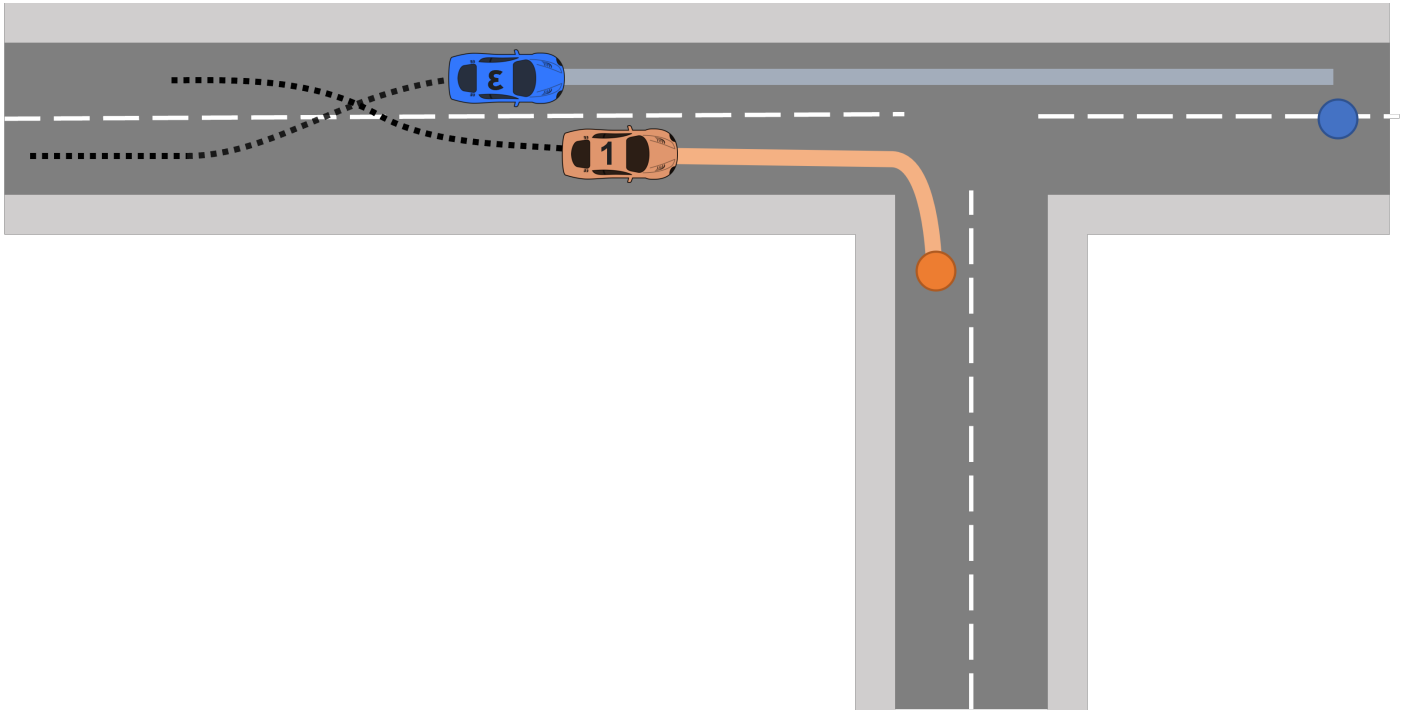
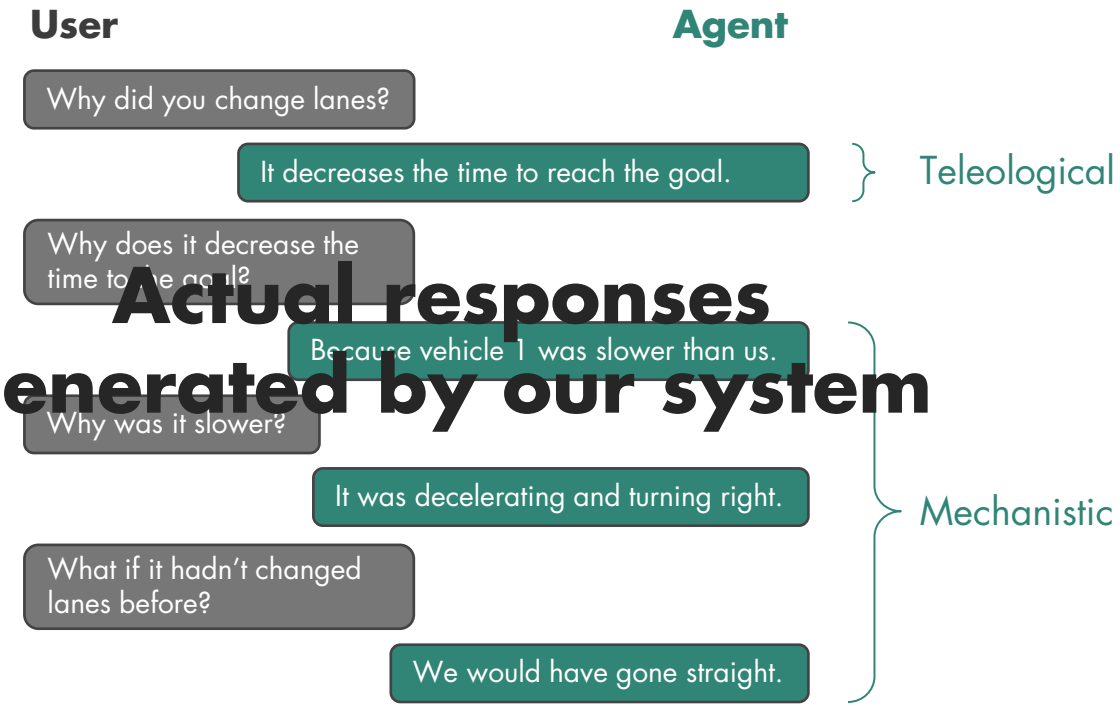
Causal Explanations for Sequential Decision-Making in Multi-Agent Systems

