Can explainable AI ever be useful for more trustworthy autonomous systems?

Bálint Gyevnár
University of Edinburgh

The Inexplicable Explanation

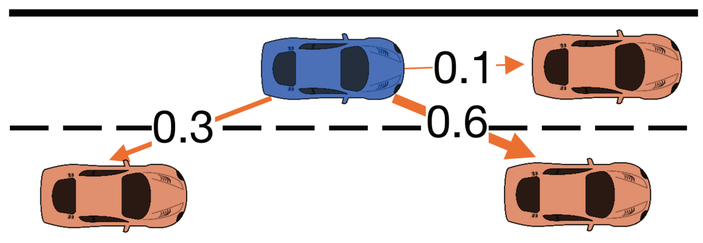
*XAI keeps **failing** people.*

Feature attribution:
Brittle, inconsistent, misleading, interpretation pitfalls;

Causal model:
Explainer's bias, complexity increase, inference time;

Representative/counterfactual example:
Misleading, unactionable;

Surrogate model:
Captures correlation not causation.



Example with lane change prediction: Strategically swapping around the attention weights can leave the prediction unaffected.

A (Transparent) Change in Mindsets

*The purpose of XAI is **misunderstood**.* ***[1]***

Social (broad) view: it is a <u>means</u> to achieve wider values
Human rights *Sustainable innovation* *Accountability*

XAI (narrow) view: it is limited to be an <u>end</u> in itself
Algorithmic understanding *Debugging* *Post-hoc justification*

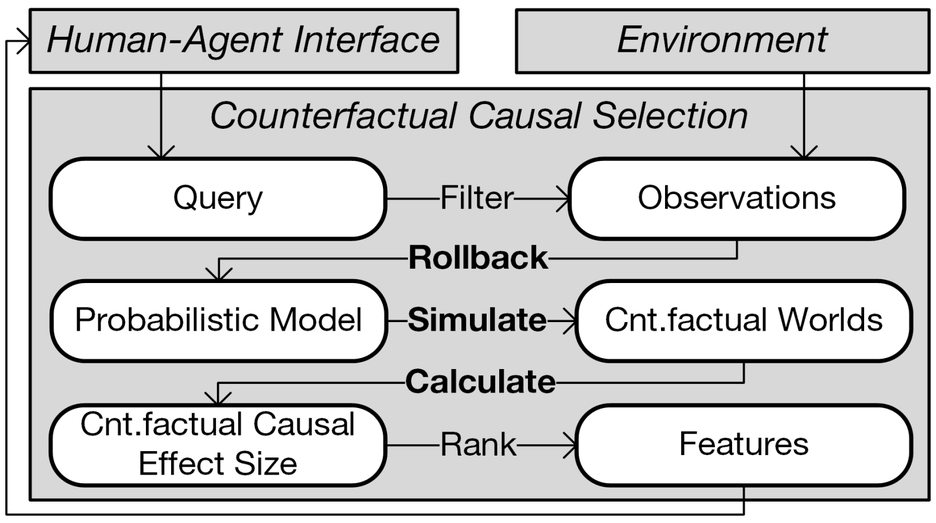
*Can we bridge the **transparency gap**?*

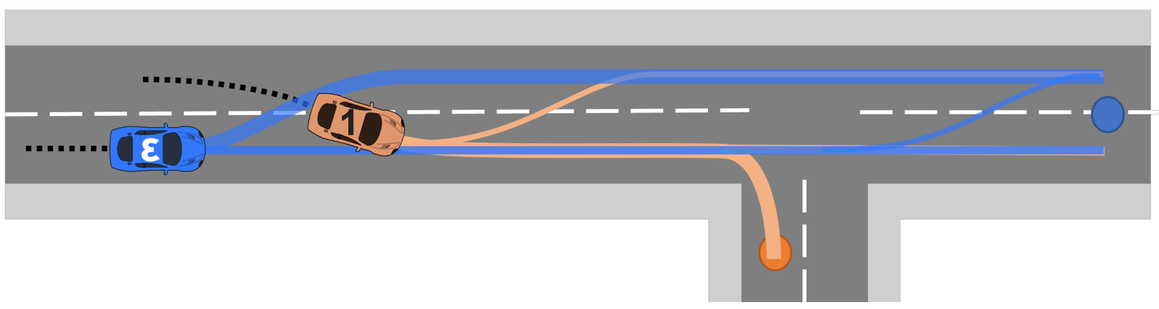
Counterfactual Causality via Simulation

More intuitive explanation based on counterfactual causality. ***[2]***

1. **<u>Simulate</u> counterfactuals grounded in the real world;**
2. **Find variables <u>correlated</u> with outcome across counterfactuals.**

Present the selected causes in natural language within a dialogue.





The blue vehicle is heading to the blue goal. It decides to change lanes after the orange vehicle cuts in front of it and begins to slow down.

[1] Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?.
Balint Gyevnar, Nick Ferguson, Burkhard Schafer; 26th European Conference on Artificial Intelligence (ECAI), October 2023.

[2] Causal Explanations for Sequential Decision-Making in Multi-Agent Systems.
Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, Stefano V. Albrecht; 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2024.

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