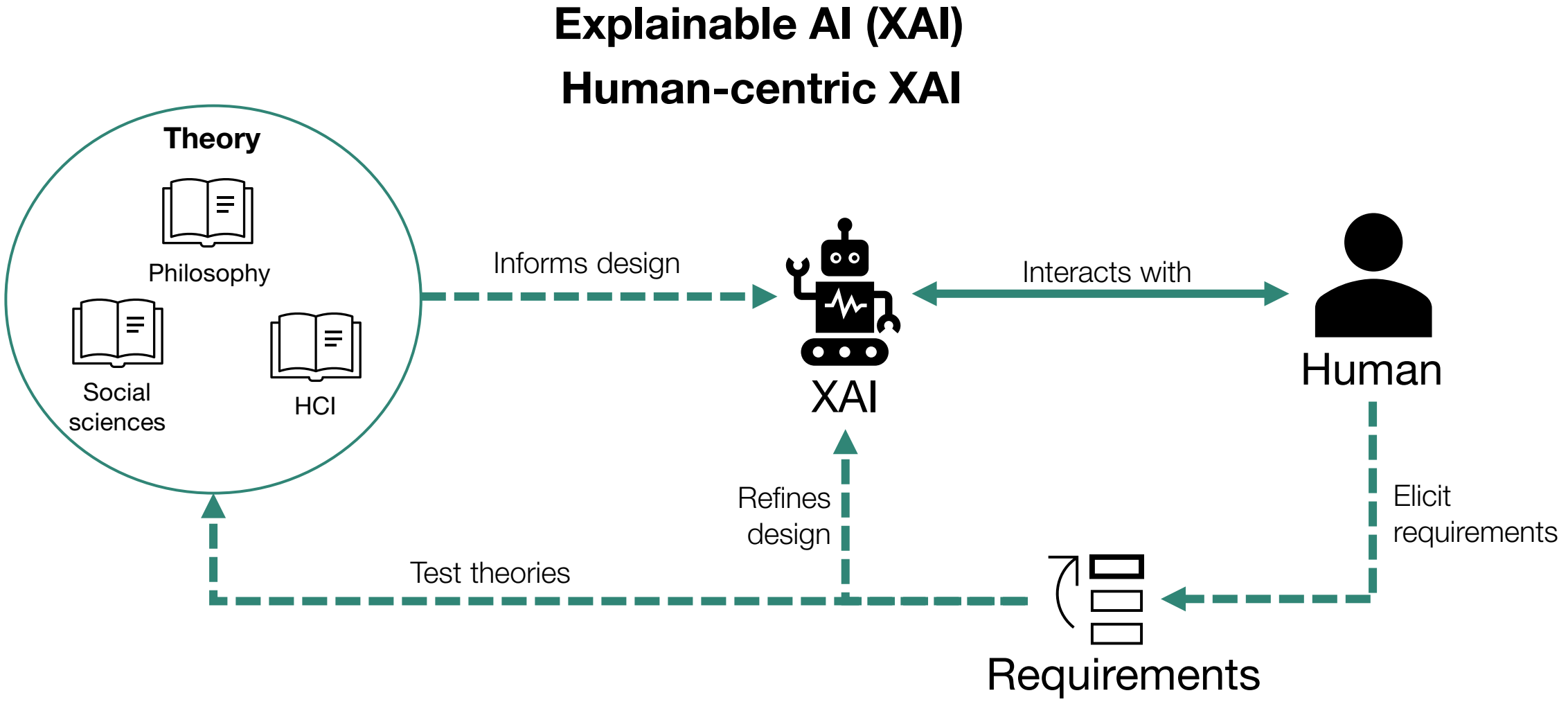
People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior

