

How do we make explainable AI work for people?

Bálint Gyevnár

25 April 2024

In Machine Learning and Modelling Seminar Series at Charles University of Prague



THE UNIVERSITY of EDINBURGH
informatics



Autonomous Agents
Research Group



UK Research
and Innovation

AGENDA

1. What are the **fundamental problems** with current explainable AI (XAI)?
2. How to **operationalise** better explanations for AI-based decision-making?



INTRO – A FIRE IN THE FOREST

Setup:

You are hiking in a forest. In the evening you make a fire to cook food.

Question:

Why is the fire burning?

Answers A:

For cooking food.

Because I collected firewood and lit them with a lighter.

Answers B:

Because firewood is flammable.

Because the moisture content of the firewood was low.

Example based on:

[1] Quillien, T., & Lucas, C. G. (2023, June 8). Counterfactuals and the Logic of Causal Selection. *Psychological Review*. Advance online publication. <https://dx.doi.org/10.1037/rev0000428>



INTRO – A FIRE IN THE FOREST

Setup:

You are hiking in a forest. In the evening you make a fire to cook food.

Question:

Why is the fire burning?

Answers A:

For cooking food.

Because I collected firewood and lit them with a lighter.

Answers B (XAI version):

Concentration of carbon had an importance of 19.84.

Water content in firewood had an importance of 13.37.



Example based on:

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

<https://dx.doi.org/10.1037/rev0000428>

PROBLEMS – A FIRE IN THE FOREST

Requires:

Domain knowledge

Model understanding

Detailed background:

What is carbon concentration?

How is water content defined?

What are the units/scale of these numbers?

How has the model arrived at its decision?

Lacks:

Causality

Interventions

Counterfactuals

Teleology

Recourse

Intelligibility

Actionability or recourse:

What is the relationship between features?

What happens if I change the features?

What if things had been different?

What is the purpose of the fire burning?

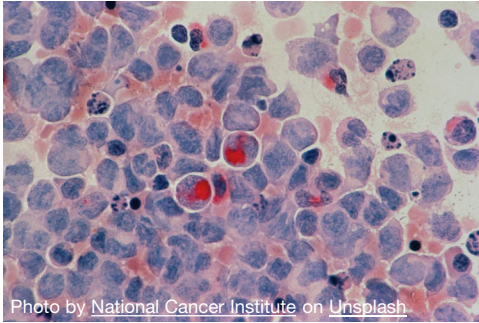
How do I make the fire stop burning?

Can stakeholders understand the explanation?

SOCIALLY/EPISTEMICALLY/SAFETY-CRITICAL SYSTEMS

Safety critical systems are not so forgiving as burning wood in a forest.

Medical diagnoses



Autonomous vehicles



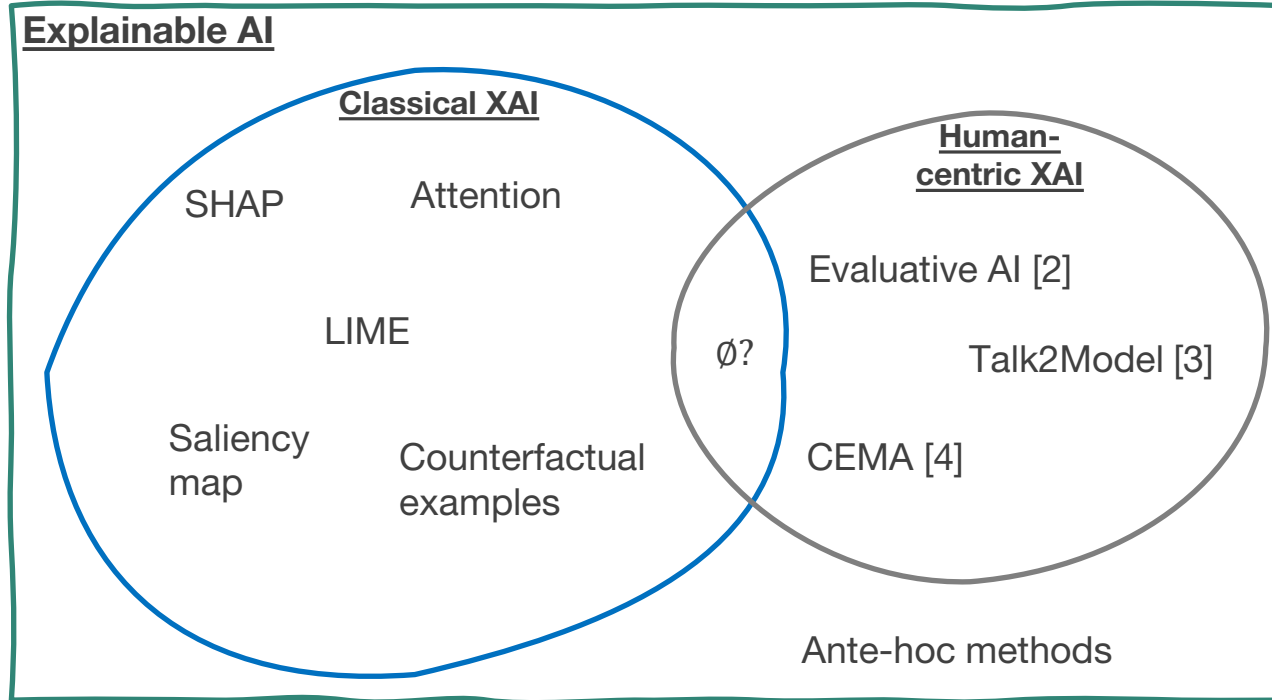
Care robots



Also: manufacturing, power grids, disaster prediction, search & rescue, etc.

Can people trust and rely on XAI for these applications? **NO.**

DISCLAIMER – FOCUS ON CLASSICAL XAI



[2] Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342.

[3] Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 5(8), Article 8.

[4] Gyevnar, B., Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024). Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. In *The 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2024)*.

AGENDA

1. What are the **fundamental problems** with current explainable AI (XAI)?
 - a. Motivation for XAI is misplaced;

1. How to **operationalise** better explanations for AI-based decision-making?



MOTIVATION FOR XAI IS MISPLACED

Big piece of the motivation:

Achieve transparency for the AI system

Why do we want transparency?

Trust, public acceptance, understanding, etc.

How do we achieve transparency?

Classical XAI?



TOWER OF BABEL OF XAI – THE PROBLEM

The Tower of Babel of XAI terms [7]:

Ethics guidelines;
Law (e.g., GDPR, AIA, DSA);
Standards;
Computer science.

Confusing and interchanging terminology:

Slows down progress and communication




The Tower of Babel by Pieter Bruegel the Elder (1563)

[8] Schneeberger, D., Röttger, R., Cabitz, F., Campagner, A., Plass, M., Müller, H., & Holzinger, A. (2023). The Tower of Babel in Explainable Artificial Intelligence (XAI). In Machine Learning and Knowledge Extraction (pp. 65–81). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40837-3_5

TOWER BABEL OF XAI – PROPOSED SOLUTION

Term	Type	Target	Example
Transparent	Emergent	Ecosystem	XAI + user manual
Explainable	Emergent	System	XAI
Interpretable	Inherent	Model	Shallow decision tree
Justifiable	Emergent	Decision	Loan prediction



[5] **Gyevnar, B.**, Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence (pp. 964 - 971). <https://doi.org/10.3233/FAIA230367>



TRANSPARENCY AS AN END

Classical XAI = AI system + Explanation

Classical XAI = ML classification + Post-hoc rationalisation

E.g., SVM + SHAP

The End

implies

Transparency = Explanation

[5] **Gyevnar, B.**, Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence (pp. 964 - 971). <https://doi.org/10.3233/FAIA230367>

BLIND RELIANCE ON BLACK-BOX AI

Black boxes where none belong:

Critical decision-making affecting lives;

Classical XAI is approximation:

Causal chain is not (well) represented;

Assumes black box is always right:

Classical XAI can mislead by justifying incorrect decisions;

More complex decision process:

Now need to debug two systems (AI + XAI).



[6] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), Article 5. <https://doi.org/10.1038/s42256-019-0048-x>

UNACTIONABLE EXPLANATIONS

Classical XAI generates unactionable explanations:

Explanations pick features that are hard to change;

For example:

“You could have received the loan if only you were 185cm tall.”

“Your marital status had the most effect on your recidivism chance.”

[7] Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable Recourse in Linear Classification. Proceedings of the Conference on Fairness, Accountability, and Transparency, 10–19. <https://doi.org/10.1145/3287560.3287566>



TRANSPARENCY AS A MEANS

XAI = AI system + Explanation + Context

Context: stakeholders, communication, deployment lifecycle, model updates, recourse, etc.

A means

Transparency = Explanation + documentation + standardisation + risk assessment + ...

[5] **Gyevnar, B.**, Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence (pp. 964 - 971). <https://doi.org/10.3233/FAIA230367>

MOTIVATION FOR XAI IS MISPLACED

This is nothing new in legal contexts:

Law is always interpreted in context;

In law, transparency itself is a means towards:

Protection of Human Rights;

Sustainable innovation.

[5] **Gyevnar, B.**, Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence (pp. 964 - 971). <https://doi.org/10.3233/FAIA230367>



TRANSPARENCY GAP

Transparency as an end

“The post-hoc rationalisation of post-hoc rationalisations”



Transparency as a means

“Explanations should serve the user not the creator”

Blindly applying XAI methods to ML systems hurts the overall system.

[5] **Gyevnar, B.**, Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence (pp. 964 - 971). <https://doi.org/10.3233/FAIA230367>

TRANSPARENCY DESIGN CHECKLIST

Design checklist:

- ✓ Who will be using the system and where?
- ✓ What effects can the AI decision have on the user?
- ✓ How does the Explanation affect the interpretation of the AI decision?
- ✓ Does the XAI system change if the AI system changes?
- ✓ Does the system handle distribution shifts and OOD examples?
- ✓ What if the AI decision is wrong?

