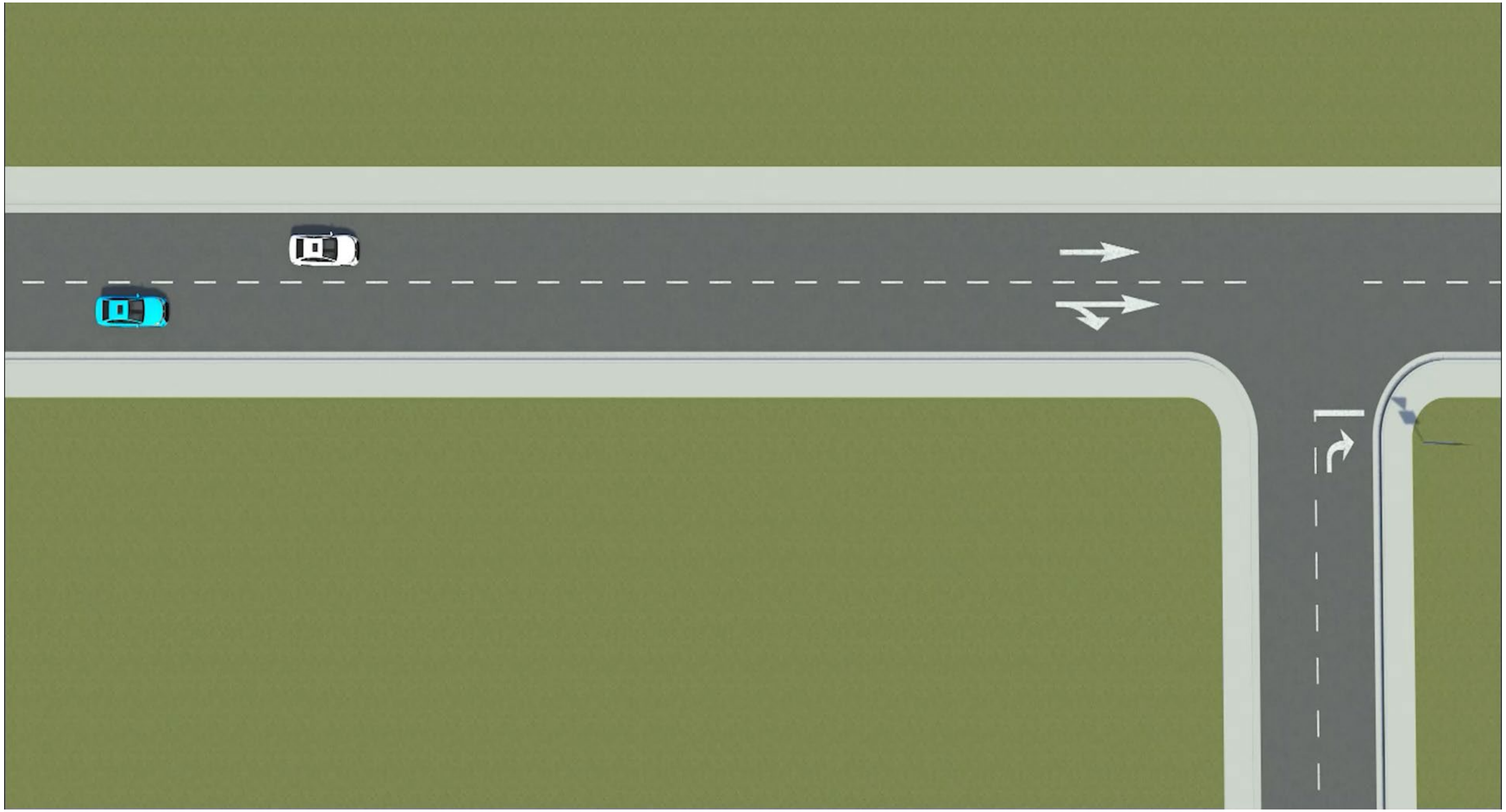
Causal Explanations for Sequential Decision-Making in Multi-Agent Systems

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

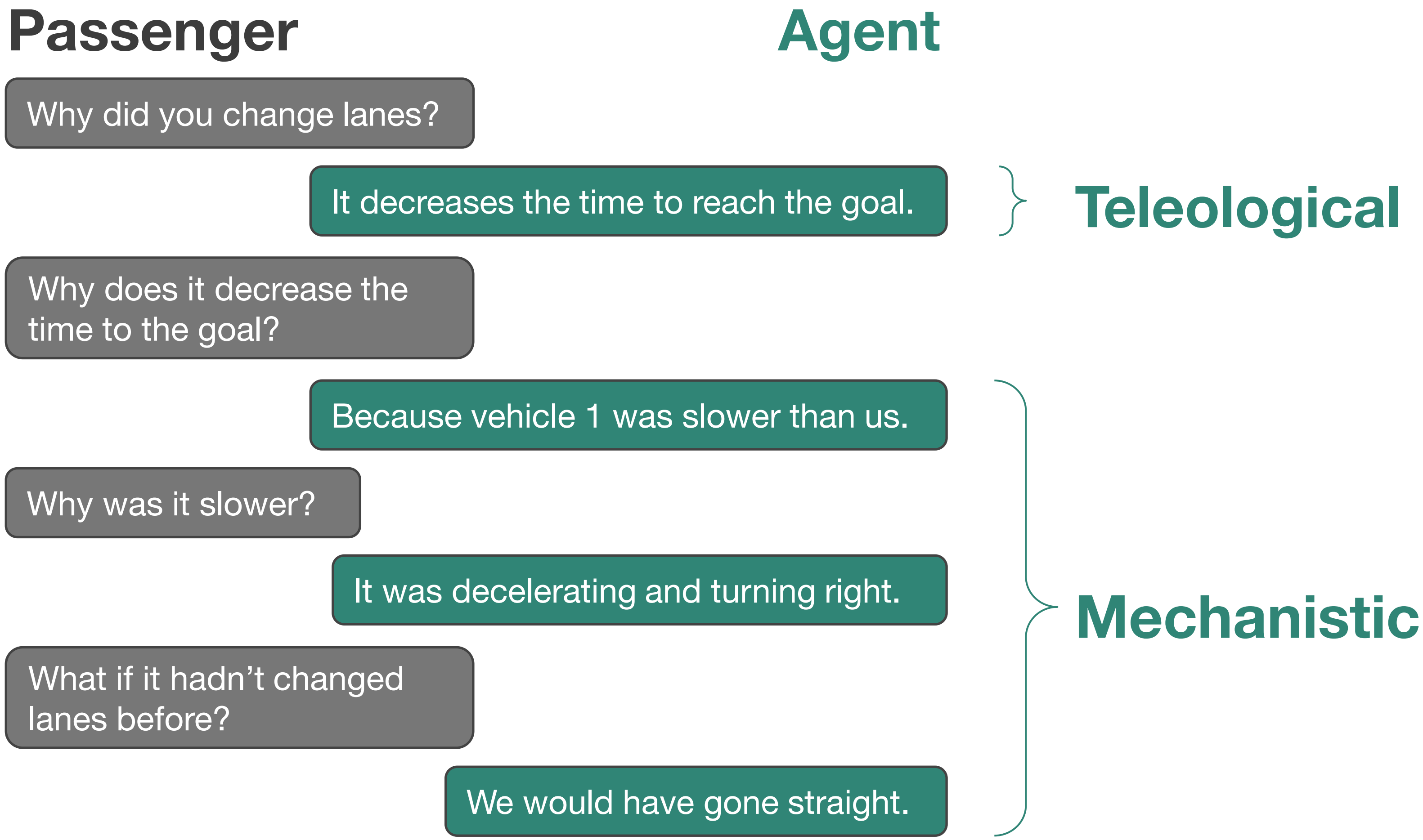
In 23rd International Conference on Autonomous Agents and Multi-Agent Systems