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

In IJCAI 2023 Workshop on Explainable Artificial Intelligence

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


3
1



Actual responses generated by our system

CEMA:

Causal Explanations in Multi-Agent system

CEMA

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

CEMA

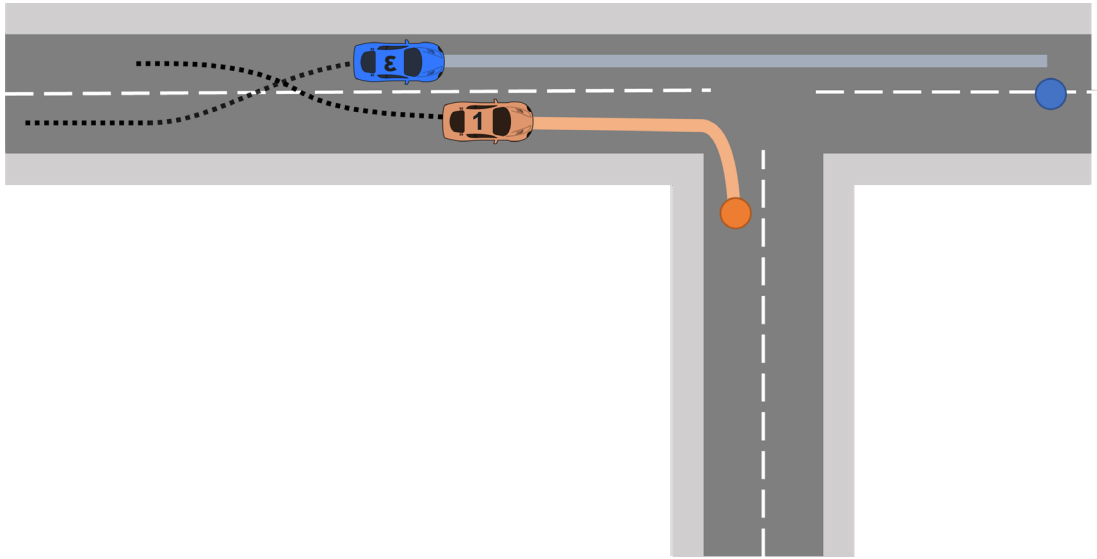
