RESEARCH-ARTICLE

People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior: Insights from Cognitive Science for Explainable AI

BALINT GYEVNAR, The University of Edinburgh, Edinburgh, Scotland, U.K.

STEPHANIE DROOP, The University of Edinburgh, Edinburgh, Scotland, U.K.

TADEG QUILLIEN, College of Arts, Humanities and Social Sciences, Edinburgh, Scotland, U.K.

SHAY B. COHEN, The University of Edinburgh, Edinburgh, Scotland, U.K.

NEIL R. BRAMLEY, College of Arts, Humanities and Social Sciences, Edinburgh, Scotland, U.K.

CHRISTOPHER G. LUCAS, The University of Edinburgh, Edinburgh, Scotland, U.K.

View all

Open Access Support provided by:

The University of Edinburgh

College of Arts, Humanities and Social Sciences

Published: 25 April 2025

Citation in BibTeX format

CHI 2025: CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3706598.3713509
ISBN: 9798400713941

People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior: Insights from Cognitive Science for Explainable AI

Balint Gyevnar
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
balint.gyevnar@ed.ac.uk

Stephanie Droop
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
stephanie.droop@ed.ac.uk

Tadeg Quillien
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
Edinburgh, United Kingdom
tadeg.quillien@ed.ac.uk

Shay B. Cohen
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
scohen@inf.ed.ac.uk

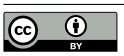
Neil R. Bramley
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
Edinburgh, Scotland, United Kingdom
neil.bramley@ed.ac.uk

Christopher G. Lucas
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
c.lucas@ed.ac.uk

Stefano V. Albrecht
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
s.albrecht@ed.ac.uk

Abstract

It is often argued that effective human-centered explainable artificial intelligence (XAI) should resemble human reasoning. However, empirical investigations of how concepts from cognitive science can aid the design of XAI are lacking. Based on insights from cognitive science, we propose a framework of explanatory modes to analyze how people frame explanations, whether mechanistic, teleological, or counterfactual. Using the complex safety-critical domain of autonomous driving, we conduct an experiment consisting of two studies on (i) how people explain the behavior of a vehicle in 14 unique scenarios ($N_1 = 54$) and (ii) how they perceive these explanations ($N_2 = 382$), curating the novel Human Explanations for Autonomous Driving Decisions (HEADD) dataset. Our main finding is that participants deem teleological explanations significantly better quality than counterfactual ones, with perceived teleology being the best predictor of perceived quality. Based on our results, we argue that explanatory modes are an important axis of analysis when designing and evaluating XAI and highlight the need for a principled and empirically grounded understanding of the cognitive mechanisms of explanation. The HEADD dataset and our code are available at: https://datashare.ed.ac.uk/handle/10283/8930.



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713509

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Cognitive science**; *Causal reasoning and diagnostics*; *Artificial intelligence.*

Keywords

Cognitive science, Explainable AI, Causality, Counterfactuals, Teleology, User study, Autonomous Driving

ACM Reference Format:
Balint Gyevnar, Stephanie Droop, Tadeg Quillien, Shay B. Cohen, Neil R. Bramley, Christopher G. Lucas, and Stefano V. Albrecht. 2025. People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior: Insights from Cognitive Science for Explainable AI. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713509

1 Introduction

The field of explainable AI (XAI) is attracting considerable and growing multi-disciplinary attention. Recent years have seen a shift from viewing XAI as a sterile scalpel for dissecting AI models toward using XAI to coordinate knowledge both between expert and non-expert stakeholders and, in a more expansive near-future vision, between natural and artificial agents [29, 72]. Cross-disciplinary work in XAI draws on, among others, the social sciences [21, 67], psychology [25, 88, 96], philosophy [98], and natural language processing [31, 62, 83]. These collaborations fuel an increasing emphasis on *human-centered XAI* [22].

A central premise of human-centered XAI research is that the design of XAI systems cannot exist in a vacuum but must involve stakeholders from the start [17, 22]. This has already resulted in a rich literature that analyzes and categorizes requirements and user preferences [44, 52, 68, 71, 97]. A common observation throughout these works is that 'participants prefer explanations that resemble human reasoning and explanations' [44, p.9] and explanations should 'align with the cognitive decision-making process that people use when making judgments' [68, p.1].

This 'cognitive approach' to designing XAI systems has been heavily popularized by Miller [67], such that significant efforts were invested into creating algorithms that provide *causal explanations* (for a comprehensive survey, see [86]). Unfortunately, most of this work lacks a robust human-centered motivation for the particular causal framework. This is apparent not only in the lack of user evaluations used to validate explanation generation methods [40] but also in the choice of quantitative metrics, which are ill-suited to assess whether people would give these explanations on their own. A major issue here is that work often focuses solely on explanation fidelity (i.e., how faithful an explanation is to the algorithm) as an objective metric, even though different people may understand an explanation differently and value different causal content depending on their context or background [22, 91].

Fidelity is a popular metric, as most algorithmic research in XAI focuses on supervised machine learning (for a meta-survey, see [80]). In contrast, autonomous decision-making systems have received considerably less attention, as these systems are usually deployed in dynamically changing, often partially observable environments, where behavior may be difficult to explain even for humans. Although explainable reinforcement learning has worked to address the algorithmic challenges of explanation generation in this domain [65], the human-centered aspect has largely been unexplored. As these systems often operate in social and safety-critical settings, understanding the cognitive processes of how humans explain autonomous behavior is essential for more effective XAI design and accurate trust calibration [20, 61].