[5] **Gyevnar, B.**, Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence (pp. 964 - 971). <https://doi.org/10.3233/FAIA230367>



AGENDA

1. What are the **fundamental problems** with current explainable AI (XAI)?
 - a. Motivation for XAI is misplaced;
 - b. Standard methods are unreliable;
2. How to **operationalise** better explanations for AI-based decision-making?



Three examples:

1. Misusing Shapley values
2. Misleading saliency maps
3. Brittle counterfactuals

Shapley values:

A method to calculate item value based on average marginal contributions;

Marginal gain from using item i with cost function C :

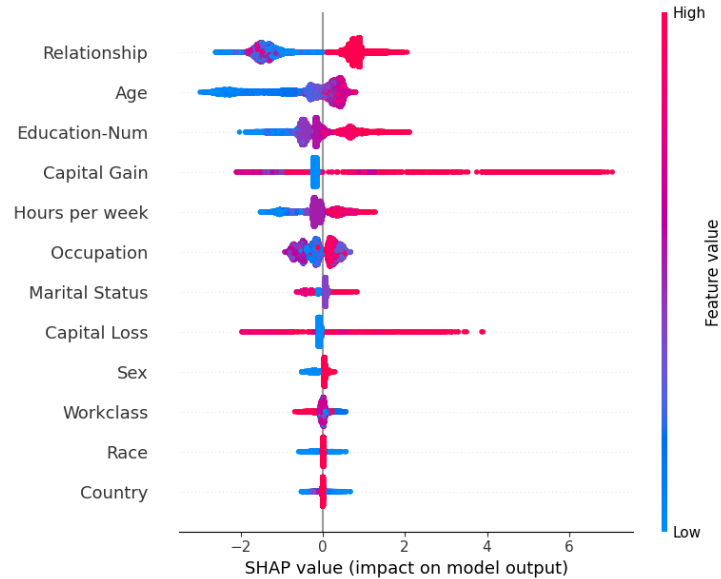
$$\Delta_C(S, i) = C(S \cup \{i\}) - C(S);$$

Average contribution of item i :

$$\phi_i = \sum_{S \in 2^{F \setminus \{i\}}} w(S) \Delta_C(S, i).$$

[9] Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. IEEE Access, 9, 144352–144360. <https://doi.org/10.1109/ACCESS.2021.3119110>

MISUSING SHAPLEY VALUES



Example taken from SHAP [documentation](#).

[9] Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. IEEE Access, 9, 144352–144360. <https://doi.org/10.1109/ACCESS.2021.3119110>

MISUSING SHAPLEY VALUES

Problem 1 (transparency gap):

Not specifically designed for ML feature selection;
Naïvely applying to ML feature selection introduces Problem 2 – 4;

Problem 2 (model averaging):

Redundant features are assigned non-zero influence;

Problem 3 (cost function):

Wrong choice of C will result in wrong explanation;

Problem 4 (correlated features):

Correlated features are assigned similar value though one may be redundant.

[9] Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. IEEE Access, 9, 144352–144360.
<https://doi.org/10.1109/ACCESS.2021.3119110>



HIGHLIGHTING MISLEADING SALIENCY MAPS

Saliency maps:

Feature importance for high-dimensional inputs;

Adversarial attacks:

Keep output same with different saliency map.

Arbitrary and cherry-picked interpretations:

Saliency maps are difficult to interpret;

Interpretations can be irrelevant.

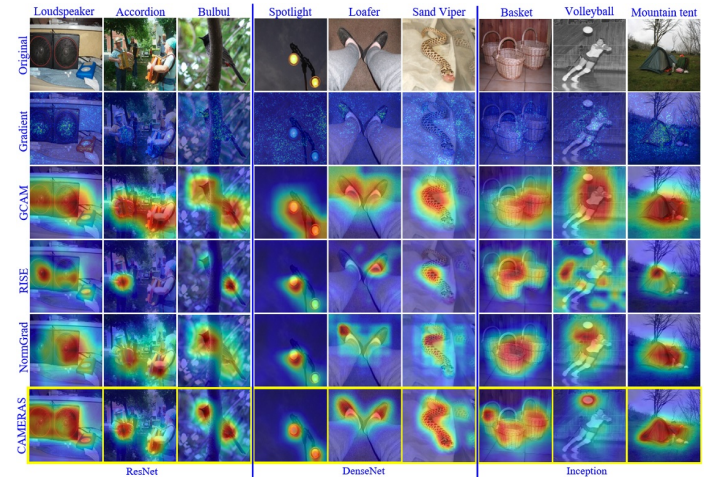


Figure from Jalwana *et al.* [11].

[10] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Haradt, M., & Kim, B. (2020). *Sanity Checks for Saliency Maps* (arXiv:1810.03292). arXiv:1810.03292 <https://doi.org/10.48550/arXiv.1810.03292>

[11]. Jalwana, M. A. A. K., Akhtar, N., Bennamoun, M., & Mian, A. (2021). CAMERAS: Enhanced Resolution And Sanity preserving Class Activation Mapping for image saliency. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16322–16331. <https://doi.org/10.1109/CVPR46437.2021.01606>

BRITTLE COUNTERFACTUALS

Brittle explanations:

A small change in input leads to different CE;

Sensitive to distance metric:

Different metrics also have different interpretations;

CE can function as adversarial examples:

Possible to game the system with information of CE.

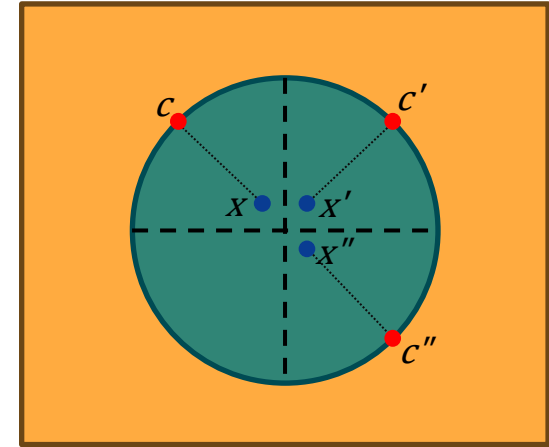


Figure 1 reproduced from Leofante & Potyka [14]. Teal circle represents the decision boundary in a binary classification setting.

[7] Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable Recourse in Linear Classification. Proceedings of the Conference on Fairness, Accountability, and Transparency, 10–19. <https://doi.org/10.1145/3287560.3287566>

[14] Leofante, F., & Potyka, N. (2024). Promoting Counterfactual Robustness through Diversity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19), Article 19. <https://doi.org/10.1609/aaai.v38i19.30127>

USE WITH CARE

This is not to say that we should never use these methods. But:

- Need to be very careful when applying them;
- Use critical thinking and know the limits of the method;
- Don't just shove XAI at everything;
- Use meaningful evaluation.



AGENDA

1. What are the **fundamental problems** with current explainable AI (XAI)?
 - a. Motivation for XAI is misplaced;
 - b. Standard methods are unreliable;
 - c. XAI evaluation is flawed;
2. How to **operationalise** better explanations for AI-based decision-making?

Evaluation in XAI is flawed:

1. Not having a strong motivation for evaluation;
2. Quantitative evaluation is ill-posed;
3. User studies are badly designed.

Ask the question: what is the purpose of my evaluating XAI?

Often missing a strong motivation:

Results from the transparency gap;

Explanation for the sake of explanation cannot be meaningfully evaluated;

Motivation and evaluation is not compatible:

Trust, transparency, etc. often mentioned as motivation;

These must not just be lofty long-distance goals;

WHY DO WE EVALUATE?

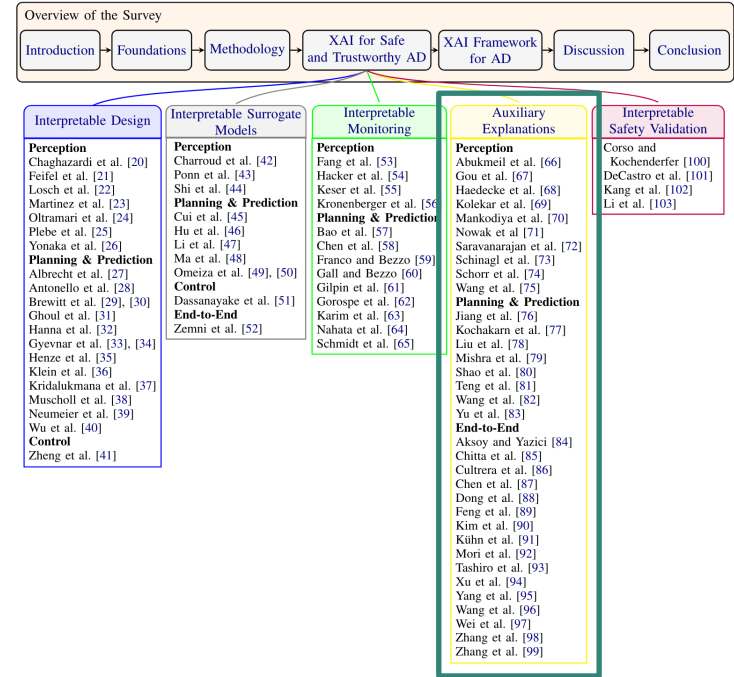
Example: Recent review of XAI for autonomous driving

Majority of methods are “auxiliary explanations”;

XAI (almost always) functions as an afterthought;

Usually, the process goes:

1. Take existing computational problem (e.g., object detection, intention prediction);
2. Train some novel model with some appendix layer;
3. Get better number and interpret results;
4. Write about importance of transparency in intro.