Rollback → Sample → Calculate

Counterfactual Effect Size Model

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

CEMA

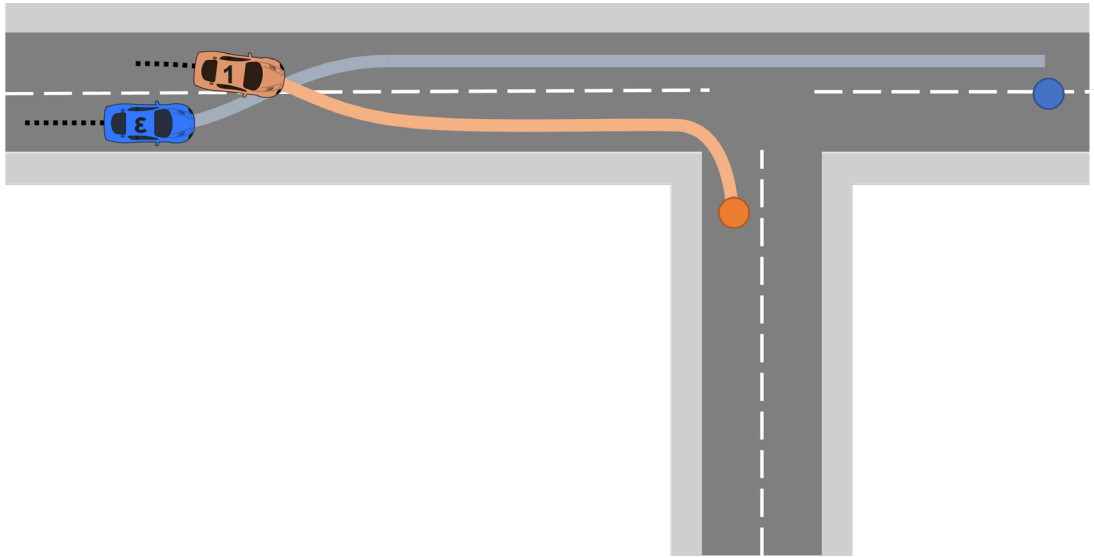
Rollback!



Rollback → Sample → Calculate

CEMA

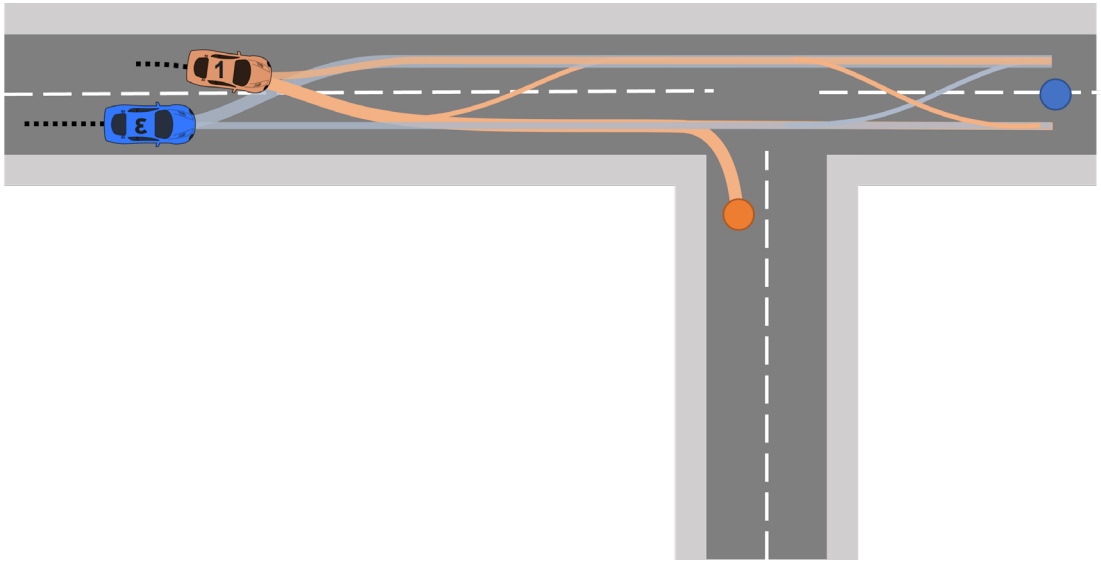
Rollback!



Rollback → Sample → Calculate

CEMA

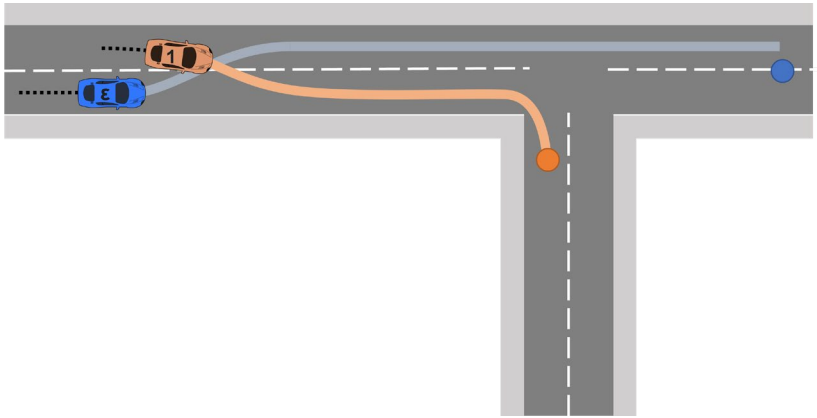
Sample!