1.1 Framework of explanatory modes

Understanding the algorithms or cognitive processes of explanation is further complicated by the facts that different academic traditions have developed their own, sometimes conflicting nomenclature of causal reasoning and that the same words can have confusingly disparate folk meanings. To aid our discussion in this work and the broader context of XAI, we have attempted to unify many complex strands under the following framework of *explanatory modes* (i.e., types of explanation).

Teleological: Makes reference to purpose, function, or an agent's goals, intentions, or desires, for example, following the pattern 'x happened in order to bring about y'. This mode is also called intentional, purposive, or functional. We include purpose and function here because the 'agent' may not be an explicit single agent but can be implied or distributed, as in phrases like 'traffic laws are there to coordinate driver behavior and prevent accidents'.

Mechanistic: This mode follows the pattern 'x happened because y happened' and implicitly or explicitly cites stable or abstract principles of how things work, assuming the same thing would happen again if the same conditions pertained. For example: the car stopped because it ran out of gas. This mode is also sometimes called *causal*, but, as that word in folk usage means any reason, and so is the superclass of teleological and counterfactual, in this work we use *mechanistic* any time we mean to emphasize factors which precede an effect.

Counterfactual: An explanatory mode that references how events could have turned out differently from how they, in fact, did. For example, following the pattern 'if x had been different, y would have happened'. In some literature, this mode is also called *contrastive*; we use contrastive in our experiment instructions to avoid jargon.

Descriptive: Describes a situation by rephrasing what was observed, without reference to stable or abstract principles. For example: the car stopped because it ground to a halt. This mode is not strictly an explanation at all but is included here because people often offer as explanation what others would class as description, both in daily life and science [8, 15, 26, 66].

1.2 Why and how we apply this framework

In this paper, we suggest that getting traction on how to give effective explanations to people about autonomous systems will involve a combination of integrating theoretical insights from cognitive science and conducting targeted empirical studies of how people generate and interpret explanations in context.

Research in cognitive science reveals that generating and interpreting causal explanations involves sophisticated computations and inferences [10, 11, 45, 54, 70, 78]. In particular, humans often adopt an *intentional stance* [16] when they explain the behavior of a complex system, ascribing goals, beliefs, and intentions to the system, for example, 'they went to the fridge because they wanted a beer [goal] and believed there was one in there [belief]'. These explanations are inherently teleological; they explain an agent's decision in terms of the purpose of that decision. In contrast, XAI systems usually generate mechanistic explanations that appeal to the mathematical, logical, or external processes and conditions involved in making a decision [84]. An explanation of the inherent purpose of the decision is often lacking.

Furthermore, to understand how people interpret causal explanations, it is also important to assess whether people tend to give mechanistic or teleological explanations, even when the agent is not a person but a machine. We also need to understand whether people's preferences for teleological and mechanistic explanations are at odds with applications of different theories of causation. This improved understanding would allow us to base the design of causal explanations on empirically validated principles.

However, research in cognitive science has traditionally been more focused on tightly controlled environments, while XAI systems, especially in the case of autonomous decision-making, need to grapple with the complexities of real-world deployment. Therefore, rather than taking a simplified toy environment, we scope our experiments on decision-making for autonomous vehicles (AV), a popular and complex application domain for XAI [48]. Against this background, we aim to answer two research questions:

RQ1: Which modes of explanation do people prefer to use and receive when explaining behavior in the complex decision-making domain of autonomous driving?

RQ2: How do people's preferences for explanatory modes change when the explained agent is an autonomous machine versus when it is a human?

To answer these questions, we discuss relevant research in cognitive science on causality, counterfactuals, and teleology as they relate to explanation and formulate hypotheses about the framework of explanatory modes. To test these hypotheses, we designed an experiment consisting of two studies with human participants recruited through the online crowd-sourcing platform Prolific (N_1 = 54; N_2 = 382). In the first study, participants were asked to watch short driving scenarios with multiple interacting vehicles and then explain, in their own words, the behavior of a particular vehicle along the different explanatory modes. In the second study, a different set of participants evaluated these explanations along various dimensions, such as perceived explanatory mode and perceived complexity, quality, and trustworthiness. This setup has the advantage of (i) generating explanations with a diverse sample of real people rather than writing them ourselves, (ii) having realistic situations where the ground truth is still relatively accessible to the explainer, and (iii) allowing us to explore various scenarios while keeping constant the overall context.

We find that the mode of explanation had a significant effect on the judgments of participants who preferred teleological and mechanistic concepts to counterfactual explanations. They were also just as likely to refer to the mental states of AVs as human drivers. In addition, perceived teleology was the best predictor of explanation quality and trustworthiness. Based on these results, our primary recommendation to the field of XAI is to consider different explanatory modes as an important axis of analysis, especially focusing on the role and effect of teleology. In summary, our main contributions are as follows:

- Discussion of causality, counterfactuals, and teleology from cognitive science as they relate to explanation, highlighting the role of different **explanatory modes** (Section 2);
- Curation of a novel **dataset of human-generated and evaluated explanations** for autonomous driving, called the Human Explanations for Autonomous Driving Decisions (HEADD) dataset[1] (Section 3);
- A human participants experiment providing evidence that **teleology is preferred** by people when explaining agents' decision, regardless of whether the agent is perceived as human or machine (Section 4);
- **Recommendations for the design and evaluation of XAI** in autonomous systems for better motivating and understanding the use of causality in human-centered XAI systems (Section 5).