[15] Kuznietsov, A., **Gyevnar, B.**, Wang, C., Peters, S., & Albrecht, S. V. (2024). *Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review* (arXiv:2402.10086). arXiv. <http://arxiv.org/abs/2402.10086>



What even is a correct explanation?

No such thing as a ground truth explanation:

Otherwise, we are just doing classification;

Fidelity:

Degree to which an explanation represents the decision-making process faithfully (but not necessarily completely);

Fidelity is NOT all you need:

You don't need to explain all factors that affected the decision;
Different explanations work better for different people and contexts.

Qualitative evaluation is essential to show:

System produces reasonable explanations;

Where and when the system fails;

Understand properties of your data;

Many papers don't have any qualitative evaluation.

USER STUDIES ARE BADLY DESIGNED

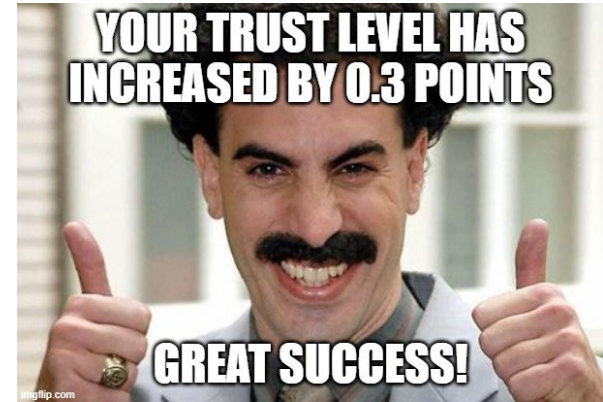
User studies are standard practice;

Based on strong assumptions:

- People need explanations;
- People engage with explanations;
- People understand domain;

With the wrong methods and goals:

- Trust improvement;
- Perceived quality and understanding;
- One-shot testing.



[2] Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342.

[16] Miller, T. (2022). *Are we measuring trust correctly in explainability, interpretability, and transparency research?* (arXiv:2209.00651). arXiv. <http://arxiv.org/abs/2209.00651>

USER STUDIES CHECKLIST

- ✓ **Focus on calibrating trust:**
 - According to the capabilities of the system;
 - User mustn't blindly trust the system;
- ✓ **Use both subjective and objective measures:**
 - Self-reporting and Likert-scales;
 - Observational measures (performance, reliance, failure prediction);
- ✓ **Perform iterative evaluation:**
 - Build explanations incrementally;
 - Explanations alter mental models.



USER STUDIES ARE BADLY DESIGNED

- ✓ **Ask whether you need a user study in the first place:**
 - Depends on stakeholders;
 - Don't treat it as gospel;
- ✓ **Let people explore the model:**
 - Propose hypotheses;
 - Interactive visualisations;
- ✓ **Think carefully about stakeholders:**
 - Explanations don't exist in a vacuum;
 - Remember the transparency gap.



AGENDA

1. What are the **fundamental problems** with current explainable AI (XAI)?
 - a. Motivation for XAI is misplaced;
 - b. Standard methods are unreliable;
 - c. XAI evaluation is unrevealing;
2. How to **operationalise** better explanations for AI-based decision-making?
 - a. Multi-agent systems (MAS);



MULTI-AGENT SYSTEMS (MAS)

What is a multi-agent systems?

Environment (actions, observations, states);
Agents (goals, rewards, policies);
Communication;
E.g., autonomous driving;

Humans can be modelled as agents;

Explanation is multi-agent interaction:

Explainer: XAI agent

Explainee: Human

Action: Request/give explanation

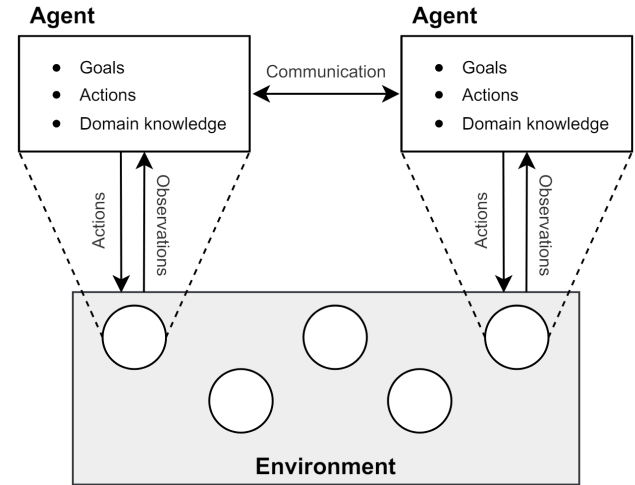


Figure by Albrecht *et al.* [17].

[17] Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press. <https://www.marl-book.com>

Why focus on MAS?

With $n = 1$ agent, reduces to classical XAI;

Otherwise:

Coupled interactions;

Conflicting goals;

Partial observability;

Communication;

Difficult to explain, even for humans.

Example domain:

Autonomous driving (AD);

Critical environment:

Socially: Driving actions are seen and judged by others;

Epistemically: Partial observability and shared rules;

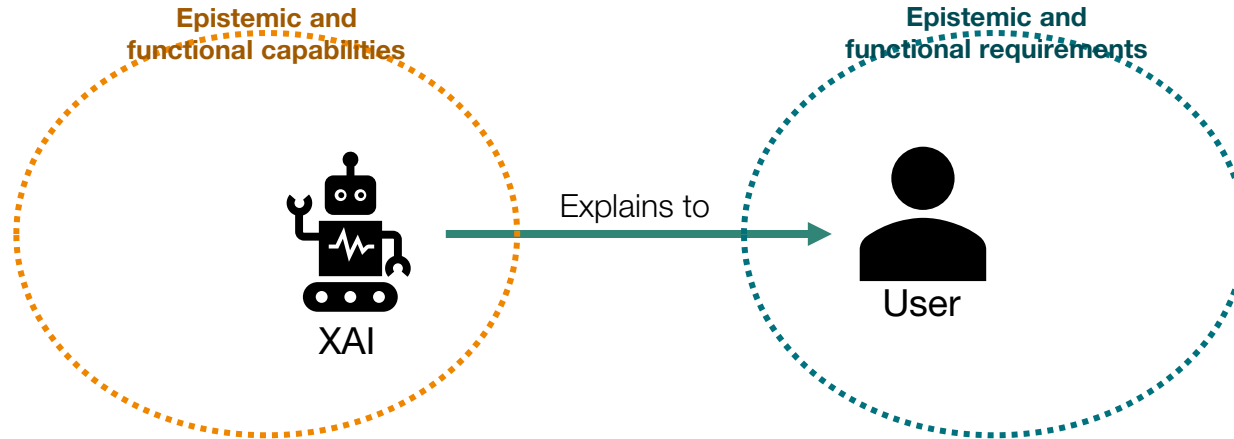
Safety: Driving can be dangerous.

AGENDA

1. What are the **fundamental problems** with current explainable AI (XAI)?
 - a. Motivation for XAI is misplaced;
 - b. Standard methods are unreliable;
 - c. XAI evaluation is unrevealing;
2. How to **operationalise** better explanations for AI-based decision-making?
 - a. Multi-agent systems (MAS);
 - b. People explain all the time; study how they do it in context;