explainable AI · human-centric design · multi-agent systems · cognitive science · natural language processing

ILLUSTRATIVE SCENARIO



EXAMPLE INTERACTION



Passenger **Agent**

Passenger: Why did you change lanes?

Agent: It decreases the time to reach the goal.

Passenger: Why does it decrease the time to the goal?

Agent: Because vehicle 1 was slower than us.

Passenger: Why was it slower?

Agent: It was decelerating and turning right.

Passenger: What if it hadn’t changed lanes before?

Agent: We would have gone straight.

Teleological

Mechanistic

CAUSAL EXPLANATIONS FOR MULTI-AGENT SYSTEMS (CEMA)

CEMA

Causal **E**xplanations for **M**ulti-**A**gent Systems

WHY MULTI-AGENT SYSTEMS?

Why multi-agent systems?

Coupled interactions;

Conflicting goals;

Partial observability;

Communication;

Often difficult to explain, even for humans.

WHY DO WE WANT EXPLANATIONS?

Critical environments:

Socially: Others react to our agents and change their behaviour;

Epistemically: Partial observability and shared rules;

Safety: Actions can harm agents/humans/environment;

Explanations help:

Explain confusing behaviour;

Highlight occluded information;

Calibrate trust.

HUMAN-CENTRED EXPLANATIONS

Theory tells us that explanations should be:

Causal;

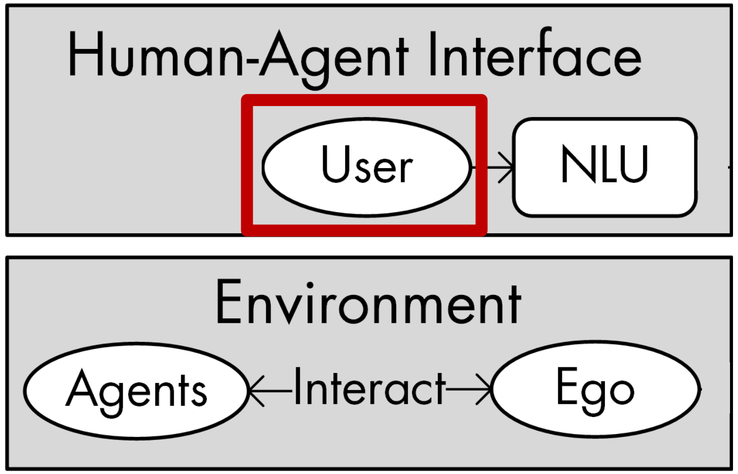
Contrastive;

Selected;

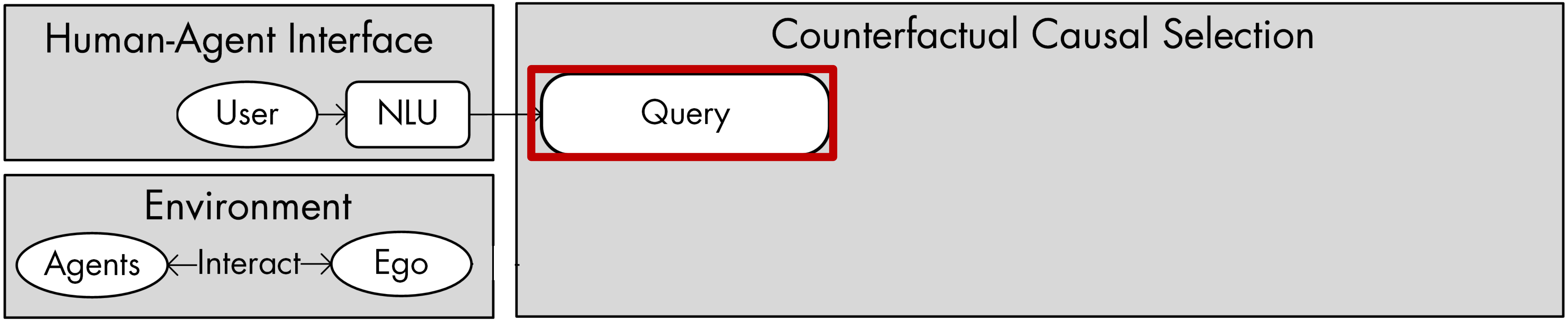
Conversational;

[2] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences.