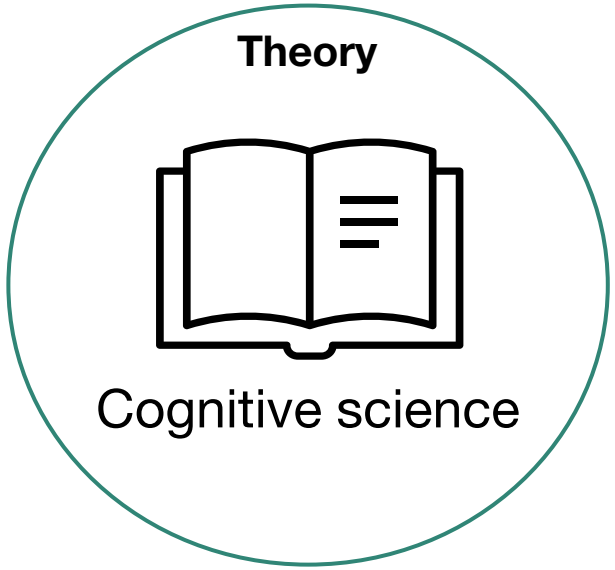
Balint Gyevnar, Stephanie Droop, Tadeg Quillien,
Neil Bramley, Shay Cohen, Chris Lucas, Stefano Albrecht

In CDTalks Series, 28 October 2024

HUMAN-CENTRIC XAI – WITH USER REQUIREMENTS



FOCUSING ON COGNITIVE SCIENCE

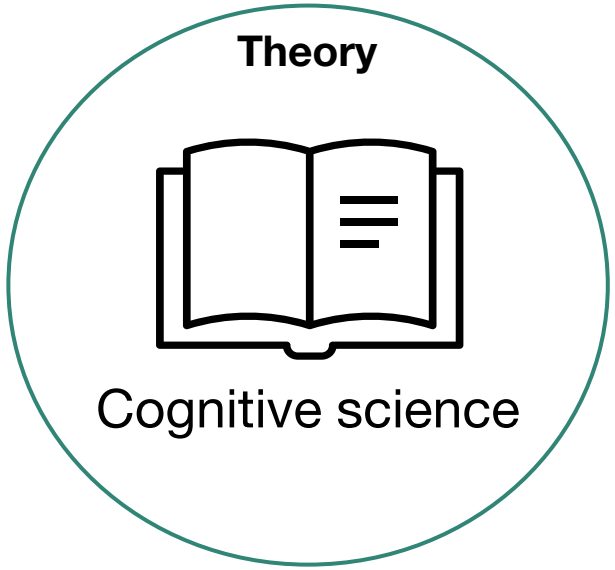


People like explanations that mimic their reasoning processes;

People use causal reasoning to explain;

But: there are many different types (or modes) of causal explanations.

FOCUSING ON COGNITIVE SCIENCE



Different modes of causal explanations:

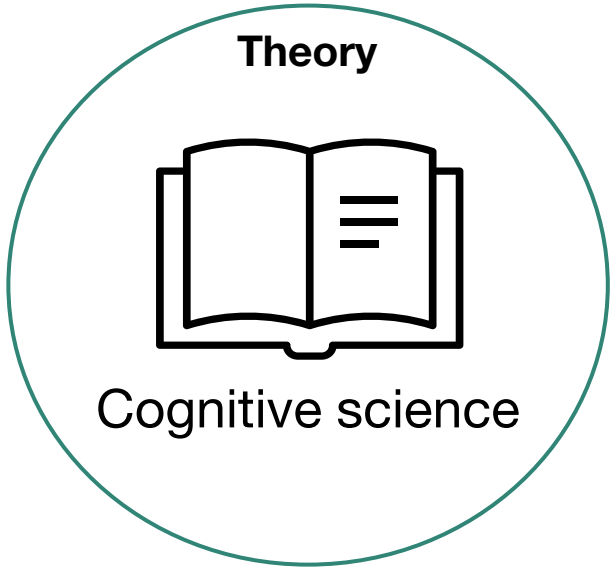
Counterfactual

if I had done x *then* y *would have happened*;

Mechanistic

y *happened because* x *happened*;

FOCUSING ON COGNITIVE SCIENCE



In sufficiently complex systems:

Intentional stance (ascribing belief, desire, intention);

***Teleological* explanation:**

Explaining in terms of the purpose of the action;

Supported by:

Still causal;

Intuitive → Arises early in development;

Robust to environmental circumstances.

Framework of Explanatory Modes:

- Counterfactual
- Teleological
- Mechanistic

A PREDICTION

Our prediction:

In sufficiently complex environments (e.g., AD) people prefer an intentional stance more than using counterfactuals.

AUTONOMOUS DRIVING – DOMAIN OF APPLICATION

Our domain:

Autonomous driving (AD);

Multi-agent system:

Coupled interactions;
Conflicting goals;
Partial observability;

Difficult to explain, even for humans;

AUTONOMOUS DRIVING – DOMAIN OF APPLICATION

Critical environment:

Socially: Driving actions are seen and judged by others;
Epistemically: Partial observability and shared rules;
Safety: Driving can be dangerous;

HOW TO ELICIT REQUIREMENTS?