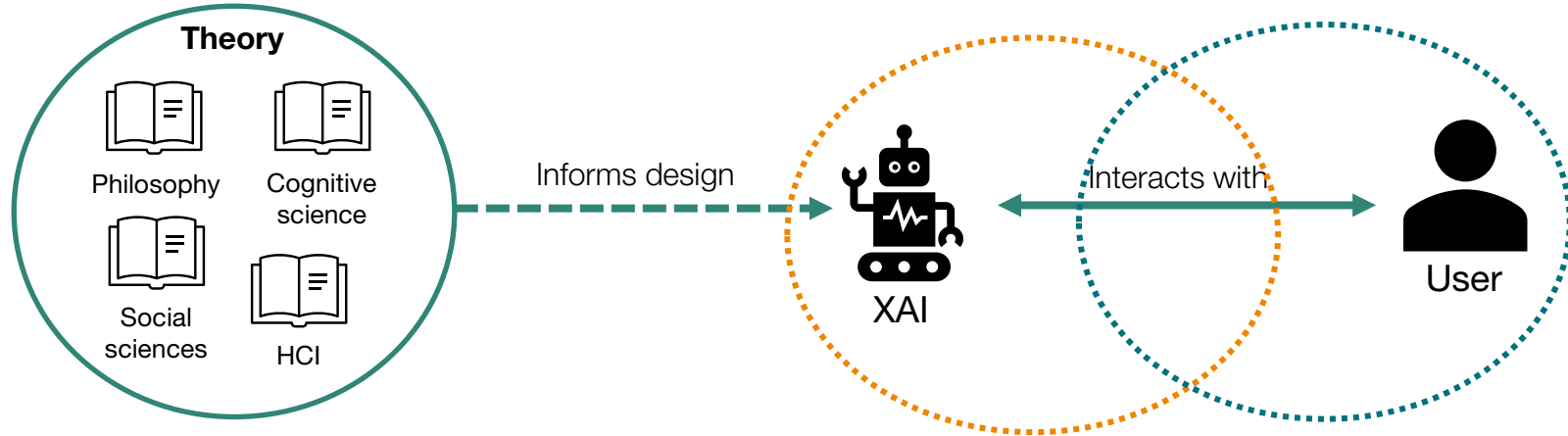


Classical XAI



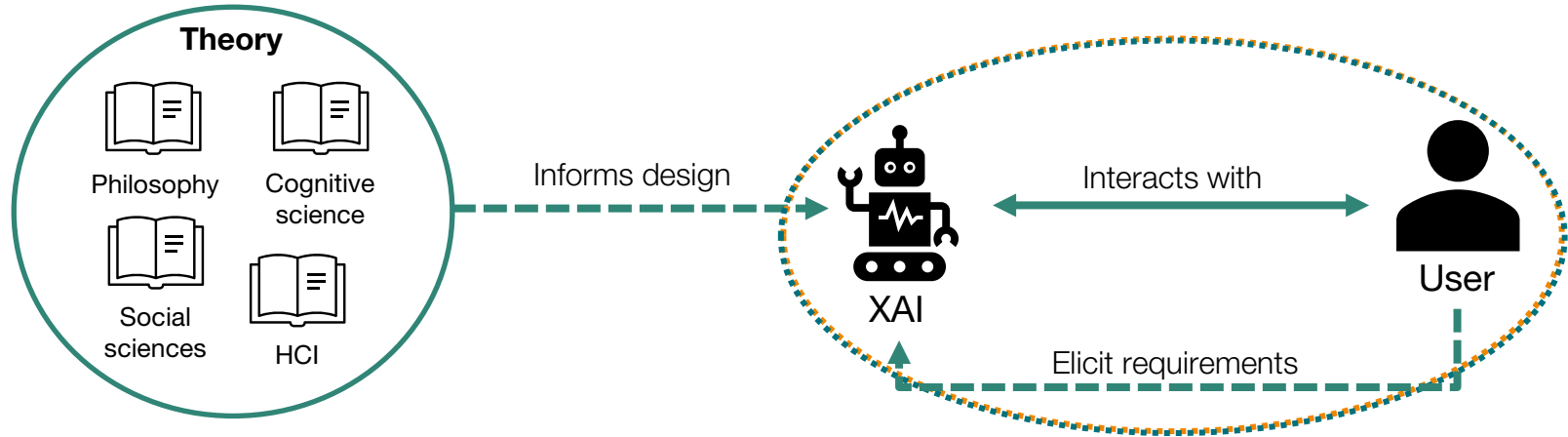
No (or very little) overlap between what XAI offers and what the user wants.

Towards human-centric XAI



Theories about how humans explain help bring capabilities and requirements closer;
But the contexts of the domain and user are left unaddressed.

Human-centric XAI



Elicit user requirements and expectations;

Allows us to confirm theory and tailor XAI to user requirements.

HOW TO ELICIT REQUIREMENTS?

But how can we ask users what sort of explanations they want?

Shouldn't ask them to rate existing explanations:

- Introduces bias;
- Restricts possible space of explanations;
- Potentially, a lot of ratings needed;

1. Ask them to write explanations themselves:

- Still possible to instruct them;
- Can interpret answers in theoretical framework;
- Gives more variety;

2. Then evaluate these explanations.



HEADD

Human Explanations for Autonomous Driving Decisions

[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior.* (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



WHY CAUSAL EXPLANATIONS?

Theory tells us that explanations should be:

Causal;

Contrastive;

Selected;

Conversational;

[19] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
<https://doi.org/10.1016/j.artint.2018.07.007>



According to Aristotle, causal explanations have two¹ main modes:

Teleological:

Focus on the purpose the actions are meant to achieve

Mechanistic:

Focus on the mechanism that gave rise to the action.

¹Technically, Aristotle defines 4 explanatory modes, but only the above two are relevant to explaining behaviour in MAS.

[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



14 unique scenarios with different driving behaviour;

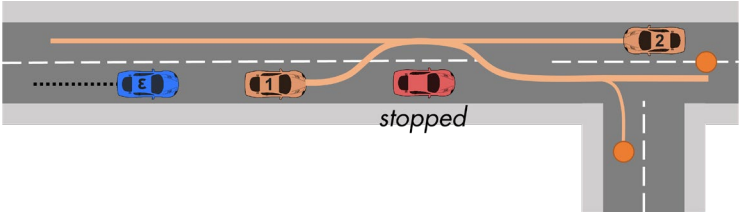
1,300+ human-written explanations;

**4 explanatory modes
(descriptive, teleological, mechanistic, counterfactual);**

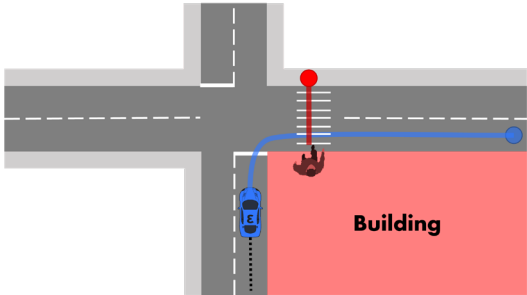
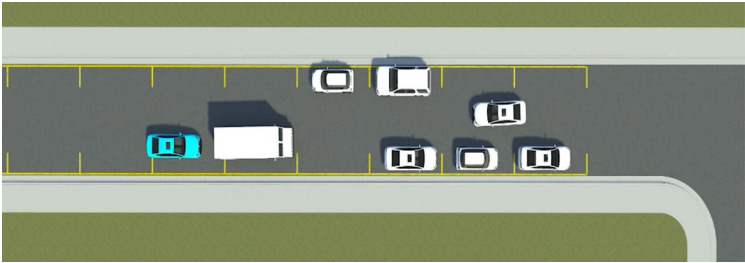
4,000+ evaluations.

[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>

HEADD – EXAMPLE SCENARIOS



Scenario 12



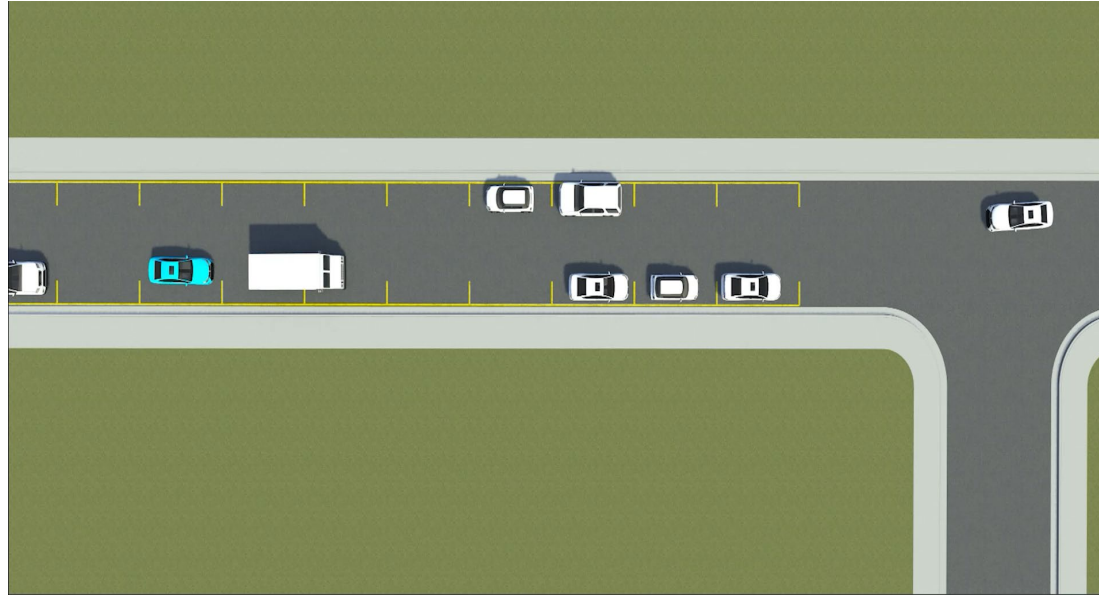
Scenario 8



[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior.* (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



HEADD – EXAMPLE SCENARIOS



[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior.* (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>

HEADD – EXAMPLE EXPLANATIONS

“The blue car was defensive. It could have overtaken the truck while the truck was waiting which could have resulted in an accident with the car approaching from the opposite side.”

(counterfactual)

“The blue car was influenced by the truck ahead of it and therefore, slowed down to wait.”

(mechanistic)

“Since there were cars parked on both sides of the street, neither direction had the legal right of way. They needed to cooperate with drivers in the other direction.”

(teleological)

“The blue self-driving car slowed down and waiting for a parked car to come out of its space, went into the passing lane, then continued to turn into another street.”

(descriptive)

[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior.* (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



At least 5 annotators for each (non-descriptive) explanations:

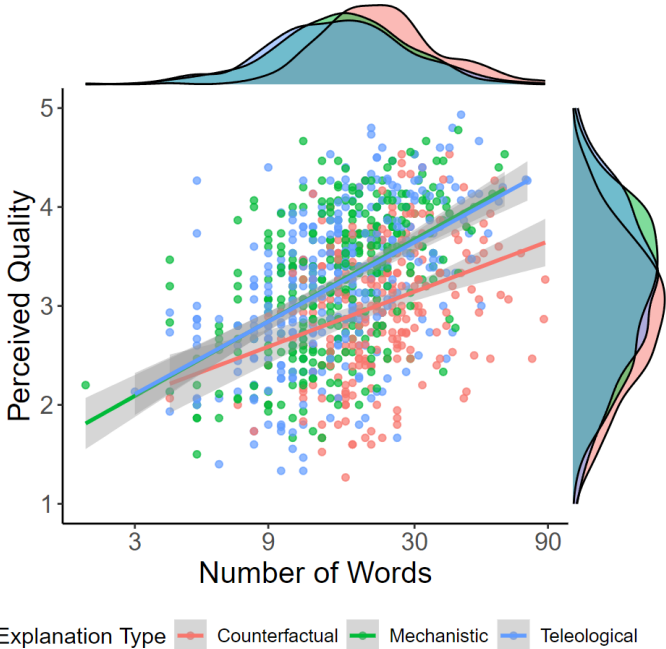
Degree of teleology and mechanistic focus;

Perceived number of causes;

Measures of completeness, sufficiency, and trustworthiness.