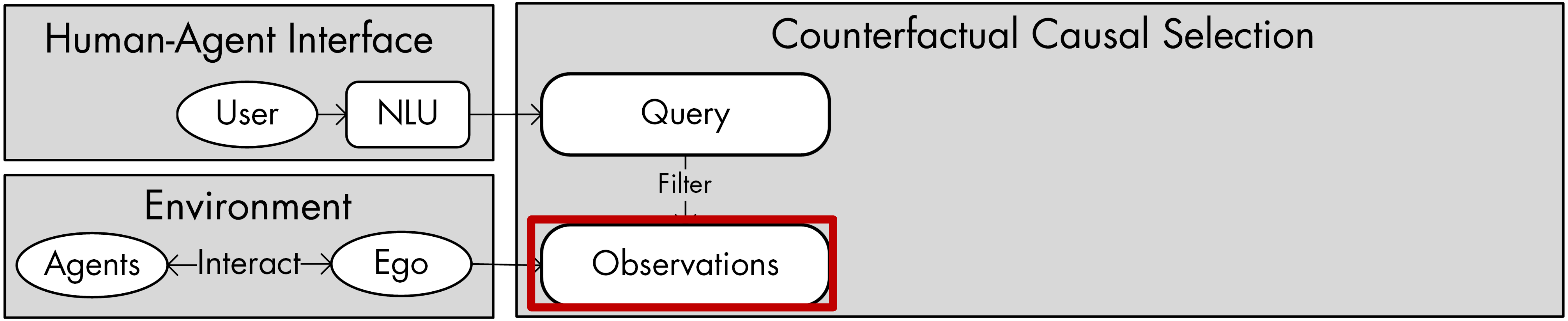
How does CEMA work?



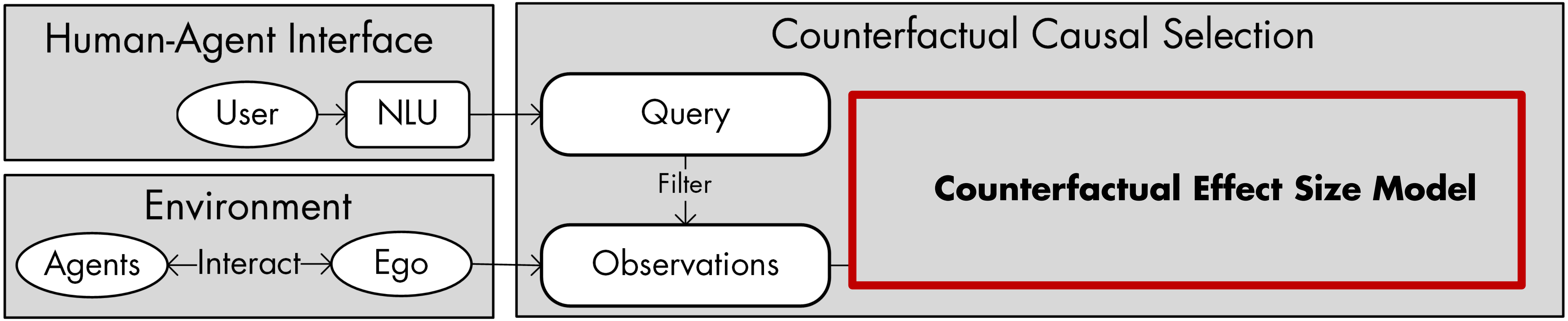
How does CEMA work?



How does CEMA work?



How does CEMA work?



COUNTERFACTUAL EFFECT SIZE MODEL (CESM)

Based on how people could select causes to explain;

People simulate worlds to select causes:

Using some prior (cognitive) distribution;
But anchored to observations;

People use correlation to select among causes:

C caused E if C is highly correlated with E across worlds.

[1] Quillien, T., & Lucas, C. G. (2023, June 8). *Counterfactuals and the Logic of Causal Selection.*