Two-stage user study with AD scenarios;

1. **Ask participants to write explanations themselves:**
   Still possible to instruct them;
   Gives wide variety;

2. **Then evaluate these explanations with other participants:**
   Perceived degree of counterfactual/teleology/mechanistic focus;
   Perceived number of causes;
   Measures of quality and trustworthiness.

HEADD

Human **E**xplanations for **A**utonomous **D**riving **D**ecisions

HEADD

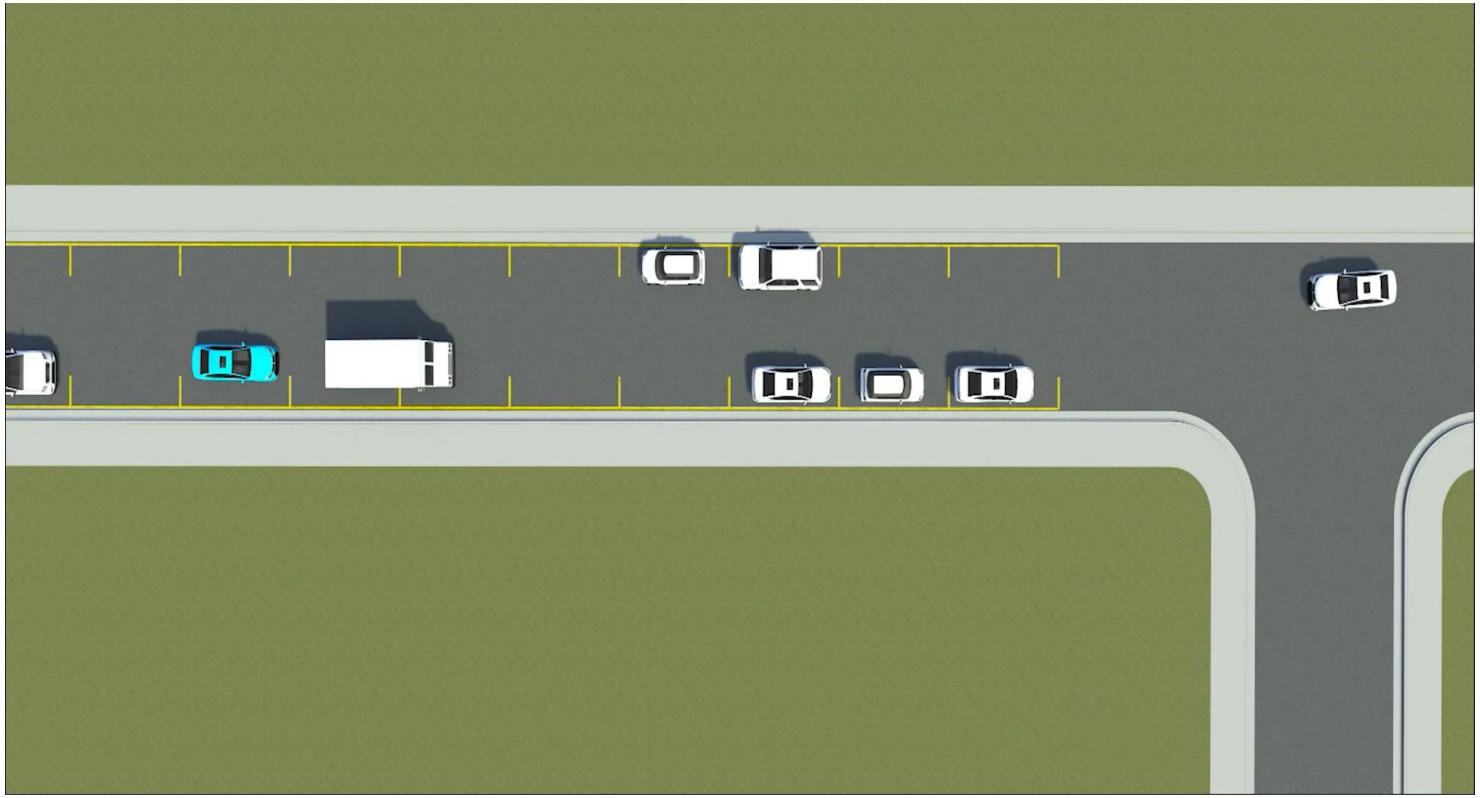
14 unique scenarios with different driving behavior;

1,300+ human-written explanations;

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

5,000+ evaluations.