2 Foundations of Explanation

Causation is a cornerstone of effective explanation, and thus XAI. Several fields investigate the notion of cause, albeit slightly differently: the XAI literature focuses on a distinction between causal explanations and counterfactuals, whereas the psychological literature has often made less of a difference and instead used counterfactuals as a means to study causal cognition, making an additional distinction between teleological explanations (focused on goals, intentions, or functions) versus causal-mechanistic ones [38, 57].

2.1 Causal and counterfactual explanations

Explanation has a close relationship with causality [67] and, although there are nuances in the details of how each is formalized [32, 34, 50], it is broadly accepted that explaining a phenomenon usually involves asserting (some of) its cause(s). In turn, causality has a close relationship with counterfactuals [49, 50, 78]. The counterfactual theory of causation, prominent in philosophy and psychology, holds that the meaning of 'C caused E' is (roughly), that if C had not happened (but other factors had still played out the way they did) then E would not have happened [50, 73, 95].

Although causal explanations implicitly involve counterfactual reasoning, it is nonetheless useful to distinguish a narrower meaning of causal explanations from counterfactual explanations. Counterfactual explanations explicitly highlight ways that things could have turned out differently (e.g., 'If I had done x, then y would have happened'), whereas causal explanations as researchers often use the term refer to a chain of events (e.g., 'y happened because x happened'). To avoid ambiguity, we here use the term *mechanistic* for explanations that cite chains of events. Empirically, when people give mechanistic explanations they tend to focus on direct causes that co-vary with an outcome, for example, 'a drunk driver caused the crash' [78]. When constructing counterfactuals, they tend to focus on controllable conditions that could have altered the outcome, for example 'the crash would not have happened if the protagonist had driven home a different way' [63].

Recent research has studied whether counterfactual or mechanistic explanations of an AI system are more effective. Empirical studies have found that users who are given a counterfactual explanation of a decision made by an autonomous system report more satisfaction with that explanation than users who are given a mechanistic explanation [12, 92, 93]. Counterfactual explanations are also more effective at improving the user's ability to predict the behavior of the system. For example, Celar and Byrne [12] showed participants the decisions made by an algorithm designed to determine whether someone's blood alcohol content (BAC) is above or below the legal limit for driving. The decisions were accompanied by a counterfactual explanation ('if the person had drunk 3 units of alcohol, they would be below the limit'), a mechanistic explanation ('drinking 5 units of alcohol caused the person to be above the limit'), or no explanation. Participants rated counterfactual explanations as more satisfying than mechanistic ones. This finding is intriguing when considering computational models of counterfactual reasoning which suggest people simulate several different counterfactual worlds as part of their process of generating explanations [18, 59, 76, 78].

However, the research that suggests an advantage for counterfactual explanations has focused on counterfactuals that typically highlight one possible alternative state of the world and on explanations generated by very simple algorithms. For example, determining whether someone is under the legal BAC limit can be done

[1]The HEADD dataset and the code used for preprocessing and quantitative analyses are all available with documentation at https://datashare.ed.ac.uk/handle/10283/8930.

by applying simple rules. The question arises as to whether this advantage of counterfactuals generalizes to more complex settings. In particular, if the system is sufficiently complex, like a self-driving car, people might take an *intentional stance* toward that system, conceiving of it as an agent.

2.2 Explanatory modes

The human mind entertains different types of causal explanations [2, 10, 28, 38]. Often, the same event can be expressed in different modes, particularly either in mechanistic terms ('the door opened because she kicked it') or in terms of the person's goals and desires ('she opened the door to let her friends in'). Explainable AI most often takes the former mechanistic stance, as the design of XAI methods is usually targeted at tracing the causal chain from input to output in terms of mathematical manipulations and algorithmic mechanisms. In contrast, the latter example corresponds to taking an *intentional stance*, whereby we can conveniently characterize and therefore predict an agent's behavior by attributing to them mental states such as beliefs, desires, and intentions [16]. The intentional stance reliably emerges very early in development [27], and its computational underpinnings are beginning to be mapped out by cognitive scientists [4, 58, 77].

Explanations that use the intentional stance are teleological: they explain something in terms of the purpose it serves. For example, saying that Mary opened the fridge *in order to* get some milk is a teleological explanation as it explains Mary's action in terms of its purpose. Teleological explanations are intuitive to the human mind, even outside the domain of psychological reasoning. They are readily produced and endorsed by children [41, 55]. Adults sometimes endorse teleological explanations even for inanimate processes, for example, when under time pressure [42, 43]. Teleological explanations are generally useful because they identify causes that are *robust* to changes in background circumstances: for example, my intention to drive home would have caused me to get home even if my usual route was closed, because I would then have taken a different route [16, 54, 56, 57]. That is, under this conception the causes of an intentional system's behavior are its goals.