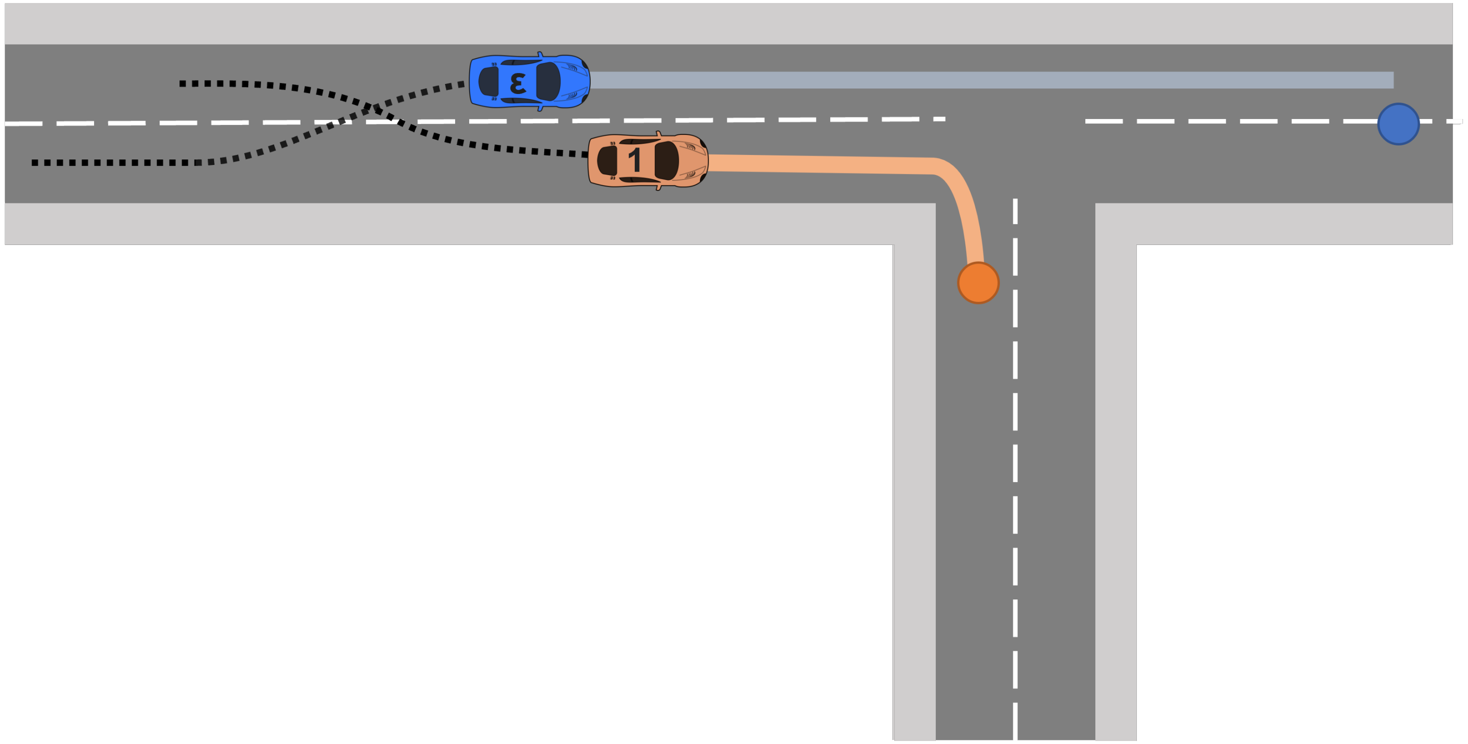
CEMA – THREE-STEP APPROACH

Counterfactual Effect Size Model in CEMA

Rollback → Simulate → Correlate

CEMA – ROLLBACK

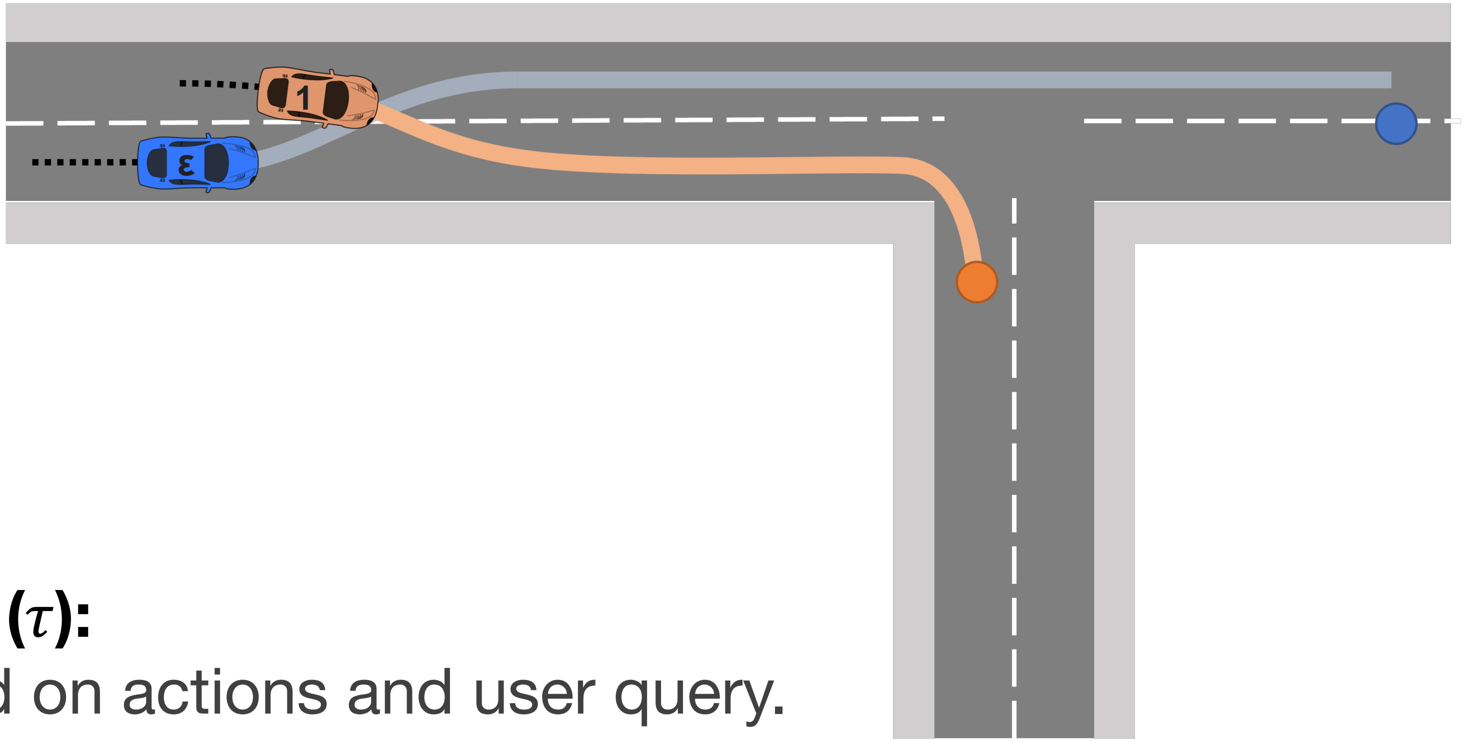
Observed trajectory: $s_{1:t}$



Rollback → Simulate → Correlate

CEMA – ROLLBACK

Rolled back trajectory: $s_{1:\tau}$



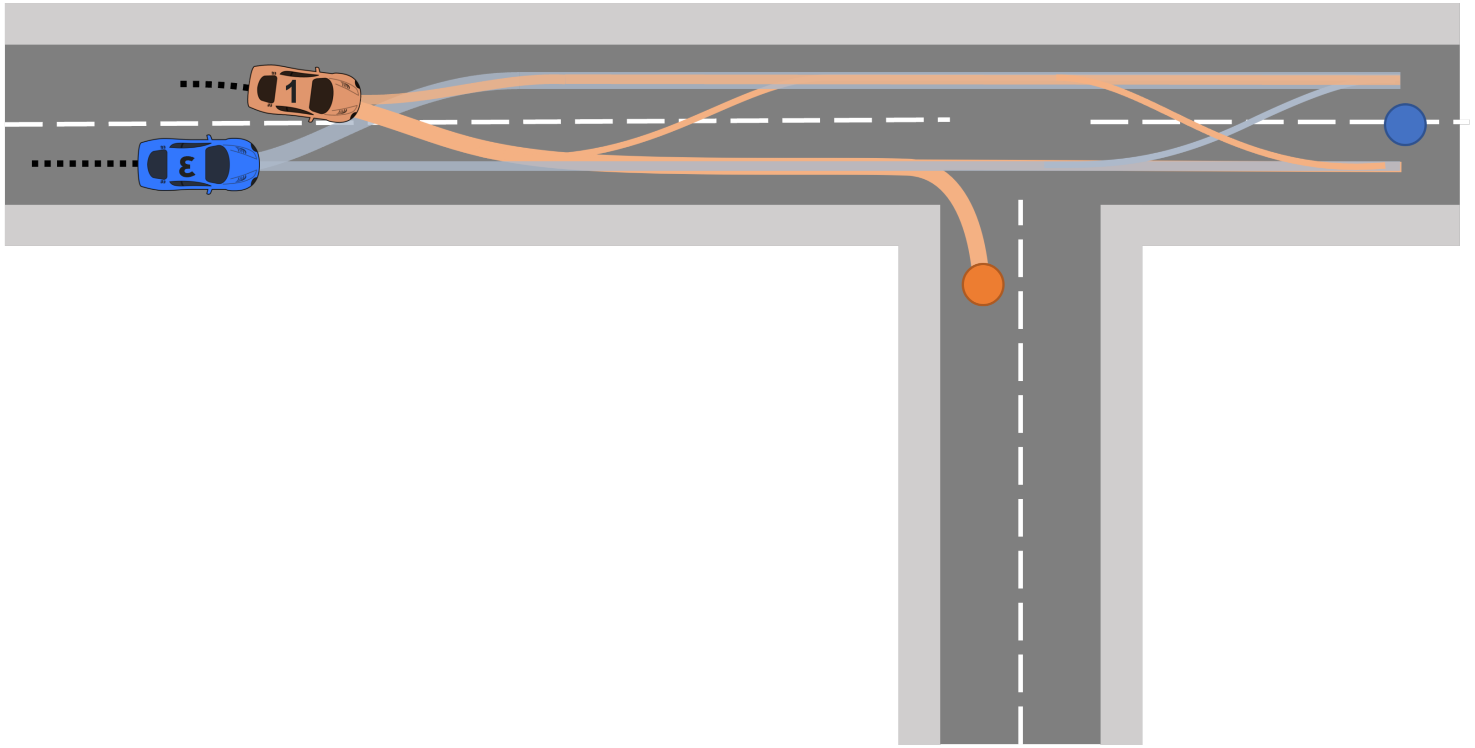
Rollback time (τ):