HEADD – EXAMPLE SCENARIOS

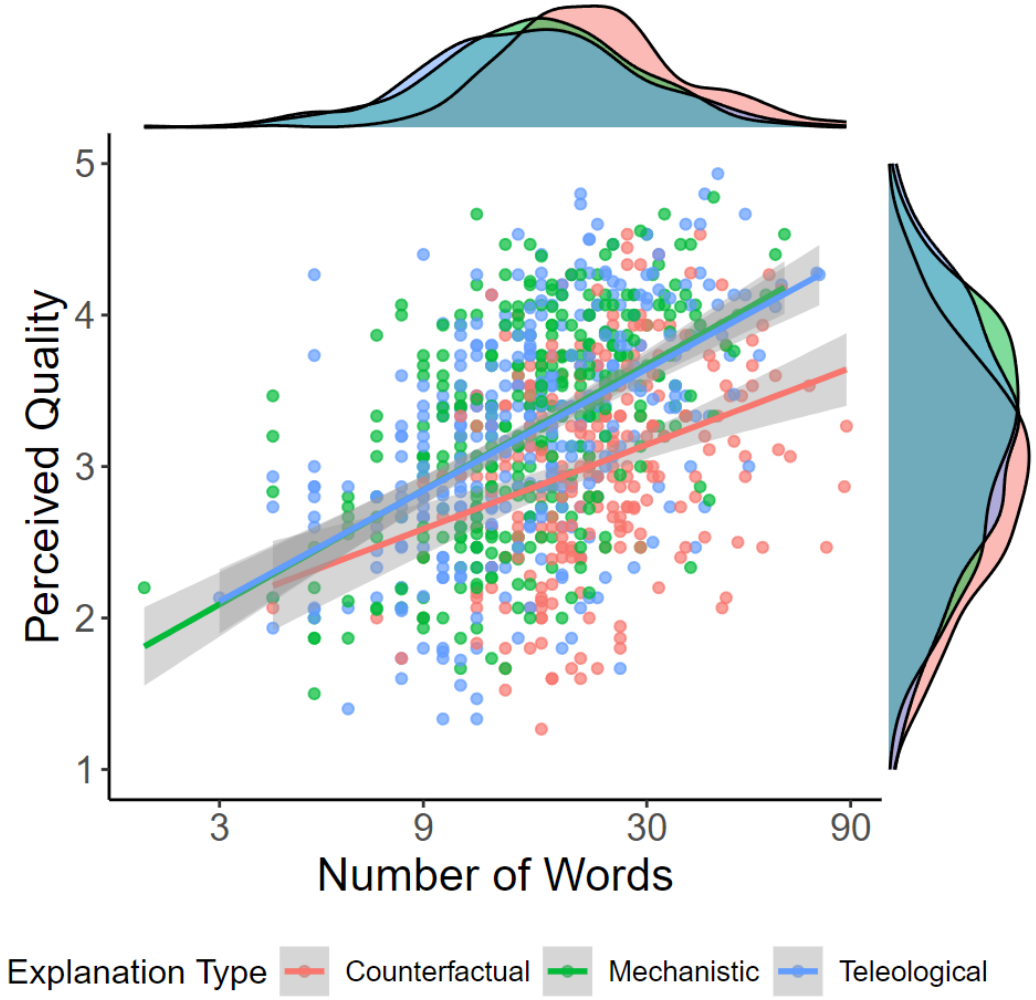


HEADD – EXAMPLE EXPLANATIONS

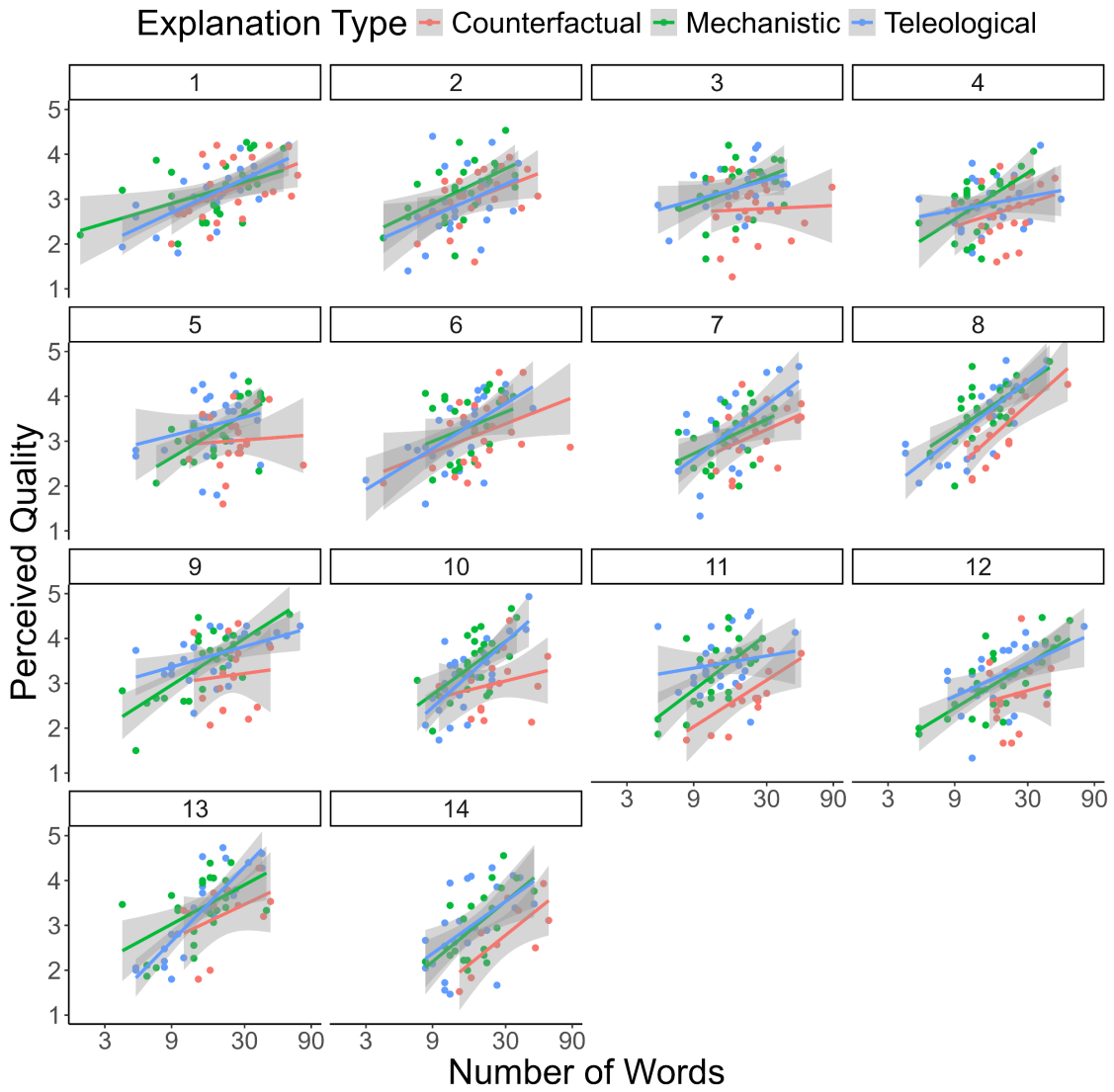
“The blue car was defensive. It could have overtaken the truck while the truck was waiting which could have resulted in an accident with the car approaching from the opposite side.” **(counterfactual)**

“It slowed down in order to prevent any form of collision.
(teleological)