Evidence suggests that people can adopt the intentional stance toward artificial systems [14, 74]. Whenever this is the case, teleological explanations of autonomous system decisions might be particularly effective, because they are consistent with the way the user intuitively represents the system [10, 100]. In our study, we test whether this is the case for autonomous driving which involves explaining the coupled decision-making of multiple agents with a mixture of both human and artificial agents. Will people prefer explanations of a self-driving car that are framed in teleological terms, or will they prefer more classic causal explanations (mechanistic or counterfactual)? In the next section, we explain the relevance of teleology to the appropriateness of counterfactual explanations.

2.3 Teleology and counterfactuals

There is an interesting potential tension between counterfactual explanations and teleology. One recipe for generating counterfactual explanations is to take inspiration from the counterfactual theory of causation and produce a counterfactual of the form 'if C had not happened, then E would not have happened' where C is the cause of outcome E. Consider, for example, a scenario where a self-driving car stops because pedestrians are crossing the road. A counterfactual explanation obeying the standard template would be 'if the pedestrians were not crossing the road, the car would not have stopped'. This explanation effectively highlights the material cause of the car's behavior but does not have teleological content.

To generate a counterfactual explanation of the car's behavior in teleological terms, we might instead say 'if the car had not stopped, it would have run over the pedestrians'. This counterfactual implicitly highlights the *reason* for the car's behavior: the car stopped because if it had not, a bad consequence would have followed. Note that this kind of counterfactual has a different structure than the standard 'if ¬C then ¬E' template: instead of altering the cause (the pedestrians crossing), we alter the effect (the car stopping). Teleological explanations are still implicitly causal: the car stopped because it computed that not stopping would have worse consequences than stopping. Nonetheless, the complexity of teleological explanations might mean that they will be difficult to express in terms of more standard counterfactuals.

These considerations suggest the following prediction: if participants intuitively conceive of self-driving cars as agents, and apply the intentional stance toward them, they might not be satisfied by counterfactual explanations of their behavior, especially if these counterfactuals are of the form 'if ¬C then ¬E'.

2.4 The present study and hypotheses

We curate a dataset, called Human Explanations for Autonomous Driving Decisions (HEADD), of human-generated explanations of the decisions of AVs as well as evaluations of these explanations by a different set of participants. We anticipate that this dataset can shed light on a variety of questions regarding both explanation generation and interpretation. Below, we focus on our main predictions as they relate to the issues we reviewed in this section.

Our key experimental manipulation was the *explanatory mode prompt*: the explanatory mode that participants were asked to use. Our **Counterfactual**[2] explanatory mode prompt requested participants to 'describe changes to the scenario so that the blue vehicle takes different actions'. That is, it requested a counterfactual of the type 'if ¬C then ¬E', which is difficult to interpret in teleological terms. Our **Mechanistic** explanatory mode prompt asked participants to 'explain how the blue car was influenced in the scenario to take these actions'. This request elicits a causal explanation and does not specifically target teleological features of the situation (although it does not preclude them). Finally, our **Teleological** explanatory mode prompt requested participants to 'explain why the blue car took these actions over different actions to reach its goal', foregrounding the agent's intentions, and emphasizing that the relevant counterfactuals are ones where the agent acts differently.

Suppose there is a robust preference for counterfactual explanation for artificial systems. In that case, we expect that explanations generated in response to the Counterfactual prompt should be rated as better than explanations generated in response to the other two prompts (replicating the results in [12, 92]). In contrast, if the behavior of a complex agent (such as a self-driving car) activates an

[2]We capitalize the explanatory mode when we refer to it as the independent variable of an experimental manipulation (i.e., prompt).

intentional stance, then the explanations generated in response to the Teleological (and possibly Mechanistic) prompt should be rated as better than the Counterfactual explanations. Additionally, if participants adopt an intentional stance, we predict that explanations containing more teleological features should be seen as more satisfying. Therefore, our first two hypotheses test people's preferences for which explanatory modes they like to *receive* when explaining complex autonomous behavior:

H1: Explanations given to the Teleological prompt are rated as better than the Counterfactual explanations.

H2: Explanations are rated as more satisfying if they are perceived to contain more teleological features.

Finally, we saw that people can readily form mental models of human agents by ascribing to them mental states such as beliefs, desires, and intentions [16, 27]. There is also evidence to suggest that people do this even for artificial systems [14, 74]. Suppose that people do not ascribe mental states to artificial agents. In this case, we expect to see a difference between the effectiveness of teleological explanations when the explanations are generated about the perceived behavior of a human versus an autonomous system. In contrast, if people do ascribe mental states to artificial systems, then we should not observe a significant difference between explanation ratings of human and autonomous agents. Therefore, our third hypothesis tests people's preferences for teleological explanations with both human and autonomous agents:

H3: Given the same degree of perceived teleology, explanations of autonomous agents are rated as less satisfying than explanations of human agents.

3 Experimental Methods

Our experimental workflow is summarized in Figure 1. The experiment consists of two human participants studies such that the first study provided the data used in the second study. In the first study, participants generated natural language sentences about the behavior of a pre-selected vehicle in various driving scenarios along the four explanatory modes of Section 1.1. In the second study, we took these sentences and asked a different set of participants to evaluate the sentences as explanations according to their causal content and subjective quality.