Rollback → Sample → Calculate

CEMA

Sample!



Action presence:
Lane change (1)

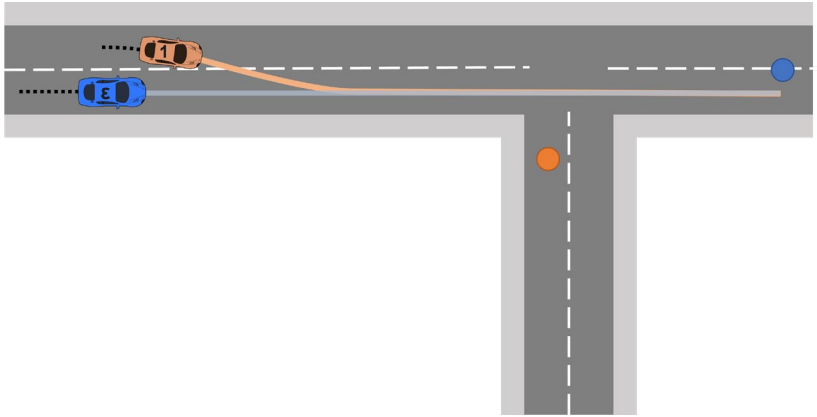
Intrinsic rewards:
Time-to-goal: 5 s
Jerk: 0.2 m/s^3
Collision: No

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

Rollback → Sample → Calculate

CEMA

Sample!



Action presence:
No lane change (0)

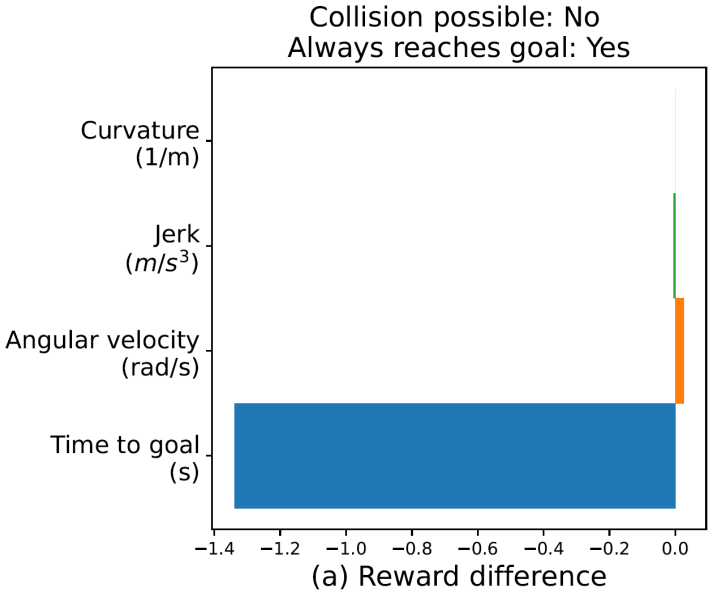
Intrinsic rewards:
Time-to-goal: 10 s
Jerk: 0.7 m/s^3
Collision: No

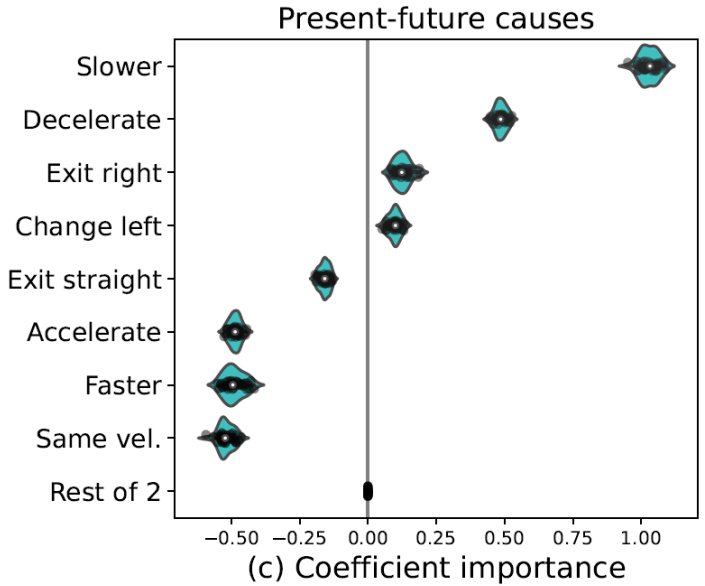
Features (binary):
{Accelerate, Continue, Faster, etc…}