[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>

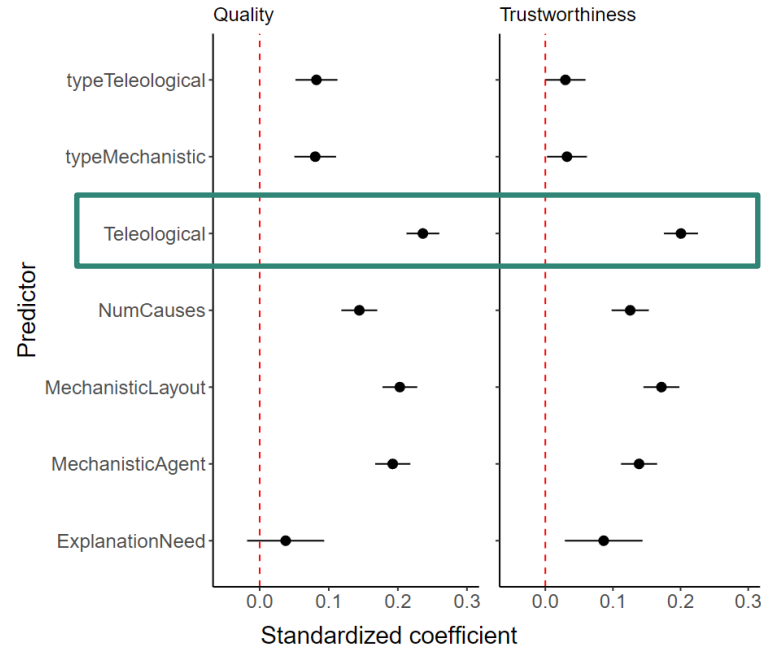
HEADD – INSIGHTS FROM THE COGNITIVE SCIENCES



[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior.* (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



HEADD – INSIGHTS FROM THE COGNITIVE SCIENCES



[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>

HEADD – TELEOLOGY IS BEST PREDICTOR OF QUALITY

Teleological explanations best predict quality and trustworthiness;

Most of XAI focuses on mechanistic explanations;

It is important to consider explanations in terms of the goals and purpose of agents.

HEADD – HUMAN OR AGENT? DOESN'T MATTER

Why did the blue car change lanes?

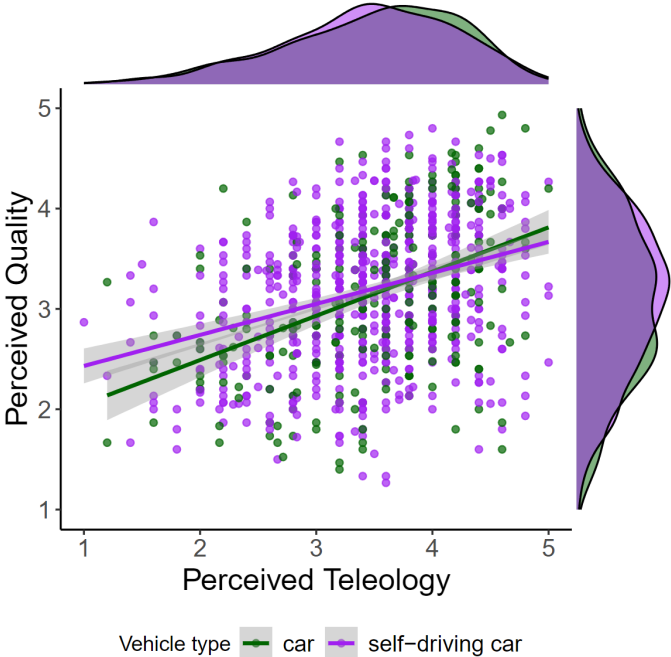


Why did the blue autonomous vehicle choose the change lane action?

[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



HEADD – HUMAN OR AGENT? DOESN'T MATTER



[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior.* (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



Doesn't matter whether human or machine;

People ascribe teleological concepts to explanations anyway.

[18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086). arXiv. <https://doi.org/10.48550/arXiv.2403.08828>



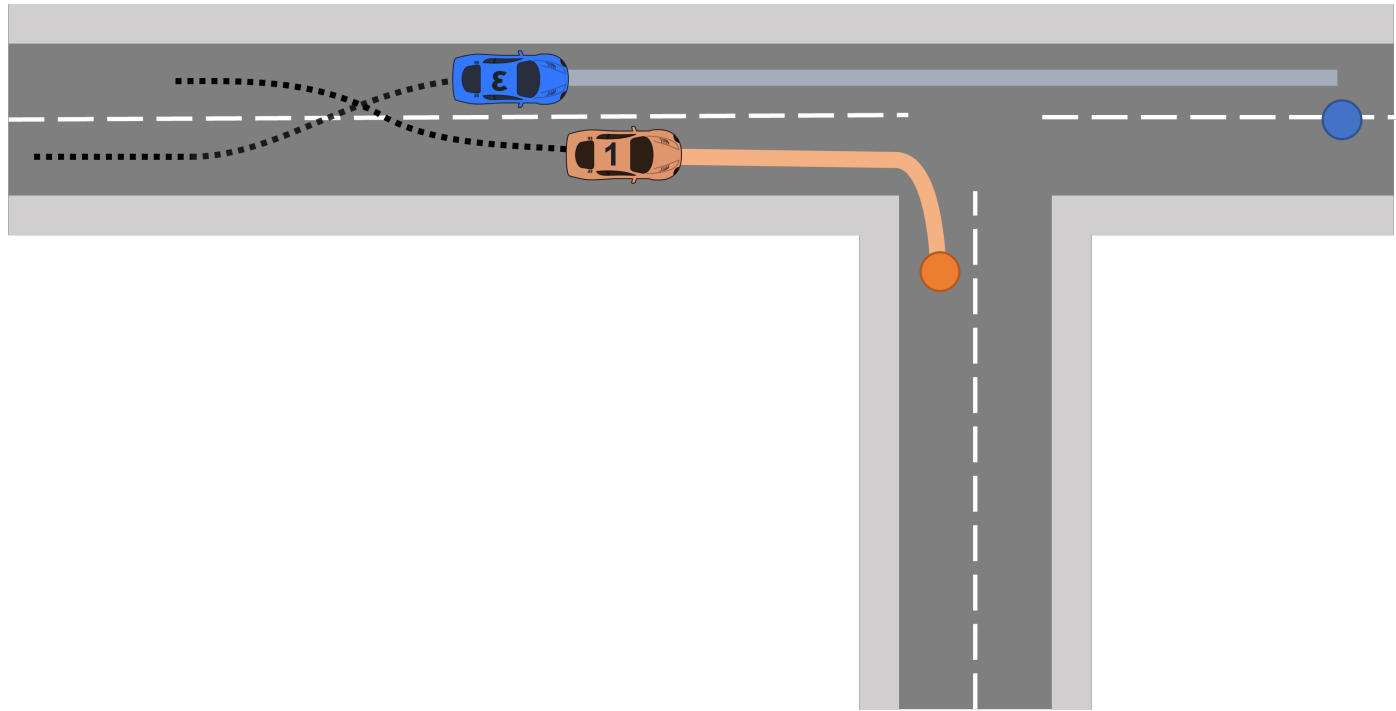
AGENDA

1. What are the **fundamental problems** with current explainable AI (XAI)?
 - a. Motivation for XAI is misplaced;
 - b. Standard methods are unreliable;
 - c. XAI evaluation is unrevealing;

2. How to **operationalise** better explanations for AI-based decision-making?
 - a. Multi-agent systems (MAS);
 - b. People explain all the time; study how they do it;
 - c. Causal Explanations for Sequential Decision-Making in MAS.

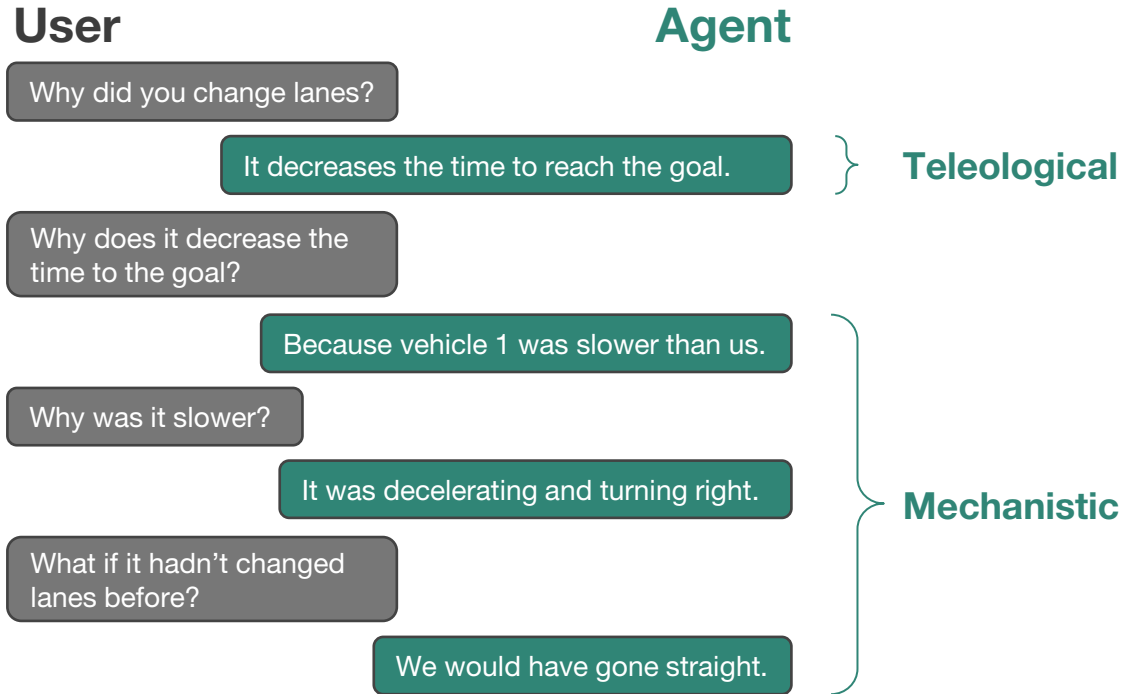


ILLUSTRATIVE SCENARIO



[20] Albrecht, S. V., Brewitt, C., Wilhelm, J., **Gyevnar, B.**, Eiras, F., Dobre, M., & Ramamoorthy, S. (2021, March 15). Interpretable Goal-based Prediction and Planning for Autonomous Driving. *IEEE International Conference on Robotics and Automation (ICRA)*. <https://ieeexplore.ieee.org/document/9560849>

EXAMPLE INTERACTION



[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

CEMA

Causal Explanations for Multi-Agent Systems

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>



WHY USE CEMA?