For both studies, we designed surveys using the Qualtrics platform and relied on the online crowd-sourcing platform Prolific to recruit participants.[3] We recruited from the USA and filtered for participants whose first language was English. Participants were paid a pro-rated fee of £11 per hour and were shown an information sheet and our ethics approval. Participants had to give their consent before they could access the studies. This study received ethics approval according to the Research Ethics Process of the School of Informatics, University of Edinburgh with reference #282628.

Our design choices for both studies, including the phrasing of prompts and the number of iterations per study, were informed by both within-institution (i.e., internal) and crowdsourced (i.e., external) pilots. For the first study, we received 10 internal participants and recruited 5 external participants. For the second study, we received 13 internal participants and recruited 15 external participants for the pilots.

3.1 Driving scenarios

We designed 14 unique driving scenarios based on previous works from autonomous driving and XAI [1, 33, 48, 94] and used these 14 scenarios in both studies. We picked scenarios to cover critical aspects of driving: safety, time efficiency, and comfort, to various degrees. For example, some scenarios were designed to focus heavily on actions related to safety, while others were designed to focus more on reaching a goal faster, that is, time efficiency. We summarize all 14 scenarios in Table 1 with full details available in the HEADD dataset.

We recreated each scenario from a top-down bird's eye view in the form of 5 to 15-second-long animated videos using the software RoadRunner 2023a by MathWorks.[3] In each scenario, a single vehicle is highlighted in blue (hereafter, 'ego vehicle'). Study 1 participants were always directed to explain the actions of the blue ego vehicle. Example snapshots from such videos are shown in Figure 2.

3.2 Study 1: Generating explanations

3.2.1 Participants. A total of 54 participants (25 male, 27 female, 2 non-binary) filled out Study 1 with a median completion duration of approximately 27 minutes. The participants' ages ranged between 19 to 73 years, with a median of 36 years. The majority of participants had some form of tertiary education (49 people) with the largest group having a Bachelor's degree (19 people). Most participants reported having a valid driver's license (48 people) and the majority of participants had been driving for at least 2 years at the time of taking the experiment (44 people).

3.2.2 Design. The workflow of Study 1 is shown in Figure 3. Based on internal comments and the external completion rates in our pilots, we showed each participant seven driving scenarios, sampled randomly and evenly from the collection of 14 scenarios. For each of the seven scenarios, participants were first shown the top-down animated video of the scenario and were told the goal (i.e., destination) of the blue vehicle. Participants were then asked to answer four questions in their own words, corresponding to the four explanatory mode prompts as mentioned in Section 2.4 and further explained below under 'Independent variables'. Participants could re-watch the video and were allowed to answer the four questions in any order.[4] We included two attention checks in the workflow and rejected 5 participants who failed at least one. At the end of the experiment, participants were asked to answer questions regarding their driving experience (driving license, frequency, annual covered distance) and demographics (age, education level, gender). We also gave them the opportunity to free form feedback.

Independent variables. We used two independent between-subjects variables and one within-subjects variable to vary the experiment:

- *Scenario* [between-subjects]: Which seven scenarios were selected for the participant. Scenarios were sampled to enforce equal coverage for all 14 scenarios;
- *AV* [between-subjects]: For half of all participants, we emphasized throughout the experiment that the ego vehicle

[3] Qualtrics: https://www.qualtrics.com/ — Prolific: https://www.prolific.com/ — RoadRunner: https://www.mathworks.com/products/roadrunner.html

[4] For analyses to rule out sequence effects, see Section 3.2.4.

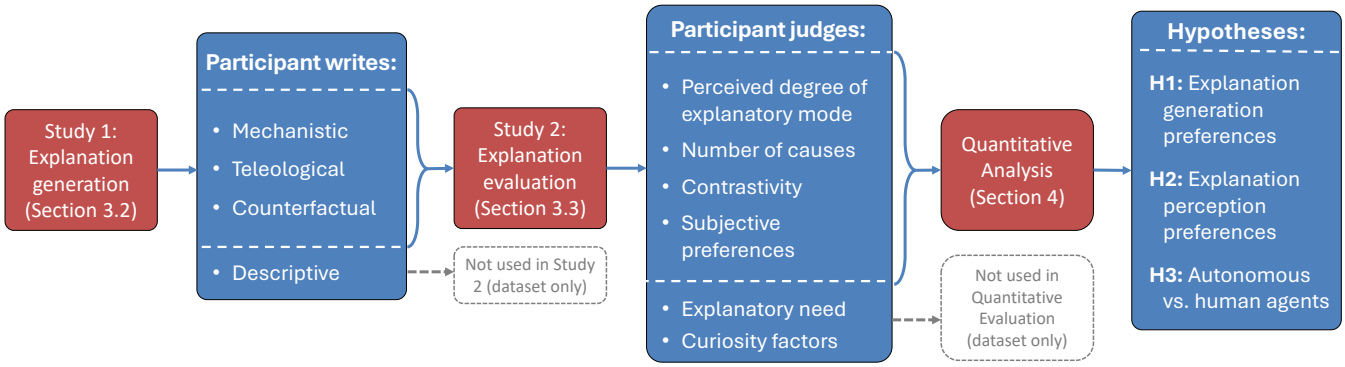


Figure 1: The workflow of our experiment with two studies. In the first study (Section 3.2), online participants generate explanations along the four explanatory modes. In the second study (Section 3.3), new participants evaluate the generated explanations along various subjective measures. Finally, we quantitatively analyze participant judgments and test our hypotheses (Section 4).

