: 10 s
Jerk: 0.7 m/s^3
Collision: No

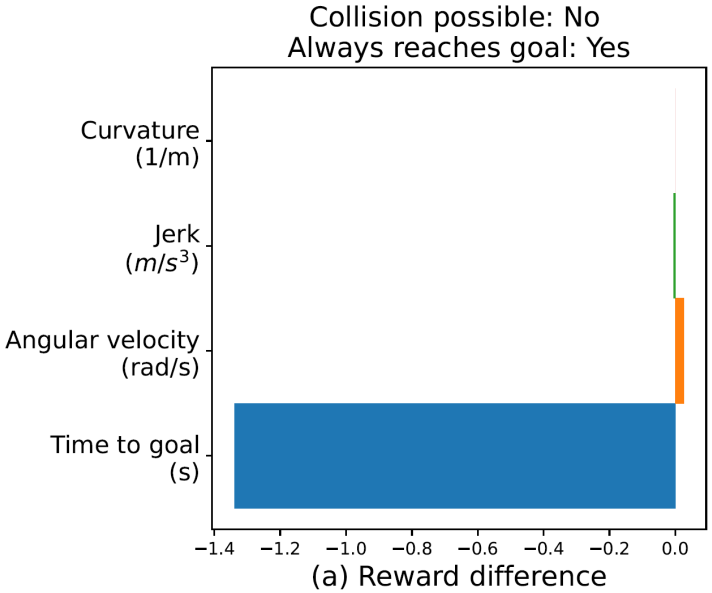
Features (binary):
{Accelerate, Continue, Faster, etc…}

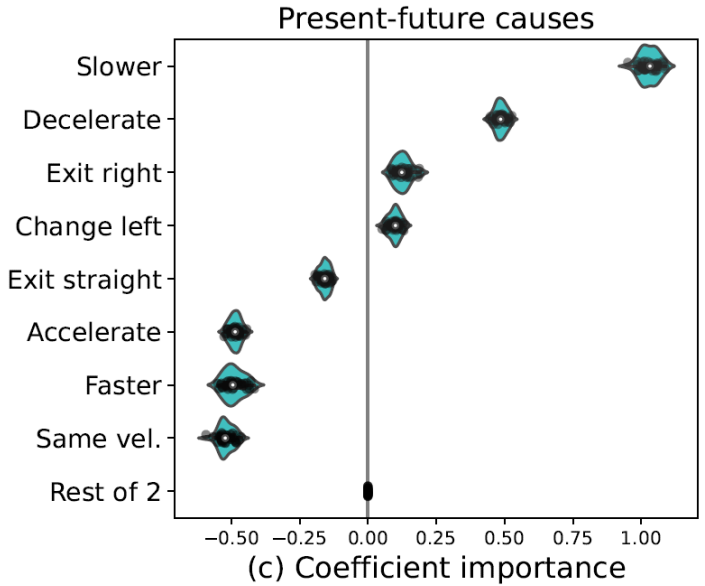
Rollback → Sample → Calculate

CEMA

Calculate!