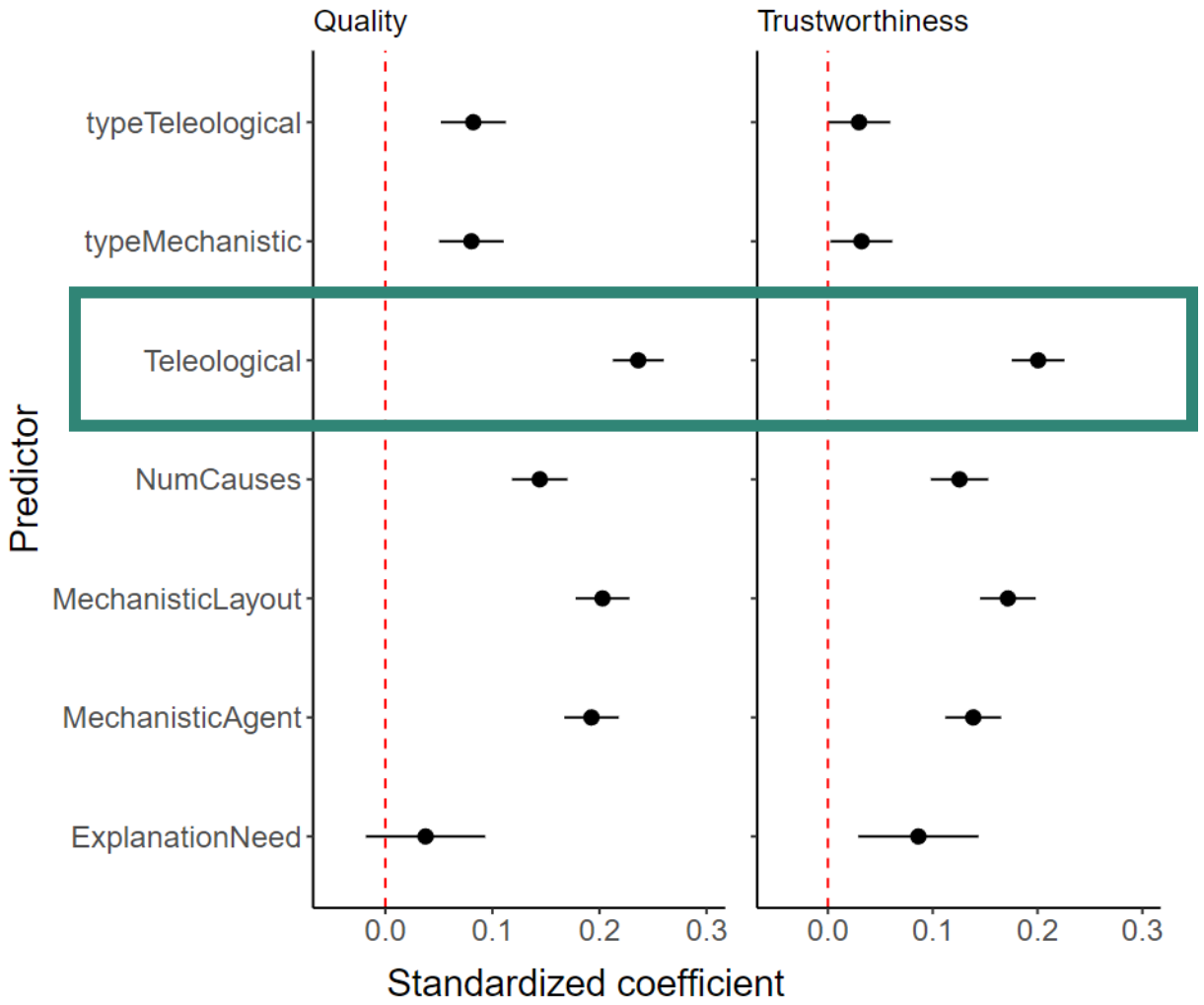
HEADD – PREFERENCES



HEADD – PREFERENCES



HEADD – INSIGHTS FROM THE COGNITIVE SCIENCES



HEADD – TELEOLOGY IS BEST PREDICTOR OF QUALITY

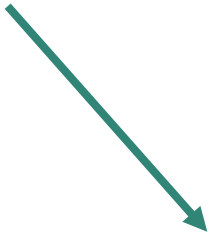
Teleological explanations best predict quality and trustworthiness;

But: most of XAI focuses on mechanistic explanations;