Selected based on actions and user query.

Rollback → Simulate → Correlate

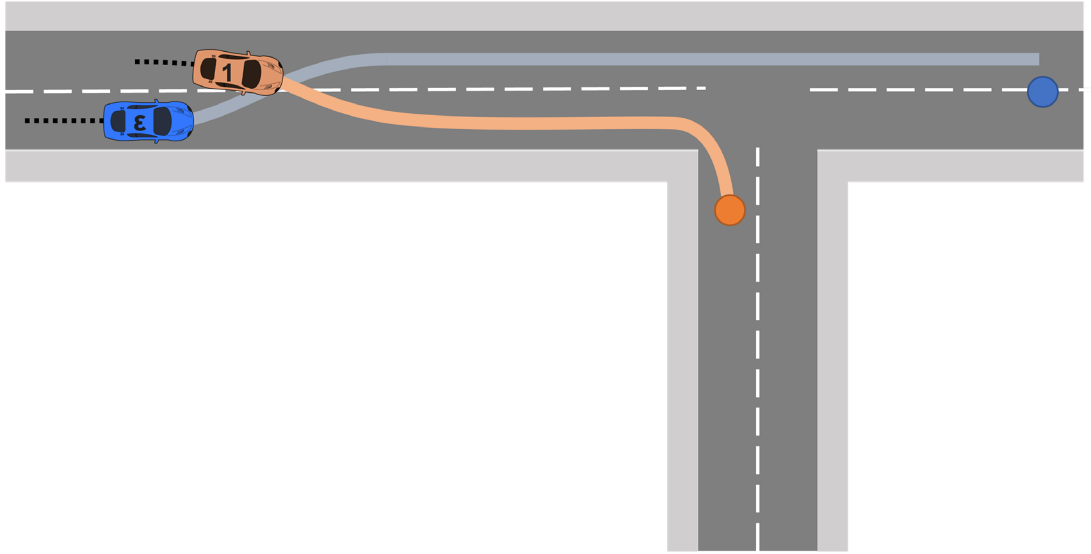
CEMA – SIMULATE (COUNTER)FACTUALS

Simulate (counter)factual worlds



Rollback → **Simulate** → Correlate

CEMA – SIMULATE FACTUAL



Action presence:

Lane change (1)

Rewards:

Time-to-goal: 7 s

Comfort (jerk): 0.5 m/s^3

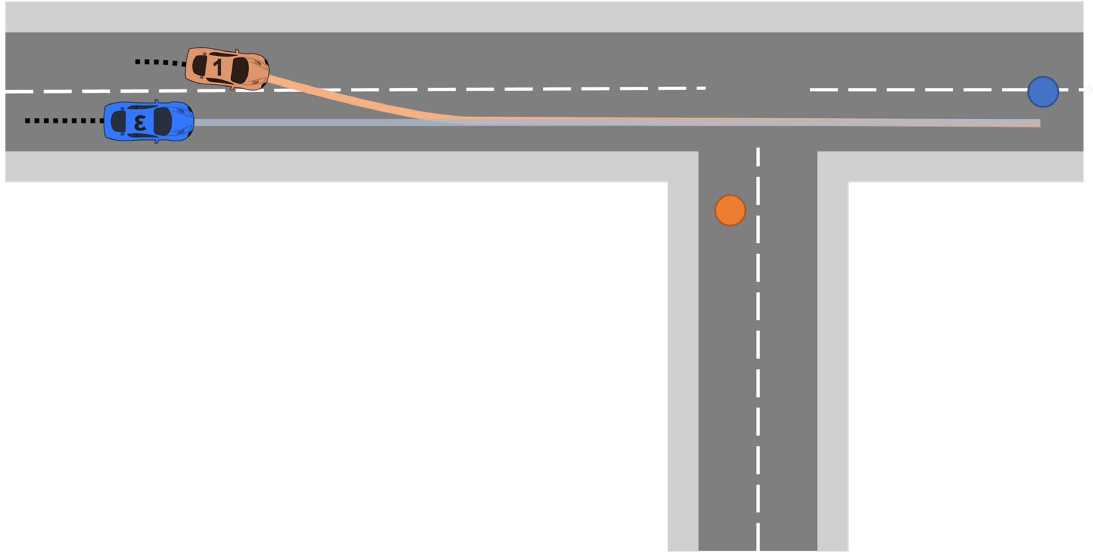
Collision: No

Features from trajectory:

{Decelerate, Turn, Slower, etc…}

Rollback → **Simulate** → Correlate

CEMA – SIMULATE COUNTERFACTUAL



Action presence:

No lane change (0)

Rewards:

Time-to-goal: 5 s
Comfort (jerk): 0.1 m/s^3
Collision: No

Features from trajectory:

{Accelerate, Continue, Faster, etc…}

Rollback → **Simulate** → Correlate

CEMA – CORRELATE – TELEOLOGICAL CAUSES

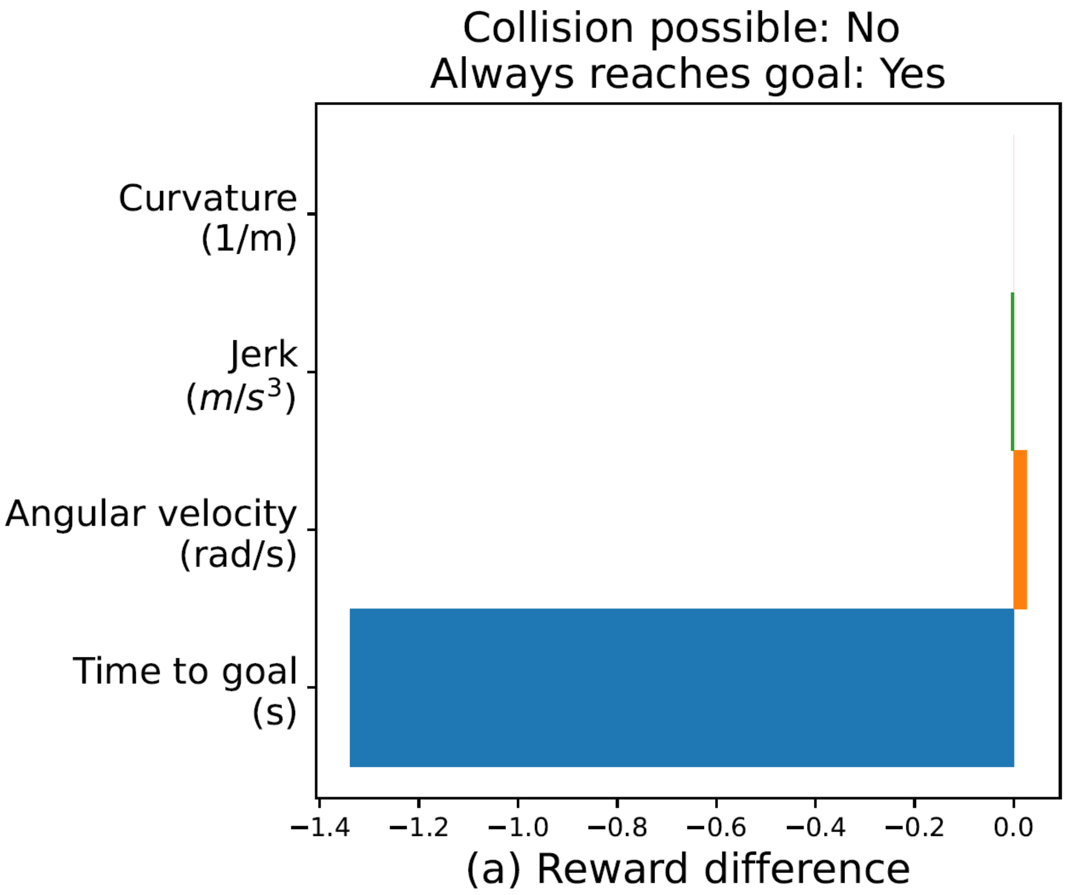
Process for teleological causes;

Difference of expected rewards between:

Worlds where queried action happened;
Where queried action did not happen.

