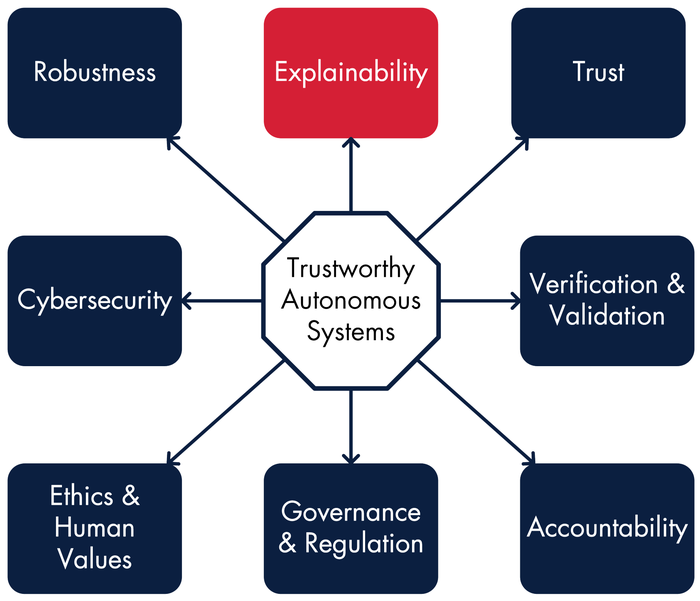
Conversational Framework for Social Explainable AI to build Trustworthy Autonomous Systems

Balint Gyevnar

in collaboration with:
Cheng Wang
Christopher G. Lucas
Shay B. Cohen
Stefano V. Albrecht

Trustworthy Autonomous Systems



Explainability

- restores agency
- enable contestation
- improves knowledge
- builds trust

Only when

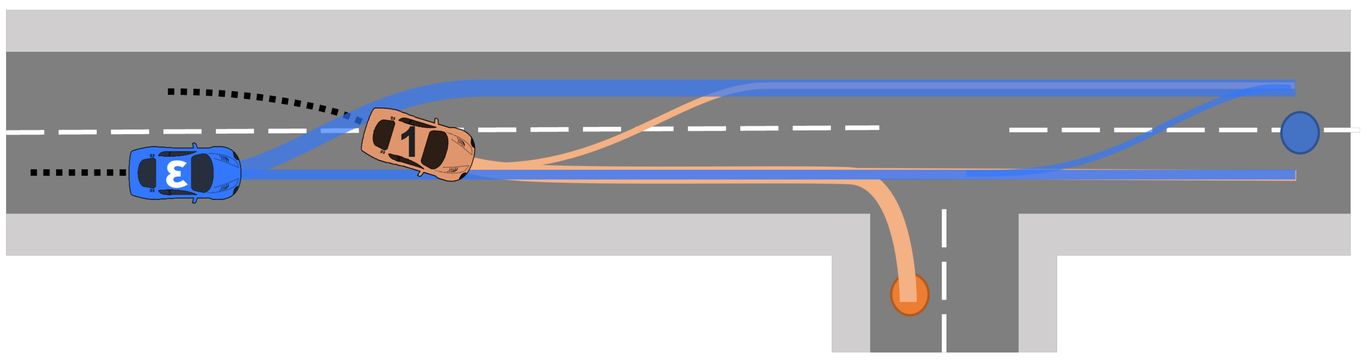
- includes causes
- contrastive
- selected for biases
- conversational

I.e., social

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

Explanations based on human-model of causal reasoning:

- Simulate counterfactuals to select causes:
  - Based on prior (cognitive) distribution
  - Anchored to observations
- Correlation to select from causes:
  - C caused E if C is highly correlated with E across counterfactuals
- Explanations satisfy social XAI criteria



The autonomous vehicle is heading to the blue goal.

It decides to change lanes after the other vehicle cuts in front of it and begins to slow down.

A passenger is surprised and the following conversation ensues:

User / **Agent**

Why did you change lanes?

It decreases the time to reach the goal.

Why does it decrease the time to the goal?

Because vehicle 1 was slower than us.

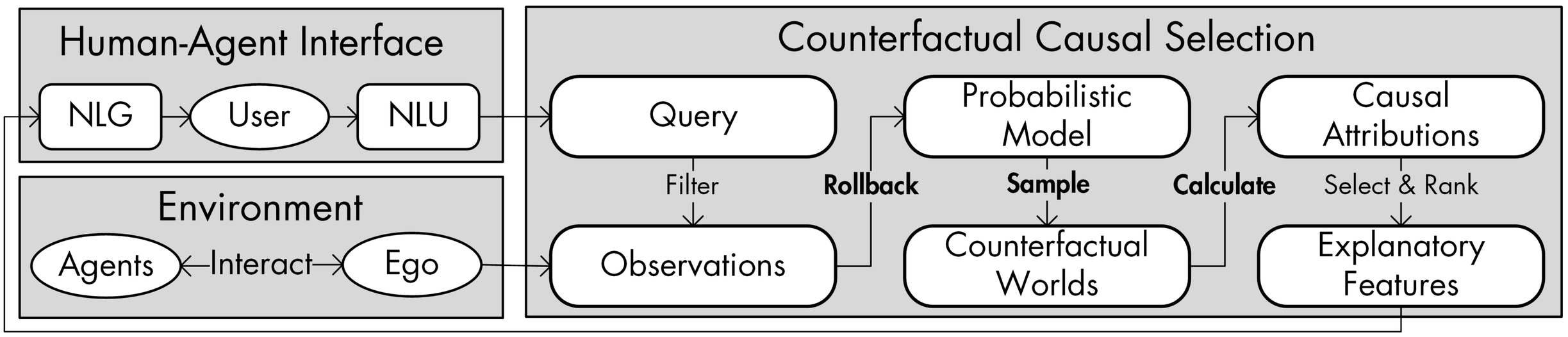
Why was it slower?

It was decelerating and turning right.

What if it hadn't changed lanes before?

We would have gone straight.

A cross-disciplinary framework



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