Applicable whenever you have:

- Probabilistic model to predict future actions of others;
- No explicit assumptions on causal structure;

Provides:

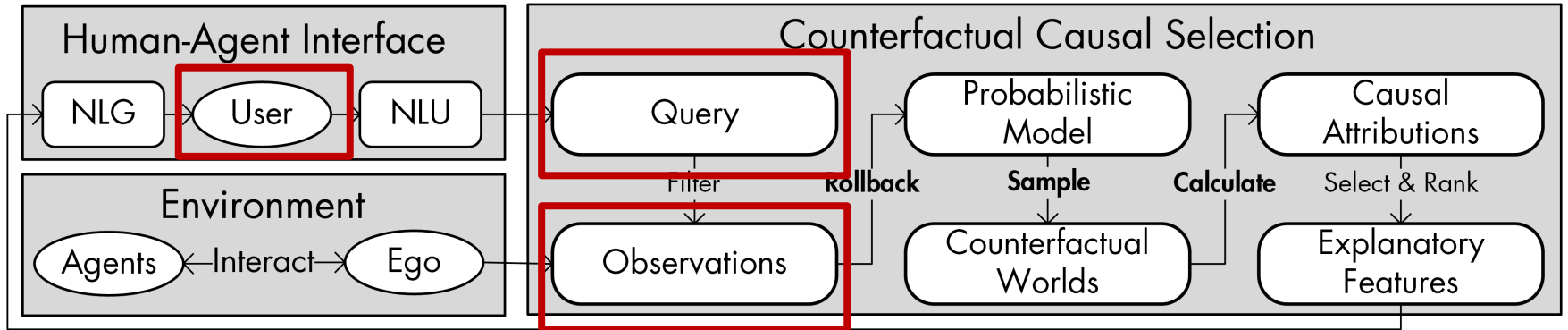
- Contrastive, causal, and selected explanations;

Designed for interaction.

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>



STRUCTURE OF CEMA



[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

Rollback → Sample → Correlate

Counterfactual Effect Size Model (CESM)

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>



Based on how people might select causes to explain;

People simulate counterfactuals to select causes:

Using some prior (cognitive) distribution;

But anchored to observations;

People use correlation to select among causes:

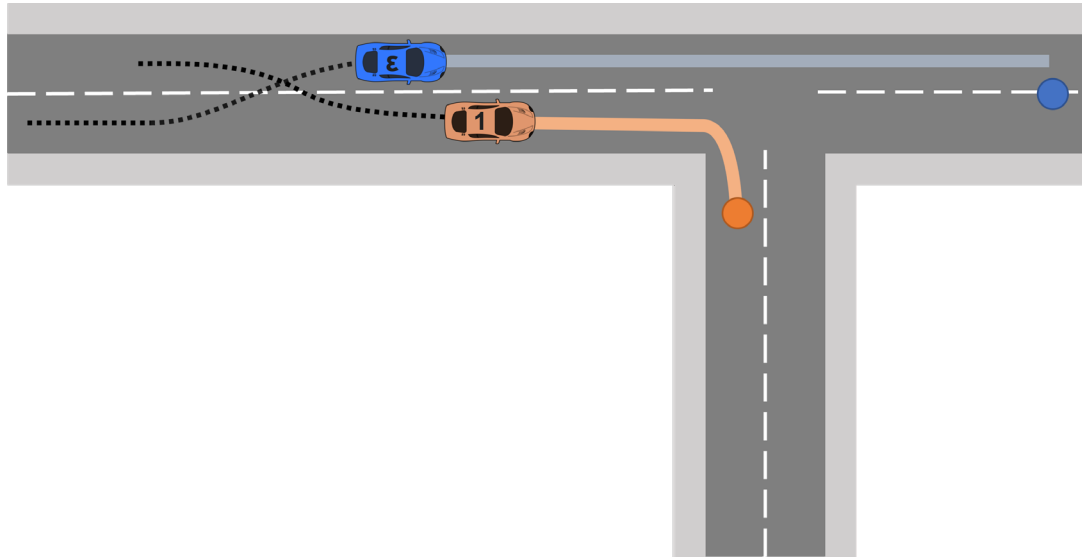
C caused E if C is highly correlated with E across counterfactuals.

[1] Quillien, T., & Lucas, C. G. (2023, June 8). Counterfactuals and the Logic of Causal Selection. *Psychological Review*. Advance online publication. <https://dx.doi.org/10.1037/rev0000428>