Rollback → Sample → **Correlate**

CEMA – CORRELATE – TELEOLOGICAL CAUSES



Rollback → Sample → **Correlate**

CEMA – CORRELATE – MECHANISTIC CAUSES

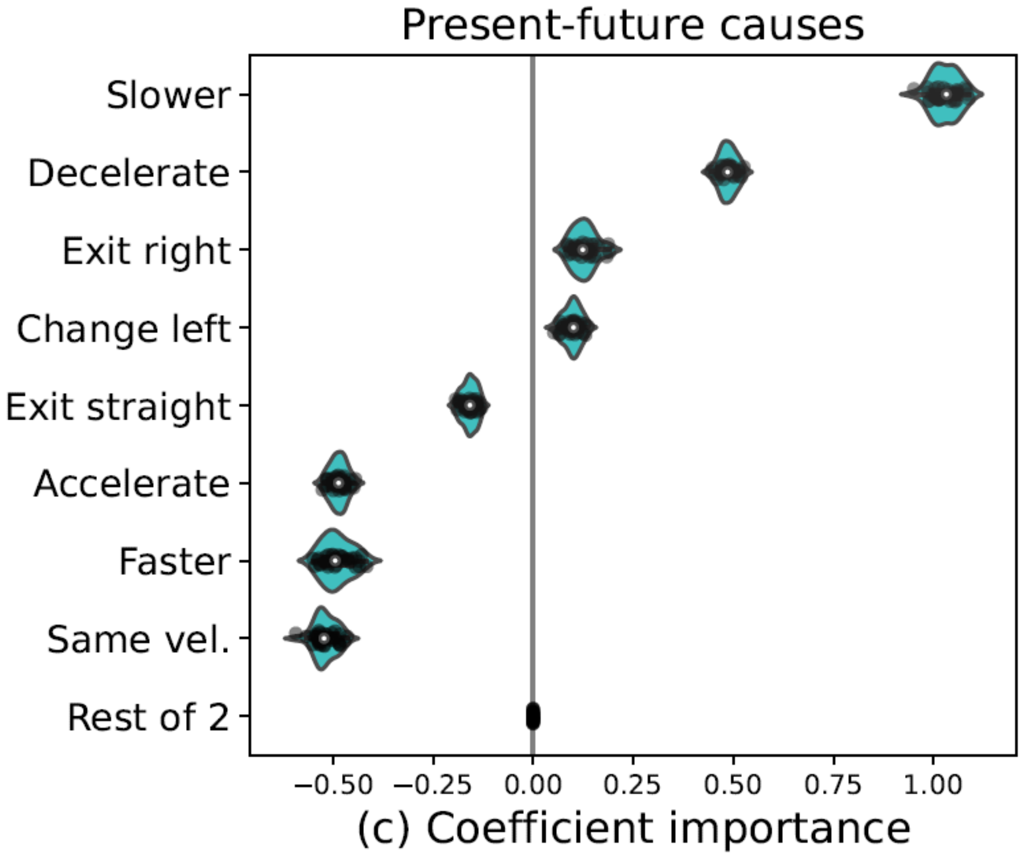
Process for mechanistic causes:

1. Fit interpretable model to trajectory features;
2. Predict presence of queried action;
3. Extract feature importance;

Counterfactual effect size of features.