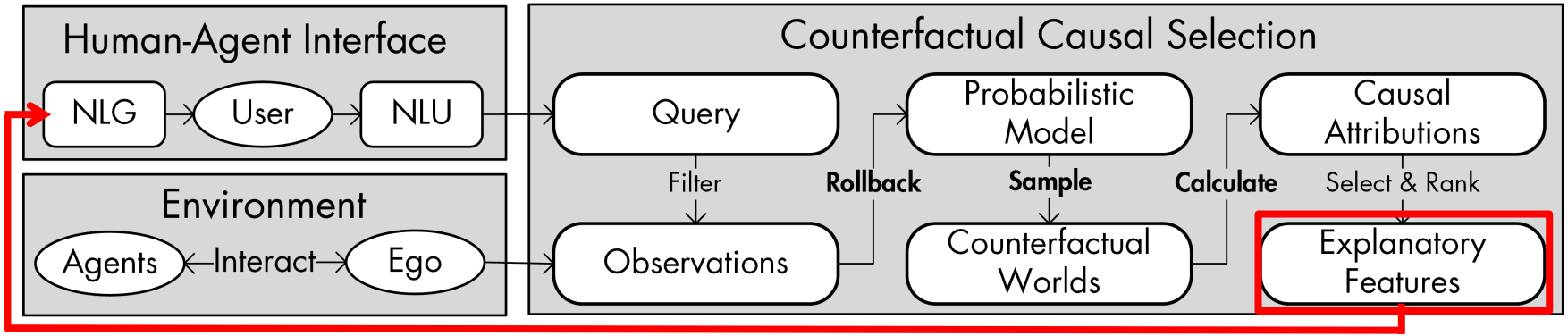


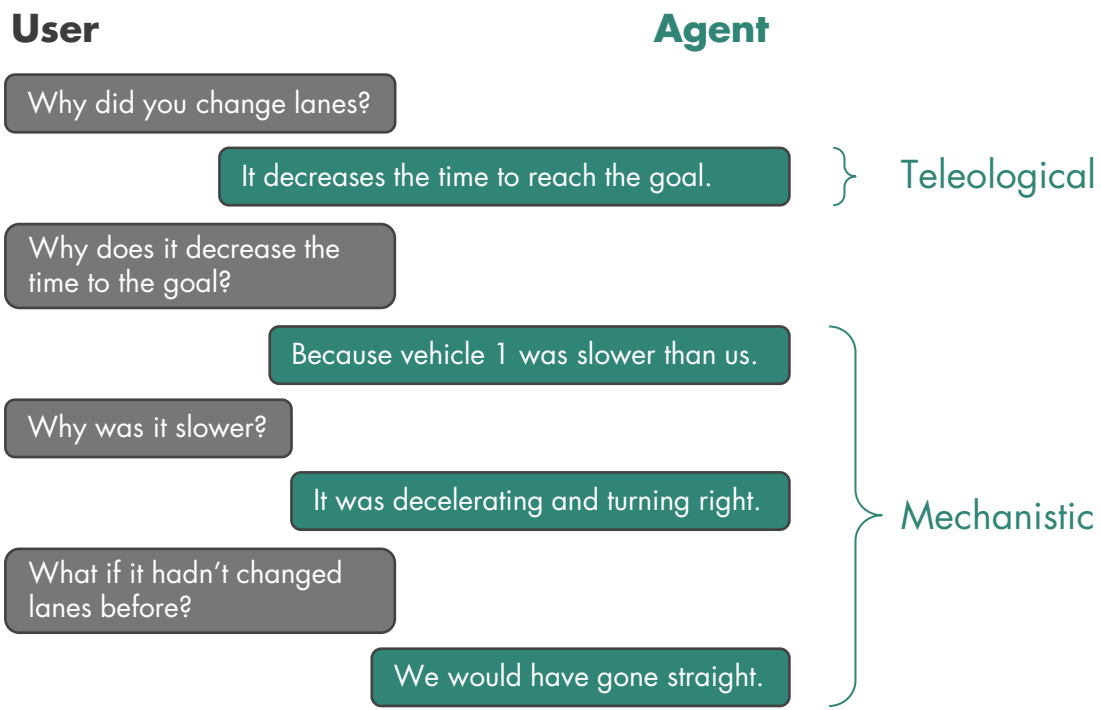




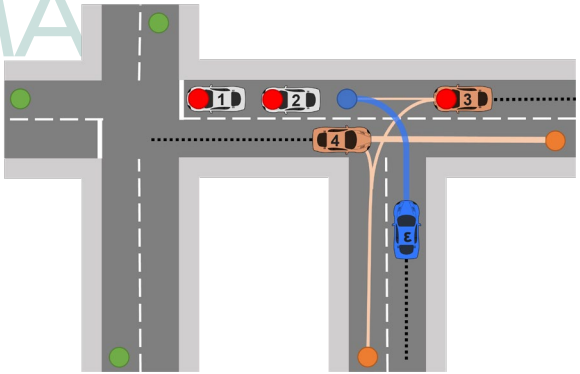
Rollback → Sample → Calculate

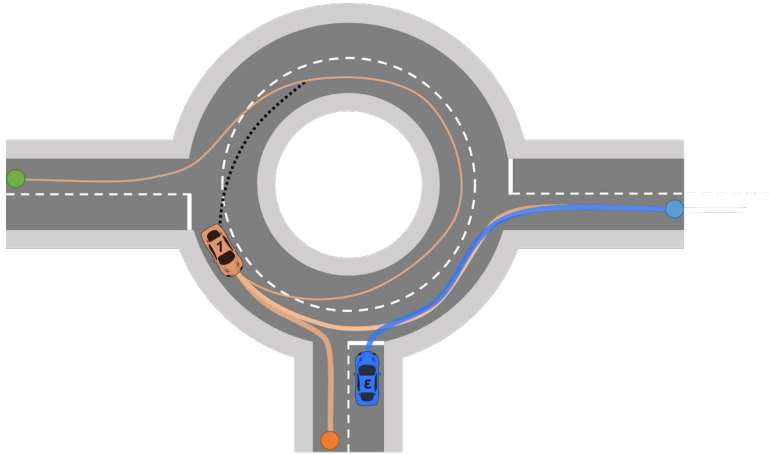
CEMA

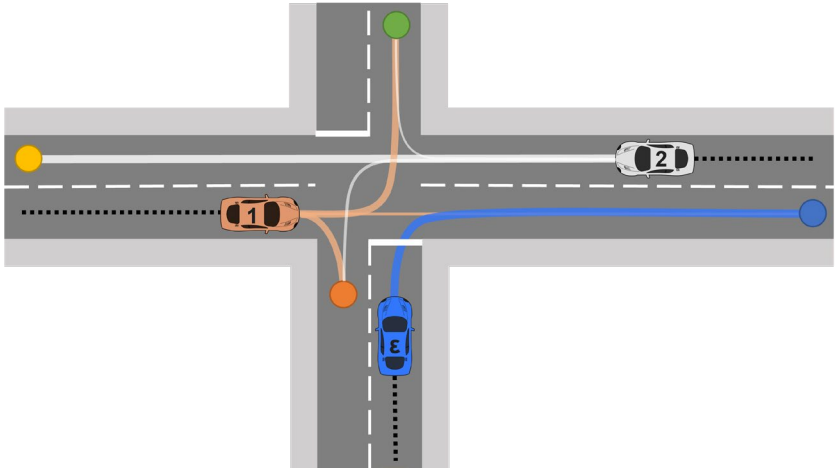

Human-Agent Interface
NLG
User
NLU
Environment
Agents
Interact
Ego
Counterfactual Causal Selection
Query
Filter
Observations
Rollback
Probabilistic Model
Sample
Counterfactual Worlds
Calculate
Causal Attributions
Select & Rank
Explanatory Features

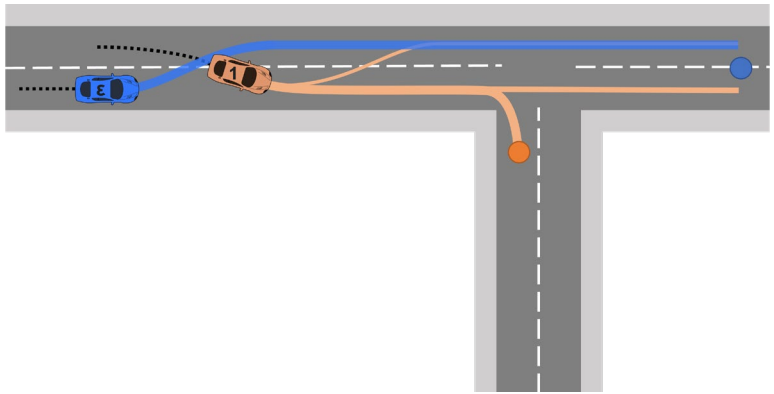

User
Agent
Why did you change lanes?
It decreases the time to reach the goal.
Teleological
Why does it decrease the time to the goal?
Because vehicle 1 was slower than us.
Why was it slower?
It was decelerating and turning right.
Mechanistic
What if it hadn't changed lanes before?
We would have gone straight.

CEMA









- Issue of lack of baselines;
- Human evaluation for goodness of explanations;
- Evaluation in more domains.

IGP2: Interpretable Goal-Based Prediction and Planning

Open-source code for CARLA simulator:
https://github.com/uoe-agents/IGP2

Blog post:
https://agents.inf.ed.ac.uk/blog/interpretable-prediction-planning-autonomous-driving/

Autonomous Agents Research Group

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

https://arxiv.org/abs/2302.10809

Contributions:

1. CEMA: A novel framework of Causal Explanations for stochastic Multi-Agent decision-making
2. Without assuming a causal structre, our method is applicable whenever predictive model is available.

In 5th International Workshop on Explainable and Transparent AI and Multi-Agent Systems

THE UNIVERSITY of EDINBURGH informatics

Autonomous Agents Research Group

UKRI CENTRE FOR DOCTORAL TRAINING

UK Research and Innovation