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

ROLLBACK

Observed trajectory: $s_{1:t}$

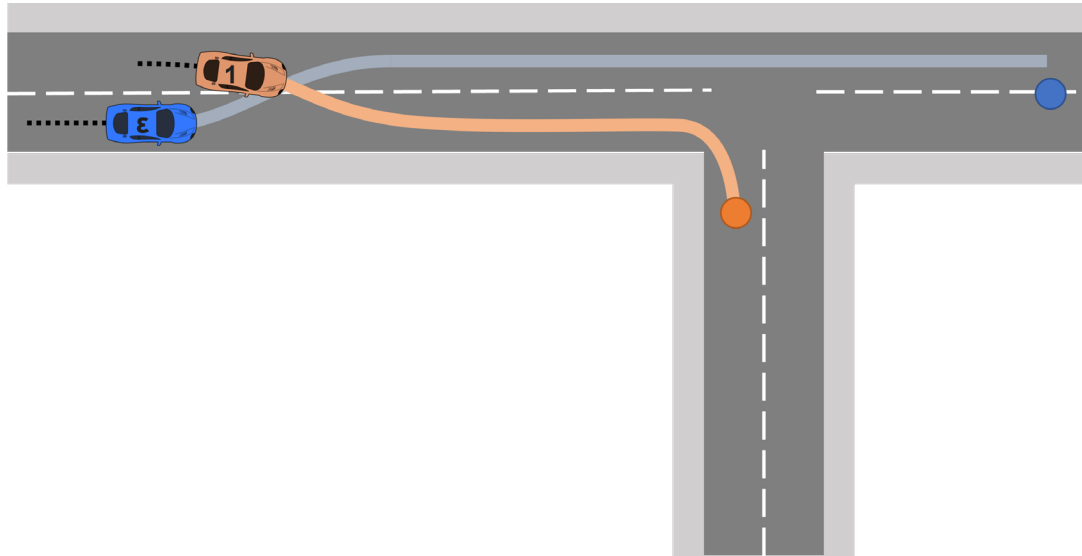


Rollback → Sample → Calculate

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

ROLLBACK

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

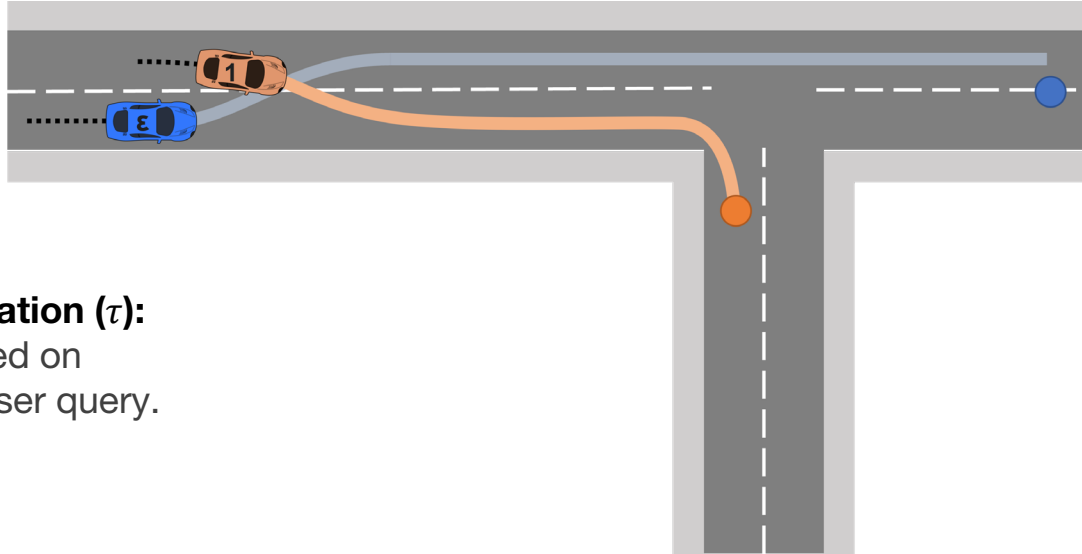


Rollback → Sample → Calculate

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

ROLLBACK

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



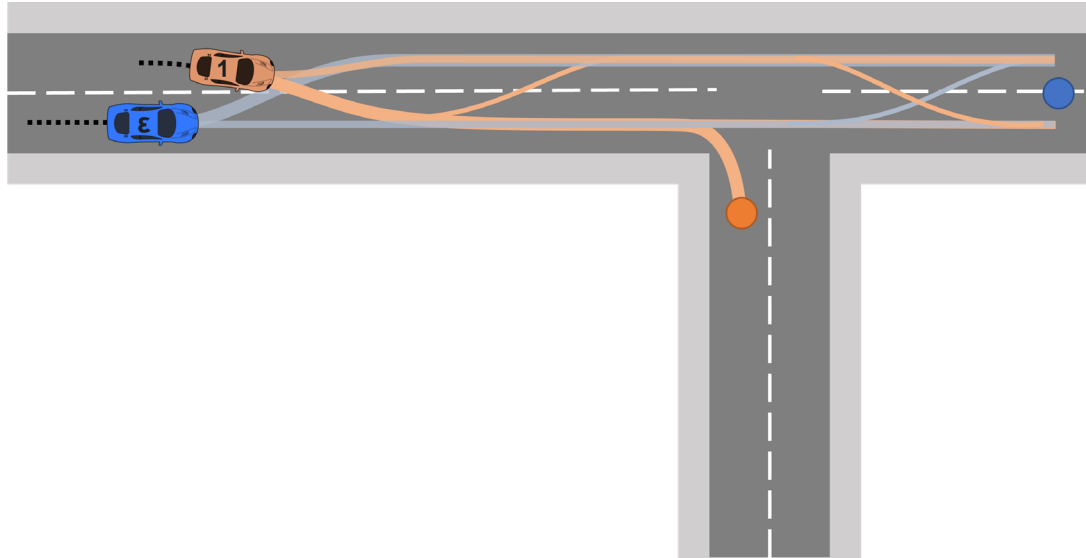
Rollback duration (τ):
Selected based on
actions and user query.

Rollback → Sample → Calculate

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

SAMPLE (COUNTER)FACTUALS

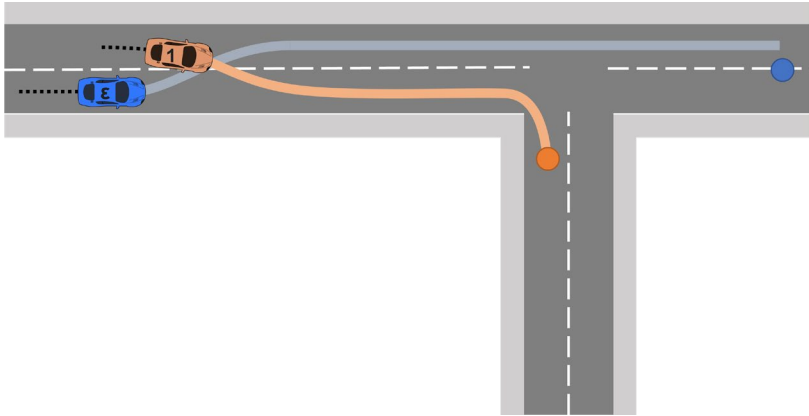
Sample (counter)factual worlds $\sim p(s_{\tau:n} | s_{1:\tau})$



Rollback \rightarrow **Sample** \rightarrow Calculate

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

SAMPLE (COUNTER)FACTUALS



Action presence (y):

Lane change (1)

Intrinsic rewards (r):

Time-to-goal: 5 s

Jerk: 0.2 m/s^3

Collision: No

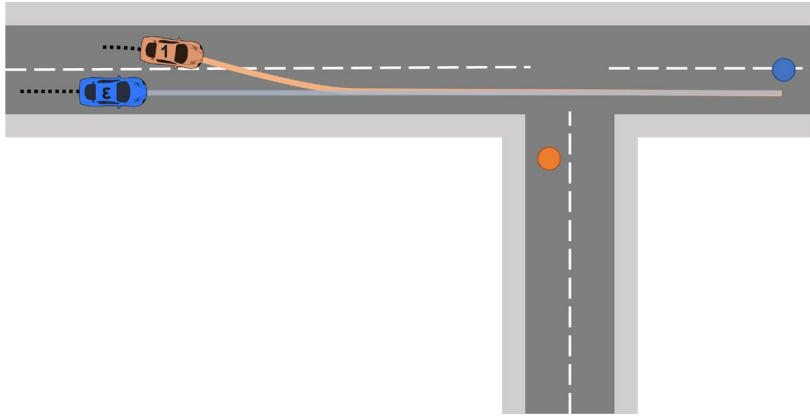
Features for trajectory:

{Decelerate, Turn, Slower, etc...}

Rollback → **Sample** → Calculate

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

SAMPLE (COUNTER)FACTUALS



Action presence (y):

No lane change (0)

Intrinsic rewards (r):

Time-to-goal: 10 s

Jerk: 0.7 m/s^3

Collision: No

Features for trajectory:

{Accelerate, Continue, Faster, etc...}

Rollback → **Sample** → Calculate

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

CALCULATE – SOME NOTATION

$$\mathbf{r} = \begin{bmatrix} \text{time-to-goal} \\ \vdots \\ \text{curvature} \end{bmatrix}$$

Reward component vector

$$\mathcal{D} = \{(s_{\tau:n}^k, y^k, \mathbf{r}^k)\}_{k=1}^K$$

Sampled set of worlds

$$\mathcal{X} = \{\mathbf{r} | y^k \in \mathcal{D} \wedge y_q = y^k\}$$

Sampled worlds where
query occurred

Rollback → Sample → **Calculate**

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>



CALCULATE – TELEOLOGICAL CAUSES

$$\Delta = \mathbb{E}_{\mathcal{X}}[\mathbf{r}] - \mathbb{E}_{\bar{\mathcal{X}}}[\mathbf{r}]$$

Expected difference of rewards between:

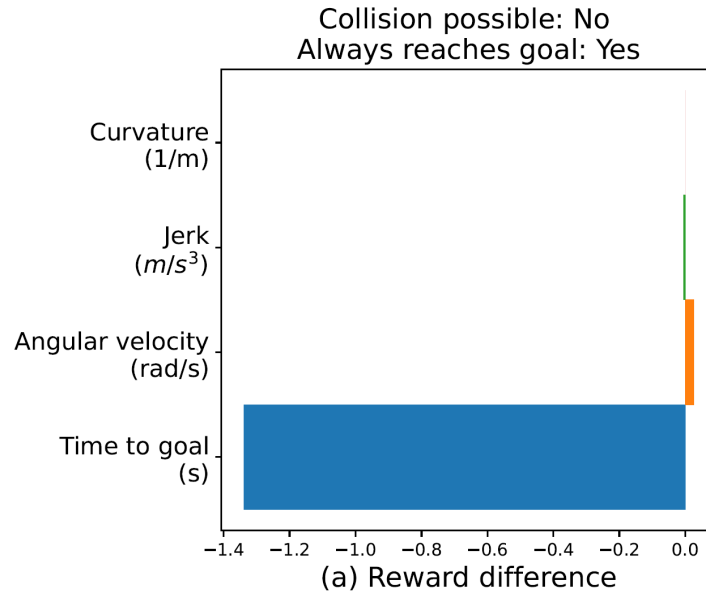
Worlds where query happened;
Where query did not happen.

Rollback → Sample → **Calculate**

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>



CALCULATE – TELEOLOGICAL CAUSES



Rollback → Sample → **Calculate**

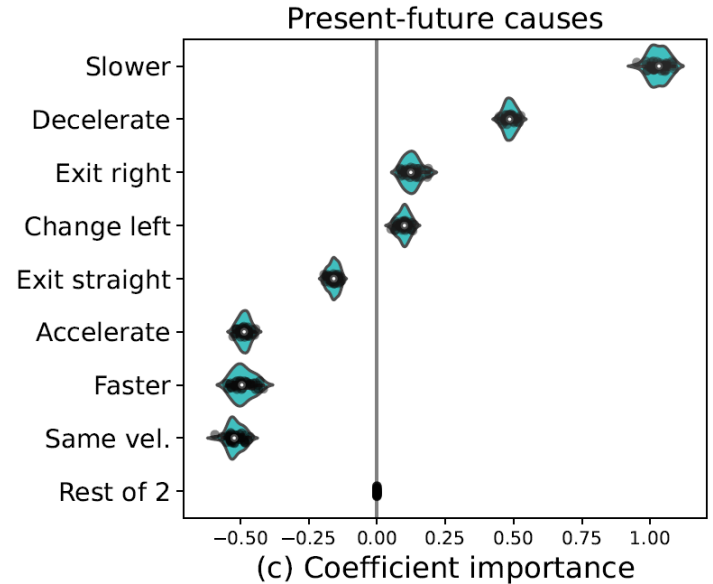
[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

CALCULATE – MECHANISTIC CAUSES

Fit interpretable model to trajectory features:

- Predict y_q from features;
- Extract feature importance;

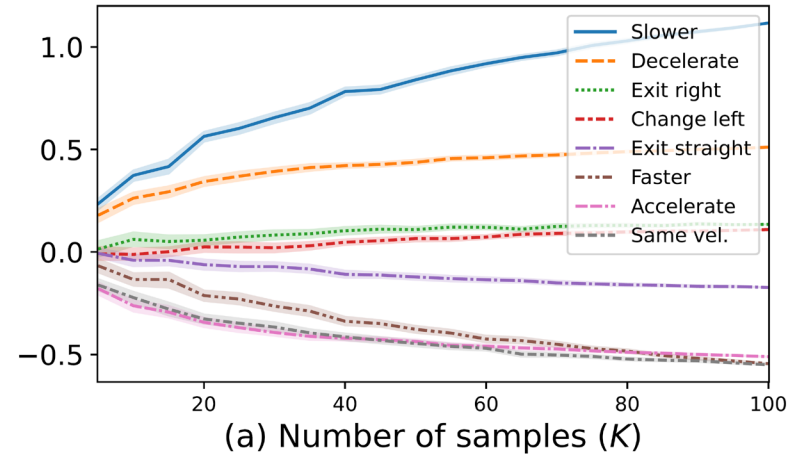
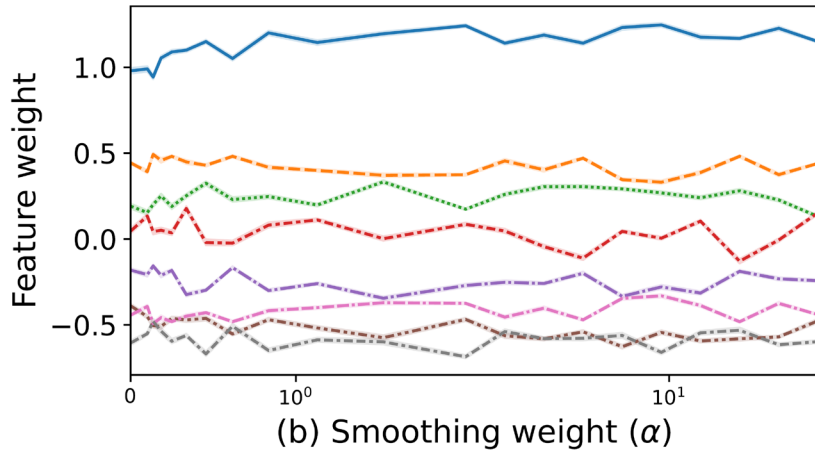
Counterfactual effect size.