Table 1: Summary table of driving scenarios. The columns Efficiency, Comfort, and Safety indicate to what degree the scenario was created with the intent to prompt an explanation related to the time efficiency, comfort, or safety of the driving action.

#	Summary	Efficiency	Comfort	Safety
1	Cutting between cars on highway to exit through off-ramp.	Medium	High	High
2	Earlier right turn as oncoming vehicle stops to give way.	High	Low	Medium
3	Merge into waiting line as oncoming vehicle leaves a gap.	High	Low	Medium
4	Enter two-lane roundabout early when oncoming vehicle enters outer lane.	High	Low	Medium
5	Take over slow moving car likely looking for parking spot.	High	Low	Low
6	Overtake decelerating vehicle as it approaches T-junction.	High	Low	Medium
7	Cautious on-ramp to highway as bushes block view.	Low	Low	High
8	Carefully approaching crossroads with an occluded crossing.	Low	Low	High
9	Giving way in roundabout with occluded center until all clear.	Medium	Low	High
10	Slowly passing line of parked cars and stopping for one pulling out.	Low	Medium	High
11	Sudden breaking when a ball rolls onto the road from behind a truck.	Low	Medium	High
12	Passing between parked cars with truck ahead blocking the view.	Medium	Low	High
13	Slowing down for a high curvature turn.	Medium	High	Low
14	Rapid deceleration to merge behind truck on highway.	Low	High	Medium

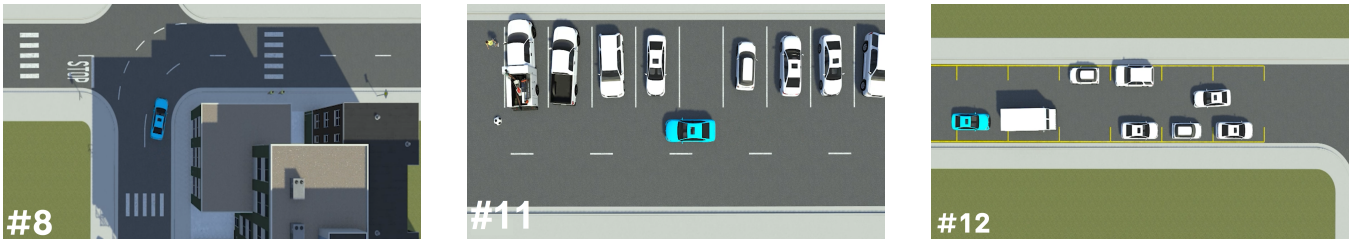


Figure 2: Three example scenarios. Participants were always asked to explain the behavior of the blue ego vehicle. (Left; #8). The blue car slows down before turning right, as its view is blocked by a building. Once the view is clear, the blue car notices pedestrians at the crossing and stops. (Mid; #11). The blue car is passing a row of parked cars when it perceives a ball rolling onto the road. It sharply breaks, as a child emerges from behind a truck. (Right; #12) The blue car waits behind a truck obscuring its vision of the road. It keeps on waiting as the truck passes between the parked cars, to potentially avoid other oncoming cars.

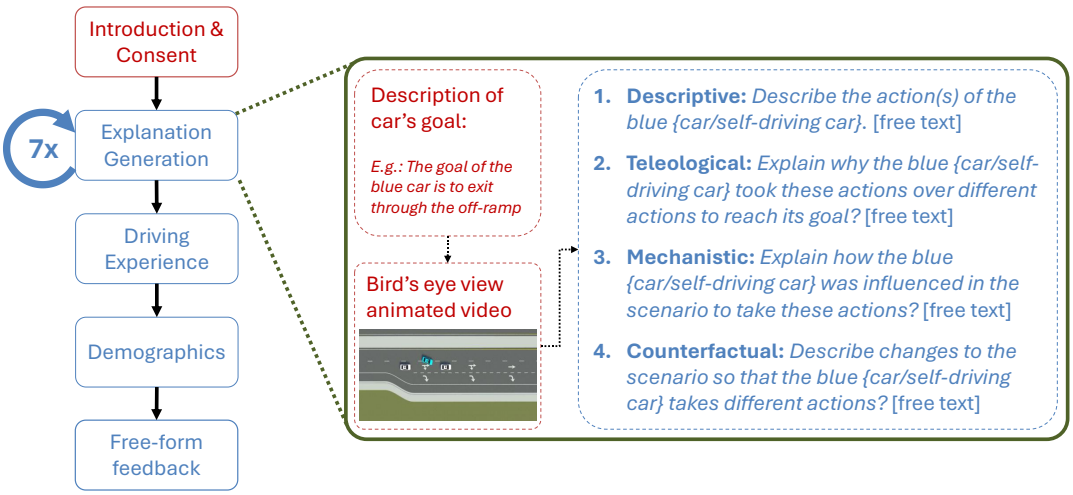


Figure 3: The workflow of Study 1. The text of each question (highlighted in italics) is copied verbatim as it appeared in the experiment to participants, with words in braces replaced according to the independent variables. Red boxes required participants to read and understand instructions. Blue boxes required input from the participants.