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

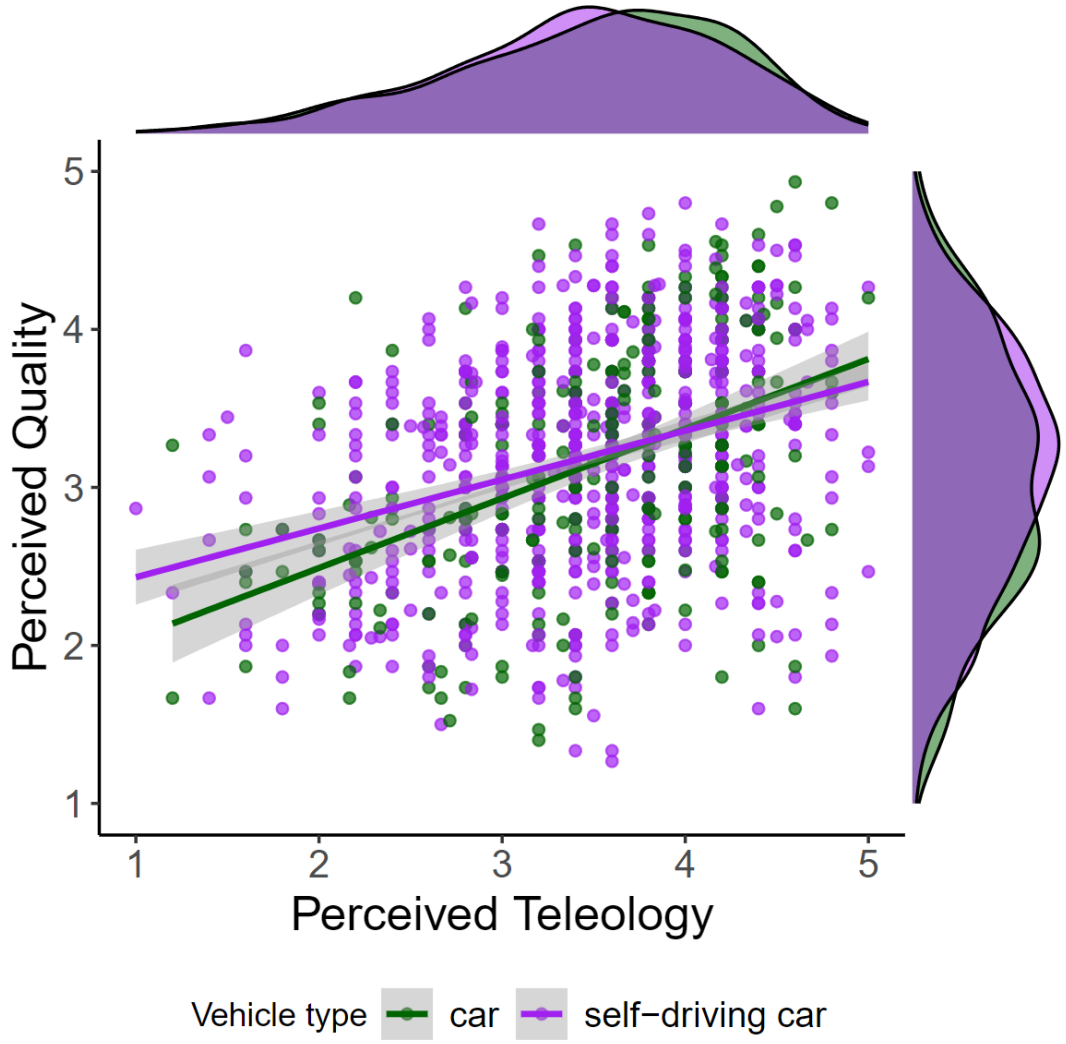
HEADD – HUMAN OR AGENT? DOESN'T MATTER

Why did the blue car change lanes?



Why did the blue self-driving car choose the change lane action?

HEADD – HUMAN OR AGENT? DOESN'T MATTER



HEADD – HUMAN OR AGENT? DOESN'T MATTER

Doesn't matter whether human or machine;

People ascribe teleological concepts to explanations and tend to take the intentional stance anyway.

TAKEAWAYS

- **Essential to <u>understand user requirements</u> in the context domain:**
  Design of XAI should start with domain knowledge elicitation;

- **The <u>framework of explanatory modes</u> provides a useful axis of analysis:**
  We design the type of causation in the explanation not the method first;

- **For complex enough domains, the <u>intentional stance</u> may be more effective:**
  Design explicitly goal-oriented explanations for systems;

- **Artificiality seems not to matter for people in complex systems.**

People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior

https://arxiv.org/abs/2403.08828

Contributions:

- Human Explanations for Autonomous Driving Decisions (HEADD) dataset 14 scenarios, 1,300+ explanations, 4,000+ annotations.
- In complex domains, the intentional stance and teleological explanations are preferred by people.

THE UNIVERSITY of EDINBURGH
informatics

Autonomous Agents Research Group