How to create **causal explanations** for sequential decision-making in multi-agent systems that **people actually like?**

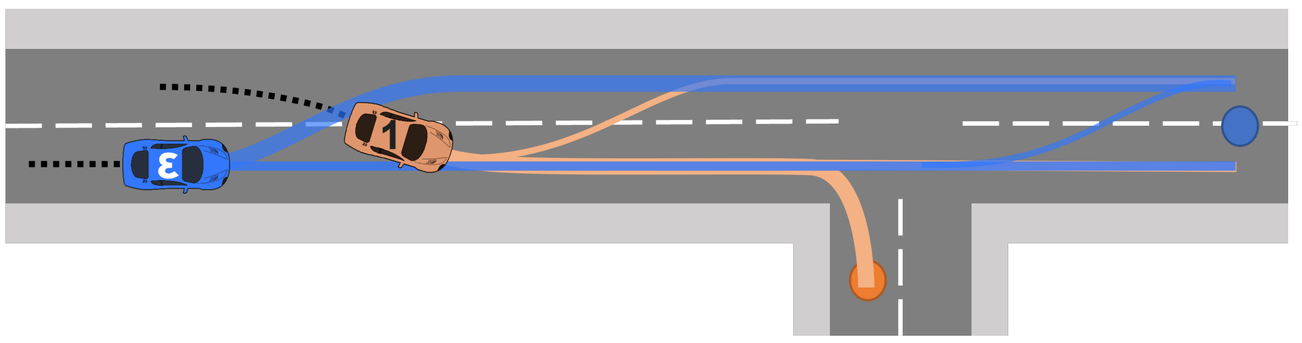
Bálint Gyevnár, Cheng Wang, Chris Lucas, Shay Cohen, Stefano Albrecht
University of Edinburgh

Multi-agent systems are difficult to explain, even for people:

Coupled interactions • Partial observability • Conflicting goals • Safety

Human-centric explanations *help to alleviate these issues:*

Causal • Contrastive • Selected • Conversational

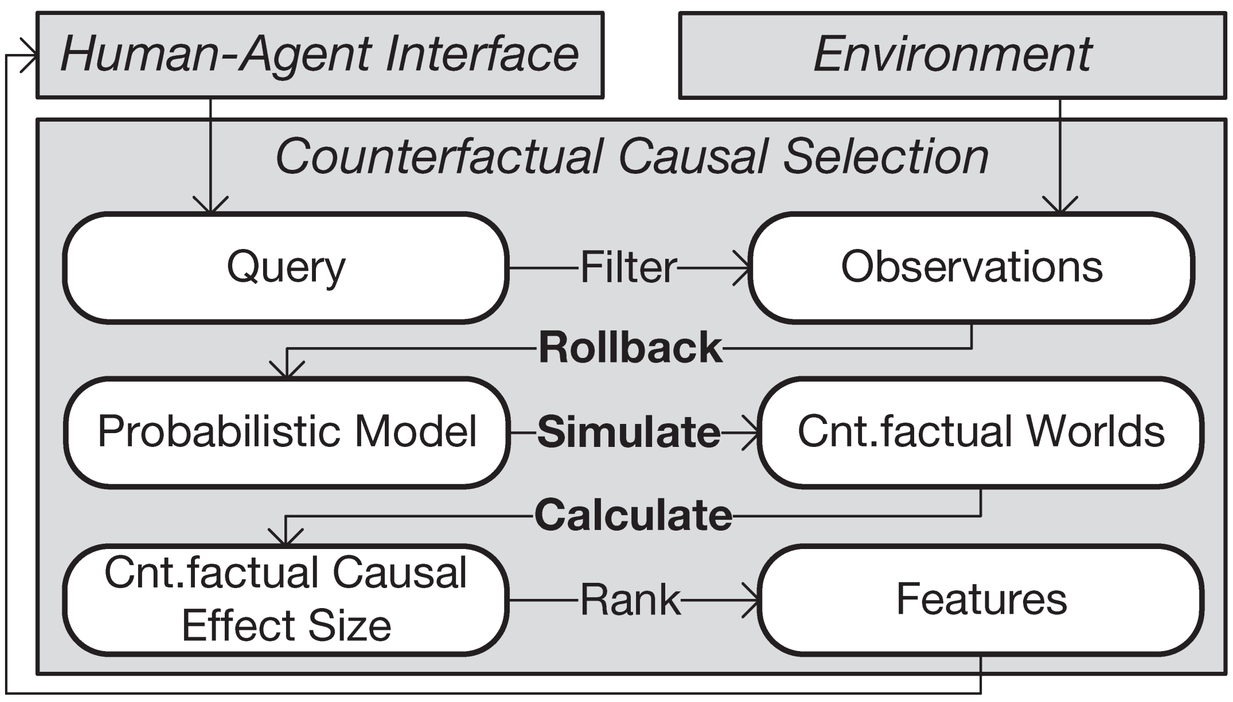


The blue car's goal is straight ahead.

It suddenly changes lanes.

How do we explain its behaviour?