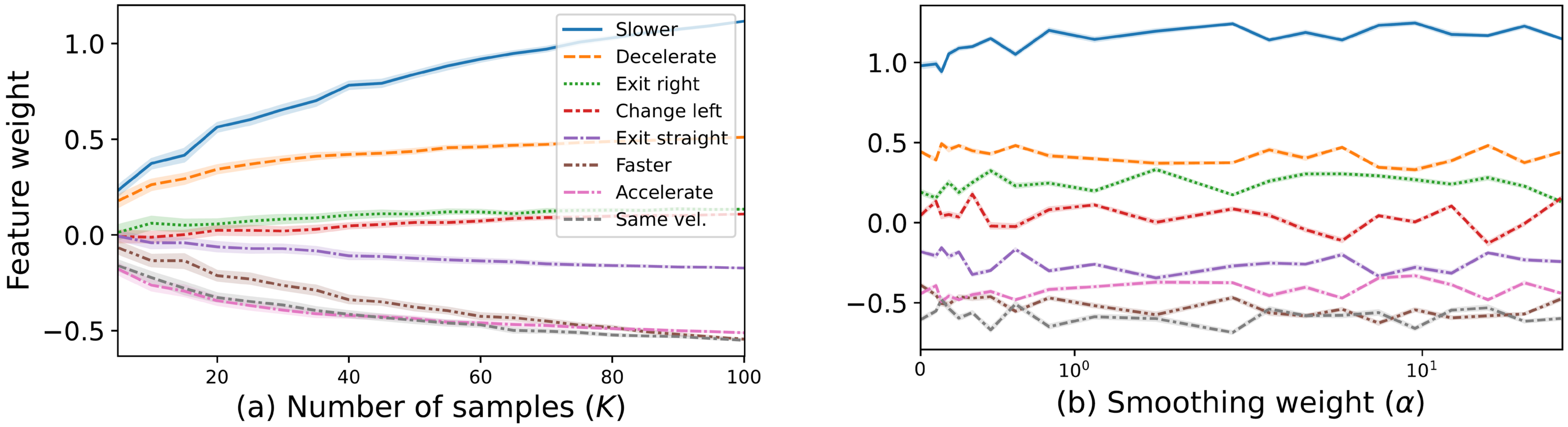
Rollback → Sample → **Correlate**

CEMA – CORRELATE – MECHANISTIC CAUSES



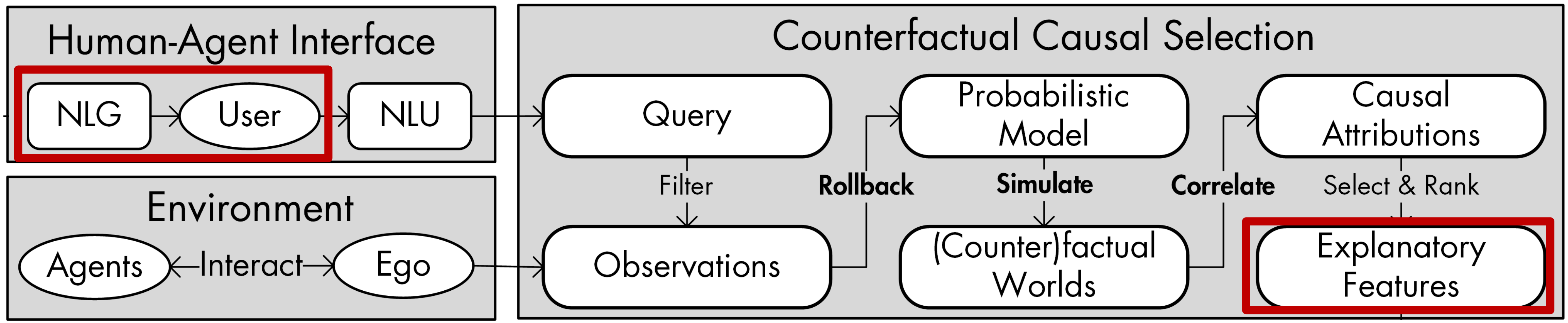
Rollback → Sample → **Calculate**

CEMA is robust

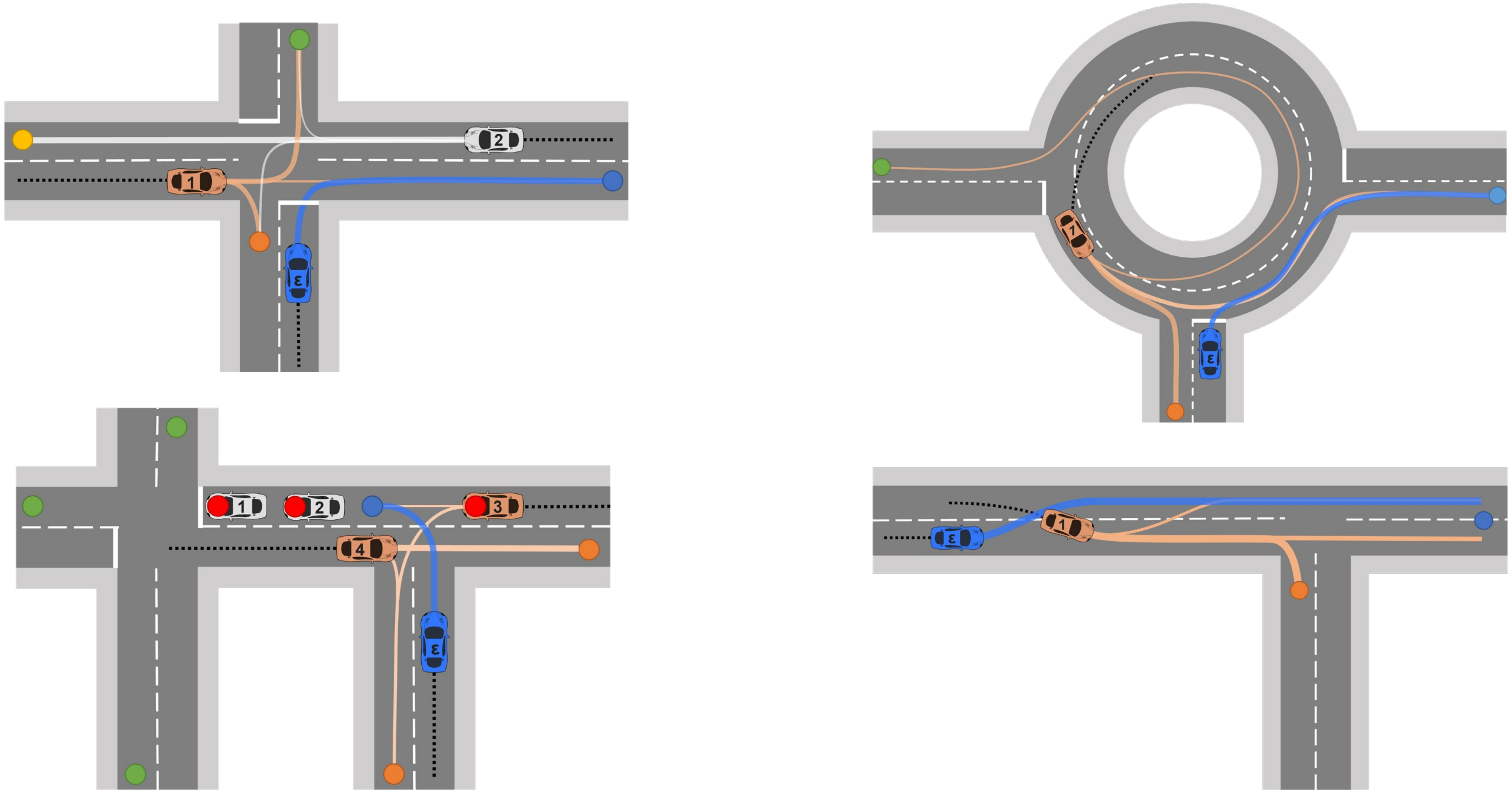


Rollback → Sample → **Calculate**

CEMA – STRUCTURE OF CEMA



FOUR SCENARIOS



CEMA – HUMAN EVALUATION

User study:

Generate human-like explanations at least as good as human-written explanations;

Method:

1. Elicit explanations from people (*HEADD: Human Explanations for AD Decisions* dataset [3]);
2. Compare human explanations to CEMA.

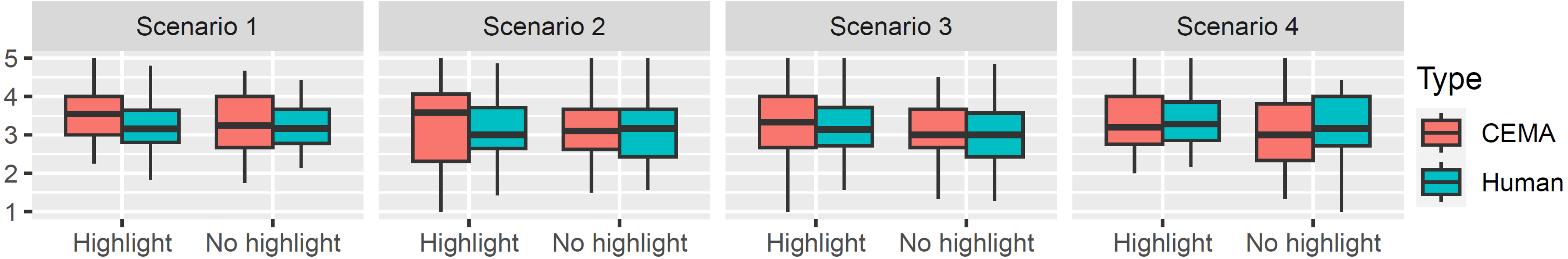
[3] **Gyevnar, B.**, et al. (2024) *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086).

Independent variables:

Scenarios (1-4)

Explanation type (CEMA/Human)

Highlighting CEMA (Y/N)



CEMA – BENEFITS

- **Generally applicable simple (3-step) framework;**
- **No explicit assumption on causal structure:**
  No need to model world with DAGs;
- **Robust causal selection based on CESM;**
- **Works for large number of agents:**
  Tested with up to 20 agents in 4 scenarios.

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

https://arxiv.org/abs/2302.10809

Contributions:

- **CEMA:** General framework for causal explanations of behaviour: Rollback time → Simulate worlds → Correlate variables with outcomes;
- Robustly generated intelligible explanations through natural language.

THE UNIVERSITY *of* EDINBURGH informatics

Autonomous Agents Research Group