was a 'self-driving car' and for the other half it was simply a 'car' in order to allow for measuring the effects of participants thinking that the ego vehicle was operated by an autonomous agent versus a human;

- *Explanatory mode prompt* [within-subjects]: The requested explanatory mode for the explanation. We copy verbatim the text of each prompt:
  (1) *Descriptive*: Describe the actions of the blue car;
  (2) *Teleological*: Explain why the blue car took these actions over different actions to reach its goal;
  (3) *Mechanistic*: Explain how the blue car was influenced in the scenario to take these actions;
  (4) *Counterfactual*: Describe changes to the scenario so that the blue vehicle takes different actions. (The new actions need not be the best actions in the scenario.)

Dependent variables. Our dependent variable in Study 1 is the free-text response given to the four different explanatory mode prompts shown above.

3.2.3 Data and preprocessing. In total, we collected 1,447 sentences across all four explanatory modes. On finishing data collection, we first removed all participants who failed at least one attention check and then manually went through each sentence and removed meaningless or unusable responses (e.g., 'I don't know', 'None', 'Nothing I can think of', etc.). Over all scenarios, this process left 1,347 sentences and the following average number of sentences per explanatory mode per scenario (with standard deviation): Descriptive: 25.78 ± 0.89; Teleological: 25.00 ± 1.24; Mechanistic: 24.21 ± 1.72; and Counterfactual: 21.21 ± 1.97. The average number of explanations given to the Counterfactual prompt is slightly lower than other Explanation prompts because in certain scenarios, especially #13 and #14, some participants struggled to think of a change that would alter the behavior of the ego vehicle. Finally, we also corrected misspellings using a standard spell checker.

3.2.4 Linguistic complexity. As we were also interested in investigating the correlation between linguistic complexity and the quality of explanations, we also performed the following linguistic preprocessing steps. First, each explanation was processed using the Spacy library [37], which tokenized and lemmatized each sentence, and provided dependency parse trees. An explanation may be composed of multiple sentences, in which case, the parsing was performed per sentence. Second, we extracted standard measures of complexity for each explanation: the number of alphanumeric characters, tokens, unique lemmas, and sentences. We calculated this both across the entire explanation and per sentence. We also determined the average separation between dependent tokens in the dependency tree of a sentence as a measure of syntactic complexity.

3.3 Study 2: Evaluating explanations

3.3.1 Participants. We recruited a total of 382 participants (187 male, 193 female, 2 non-binary) to participate in Study 2 with a median duration of completion of approximately 22 minutes. The set of participants of Study 1 and 2 were mutually exclusive. Participants' ages ranged between 19 to 83 years, with a median of 38 years. The majority of participants had some form of tertiary education (320 people) with the largest group having a Bachelor's degree (150 people). Most participants reported having a valid driver's license (343 people) and the majority of participants had been driving for at least 2 years at the time of the experiment (353 people).

3.3.2 Design. The workflow of Study 2 is shown in Figure 4. Study 2 started with a teaching phase during which participants were guided through a simple driving scenario that explained to them the core concepts required to evaluate the explanations from Study 1. This consisted of teaching them the various explanatory modes and a description of what causation is. To keep Study 2 accessible, we aimed to minimize the amount of jargon used. This means, for example, that we refer to explanatory modes as 'explanation targets'