Rollback → Sample → Calculate

CEMA

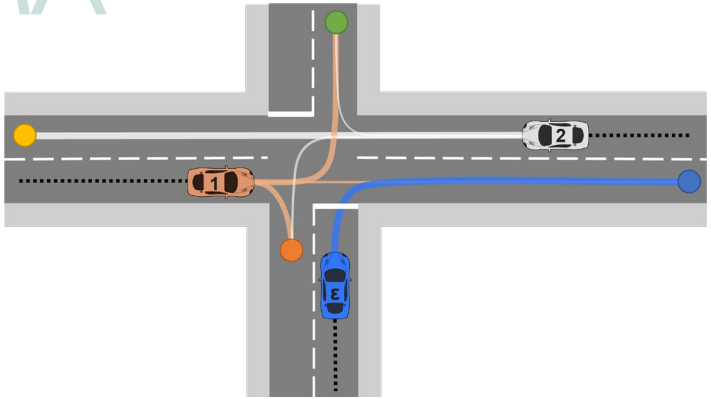
Calculate!

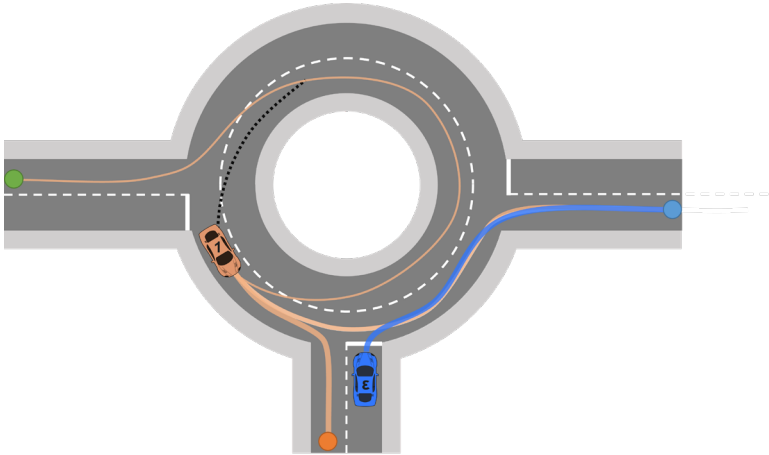


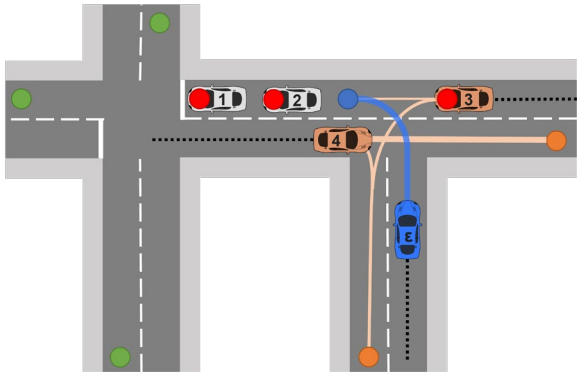


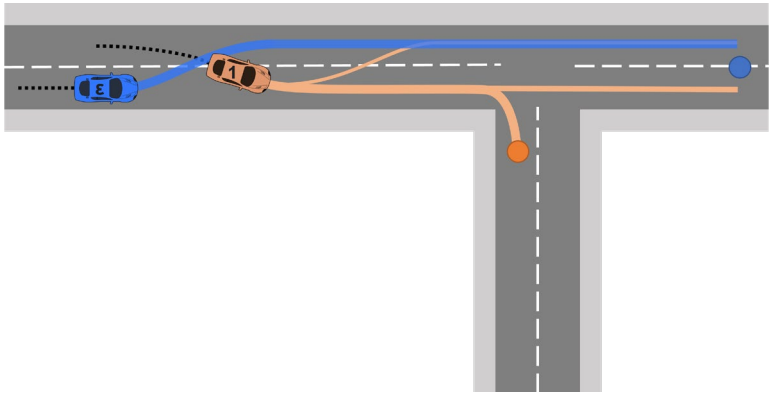
Rollback → Sample → Calculate

CEMA









- Works for large number of agents:
  - Tested with up to 20 agents;
- Two-stage human evaluation for goodness of explanations:
  - Compared against human-written baseline from participants;
- TODO:
  - Evaluation in more domains;
  - Integration with NLP.

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 IJCAI 2023 Workshop on Explainable Artificial Intelligence

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