Rollback → Sample → **Calculate**

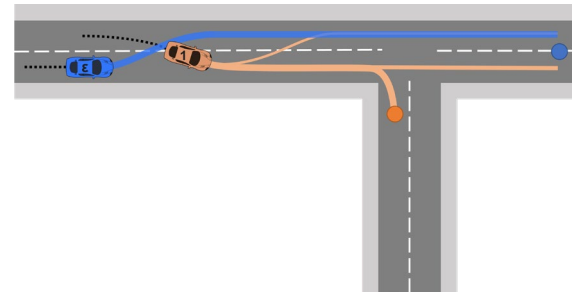
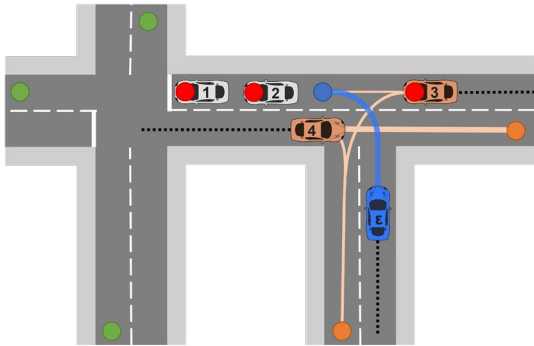
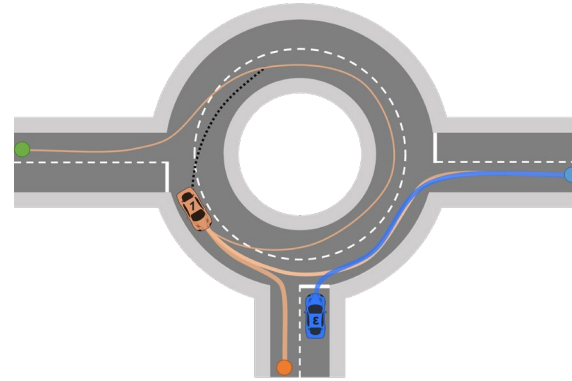
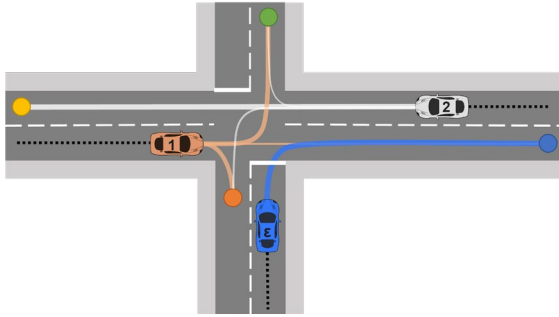
[4] Gyevnar, B., Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

ROBUSTNESS – MECHANISTIC CAUSES



[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

FOUR SCENARIOS



[4] Gyevnar, B., Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

Motivation:

Create intelligible explanations for humans;

Goal:

Generate human-like explanations that people find high quality;

Method:

1. Elicit explanations from people (HEADD);
2. Compare human explanations to CEMA.

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

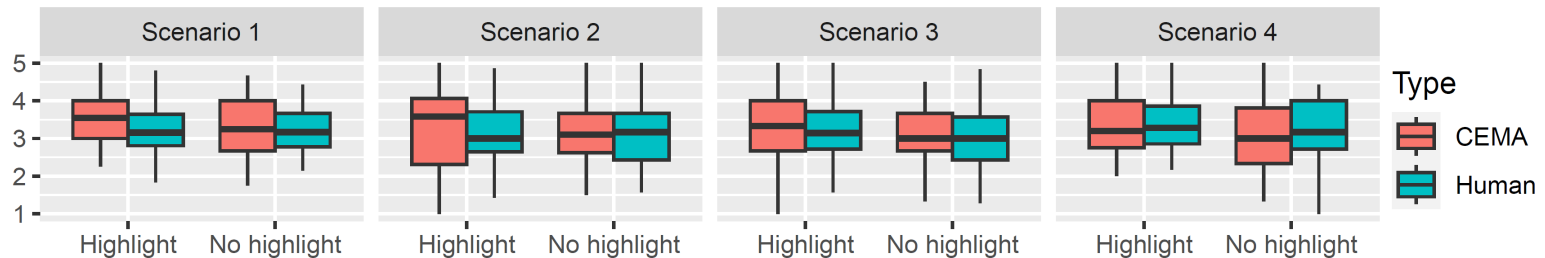


Independent variables:

Scenario (1 - 4)

Explanation type (CEMA/Human)

Highlighting CEMA (Y/N)



[4] Gyevnar, B., Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>

BENEFITS

- ✓ **Generally applicable simple interactive 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.

[4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2024, May). *Causal Explanations for Sequential Decision-Making in Multi-Agent Systems*. 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). <https://doi.org/10.48550/arXiv.2302.10809>



WHAT IS NEXT?

Better natural language generation;

More interactive evaluation;

Learning to explain with guarantees in MAS;

Explanations for more optimal planning in large MAS.



TAKEAWAYS

- **XAI does not live in a vacuum.**
Ask yourself for whom, in what context, and how it is best to explain AI systems;
- **No such thing as improving trust.**
Calibrate it according to the capabilities of the AI system;
- **Multi-agent systems are your friend.**
They are ripe with tough-to-explain environments;
- **Ask your users** how they would explain, then **learn from them**;
- **Causal explanations** and natural language is effective.

ACKNOWLEDGEMENTS

My work would not have been possible without my supervisors:

Stefano Albrecht



Chris Lucas



Shay Cohen



And without my collaborators:

Cheng Wang

Anton Kuznietsov

Massimiliano Tamborski

Stephanie Droop

Nick Ferguson

Tadeg Quillien

Burkhard Schafer

Members of AARG

Thank you!
Read more about my work:



<https://gbalint.me/>



Autonomous Agents
Research Group



THE UNIVERSITY of EDINBURGH
informatics



UK Research
and Innovation

REFERENCES

- [1] Quillien, T., & Lucas, C. G. (2023, June 8). Counterfactuals and the Logic of Causal Selection. *Psychological Review*. Advance online publication.
- [2] Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342.
- [3] Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 5(8), Article 8.
- [4] **Gyevnar, B.**, Wang, C., Lucas, C. G., Cohen, S. B., & Albrecht, S. V. (2023). Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. *In The 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2024)*.
- [5] **Gyevnar, B.**, Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In *Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence* (pp. 964 - 971).
- [6] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), Article 5
- [7] Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable Recourse in Linear Classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19. [8] Schneeberger, D., Röttger, R., Cabitza, F., Campagner, A., Plass, M., Müller, H., & Holzinger, A. (2023). The Tower of Babel in Explainable Artificial Intelligence (XAI). In *Machine Learning and Knowledge Extraction* (pp. 65–81). Springer Nature Switzerland.



REFERENCES

- [9] Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. *IEEE Access*, 9, 144352–144360.
- [10] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2020). *Sanity Checks for Saliency Maps* (arXiv:1810.03292). arXiv.1810.03292.
- [11]. Jalwana, M. A. A. K., Akhtar, N., Bennamoun, M., & Mian, A. (2021). CAMERAS: Enhanced Resolution And Sanity preserving Class Activation Mapping for image saliency. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16322–16331.
- [12] Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3543–3556). Association for Computational Linguistics.
- [13] Wiegrefe, S., & Pinter, Y. (2019). Attention is not not Explanation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 11–20). Association for Computational Linguistics.
- [14] Leofante, F., & Potyka, N. (2024). Promoting Counterfactual Robustness through Diversity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19), Article 19.



REFERENCES

- [15] Kuznietsov, A., **Gyevnar, B.**, Wang, C., Peters, S., & Albrecht, S. V. (2024). *Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review* (arXiv:2402.10086). arXiv.
- [16] Miller, T. (2022). *Are we measuring trust correctly in explainability, interpretability, and transparency research?* (arXiv:2209.00651). arXiv.
- [17] Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press.
- [18] **Gyevnar, B.**, Droop, S., & Quillien, T. (2024, March 11). *People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior*. (arXiv:2402.10086). arXiv.
- [19] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- [20] Albrecht, S. V., Brewitt, C., Wilhelm, J., **Gyevnar, B.**, Eiras, F., Dobre, M., & Ramamoorthy, S. (2021, March 15). Interpretable Goal-based Prediction and Planning for Autonomous Driving. *IEEE International Conference on Robotics and Automation (ICRA)*