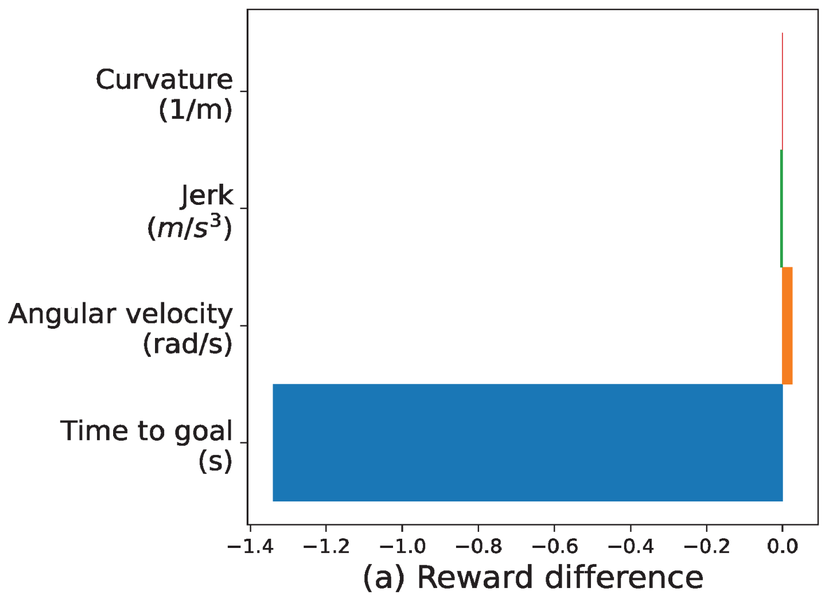
CEMA *Causal Explanations for Multi-Agent systems*



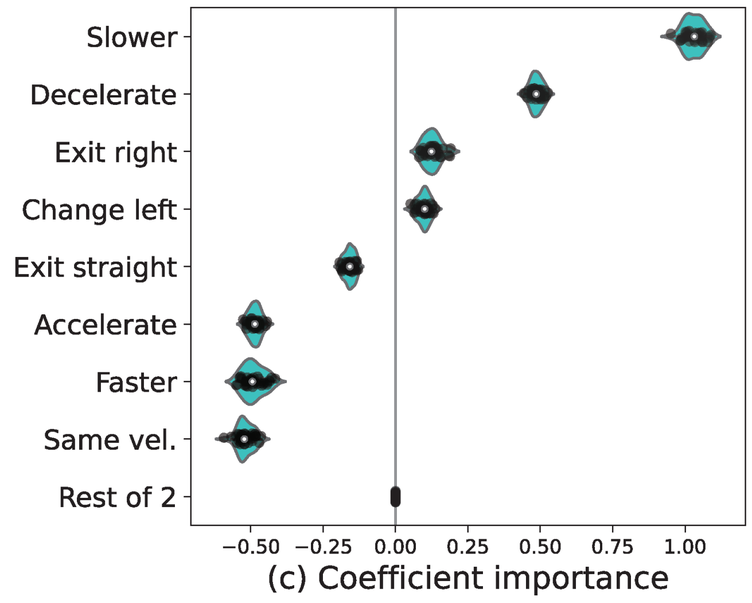
1. Rollback time to before the action you want to explain;
2. Simulate counterfactuals grounded in the real world;
3. Correlated variables with behaviour across counterfactuals rank causes.

CEMA generates explanations with two modes:

Goal-oriented:

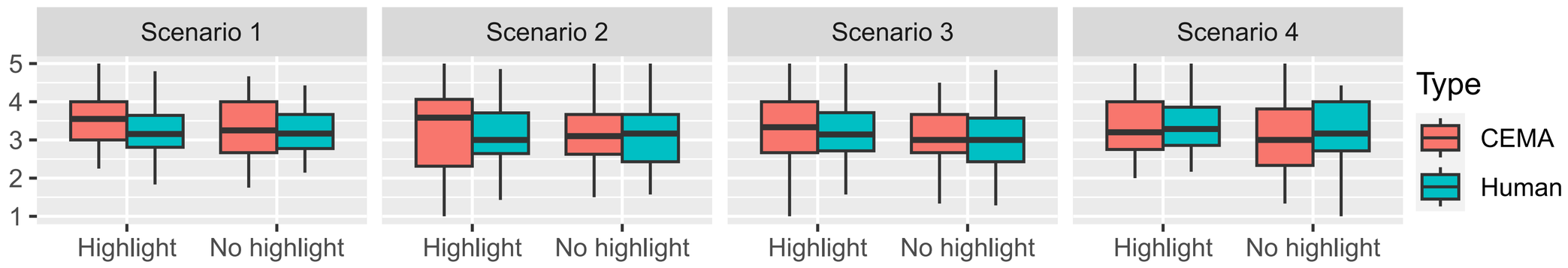


Mechanistic:



Explanations are also:

- Faithful;
- Robust to simulation quality;
- Robust to sample size;
- Tested with up to 20 agents;
- Rated to human-quality.



Read the paper for more

THE UNIVERSITY of EDINBURGH informatics
Autonomous Agents Research Group
UKRI CENTRE FOR DOCTORAL TRAINING