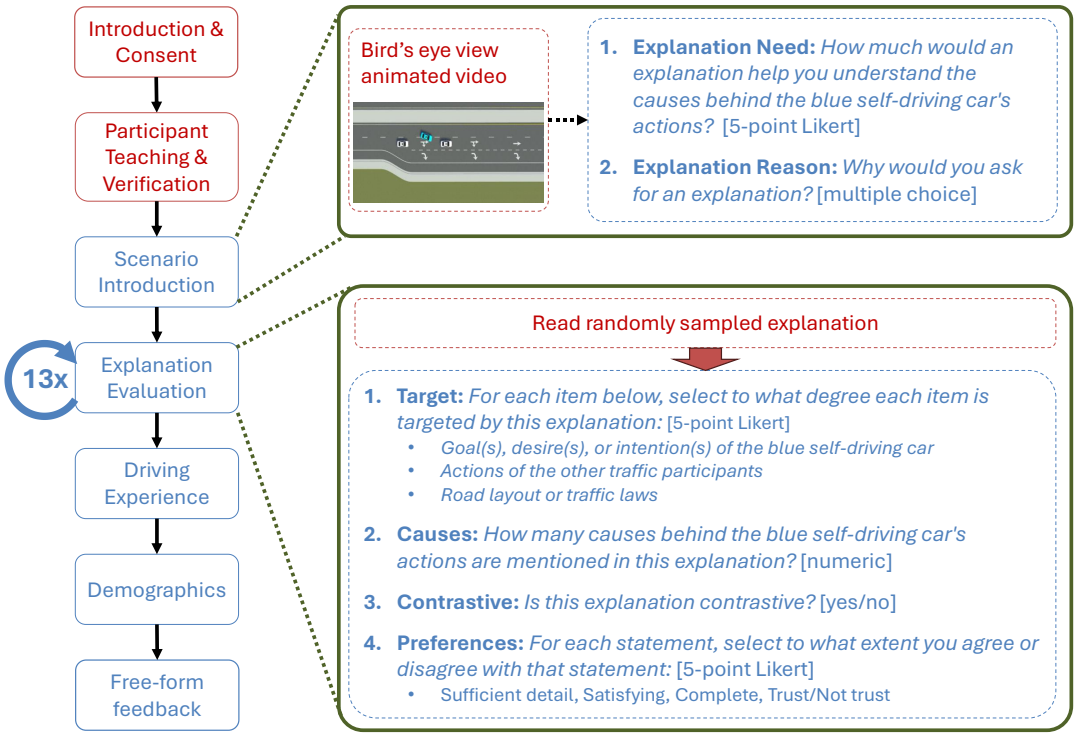


Figure 4: The workflow of Study 2. The text of each question (highlighted in italic) is verbatim as it appeared in the experiment. Red boxes required participants to read and understand instructions. Blue boxes required input from the participants.

and counterfactual as 'contrastive'. To verify whether participants had read and understood the content of the teaching phase, we asked them three single-choice questions regarding the definitions of explanatory mode, causation, and contrastivity. We rejected participants who could not answer any of the three questions correctly given two tries for each question.

Following the teaching phase, we picked at random one of the 14 scenarios, ensuring that all scenarios were picked evenly. We showed participants the video of the scenario without telling them the goal of the ego vehicle. Participants were then asked to answer two questions in order for us to better understand the factors behind whether and what sort of explanations they wanted.

After this, we randomly sampled 13 explanations for the picked scenario. We picked 13 based on feedback from the internal pilot and the completion rates in the external pilot. We only included explanations from Study 1 which were given to the Teleological, Mechanistic, or Counterfactual prompt. That is, Study 2 excluded the responses given to the Descriptive prompt of Study 1. We did this because our focus is on understanding the causal content of explanations, but participants of Study 1 were not explicitly instructed to include causal information for the Descriptive prompt. For each of the 13 explanations, participants were asked to answer several questions regarding its causal content and their subjective preferences, see 'Dependent variables' below.

We included 3 attention checks and removed the data of all participants who failed at least one of them. In total, we excluded 26 participants this way. Finally, participants were asked the same driving experience and demographics-related questions as in Study 1 and had the opportunity to leave feedback.

Independent variables. We manipulated two independent variables in Study 2. No other manipulations were performed as the goal of Study 2 was to provide rich evaluations of the explanations from Study 1:

- *Scenario* [between-subjects]: The scenario for which the participant had to evaluate the explanations;
- *Explanation*: The 13 different non-descriptive explanations sampled randomly and evenly for the given scenario.

Dependent variables. We measured the following dependent variables in Study 2, named as they appear in our results in Section 4. We copy verbatim the survey question text of each variable below.

First, we measured the need for an explanation for the given scenario. These were not used for our analyses later but are included in the dataset to enable more fine-grained future analyses:

(1) *ExplanationNeed* [5-point Likert]: How much would an explanation help you better understand the causes behind the blue self-driving car's actions?
(2) *ExplanationCuriosity* [multiple choice]: Why would you ask for an explanation? Select all that apply: (i) To know that I understand the self-driving car correctly; (ii) To understand what the self-driving car would do next; (iii) To know why the self-driving car did not make some other decision; (iv) To know what the self-driving car would have done if something

had been different; (v) I was surprised by the self-driving car's actions and want to know what I missed.

Second, we measured how much the displayed explanation targets an explanatory mode. We split the mechanistic explanatory mode into two scales to keep the question more accessible while also providing an opportunity for more fine-grained analysis:

(3) *TeleologyPerceived* [5-point Likert]: How much does the explanation target the goal(s), desire(s), or intention(s) of the blue car?
(4) *MechanisticAgent* [5-point Likert]: How much does the explanation target the actions of the other traffic participants?
(5) *MechanisticLayout* [5-point Likert]: How much does the explanation target the road layout or traffic laws?
(6) *NumCauses* [non-negative integer]: How many causes behind the blue self-driving car's actions are mentioned in this explanation?
(7) *Contrastive* [boolean]: Is this explanation contrastive?

Third, we took preference judgments of the participants regarding the following aspects [all 5-point Likert]:

(8) *SufficientDetail*: This explanation of why the self-driving car behaved as it did has sufficient detail;
(9) *Satisfying*: This explanation of why the self-driving car behaved as it did is satisfying;
(10) *Complete*: This explanation of why the self-driving car behaved as it did seems complete;
(11) *Trust*: This explanation lets me judge when I should trust and not trust the self-driving car.

3.3.3 Data and processing. We collected 5,222 annotations for all 986 non-descriptive explanations generated in Study 1. We filtered out the data of participants who failed at least one attention check but kept those who reported not having a valid driver's license as our results were not affected by their inclusion. This gave 5 to 7 independent evaluations per explanation (Mean/Std: 5.3 ± 0.58). To verify that we have sufficient statistical power, we ran a post-hoc power analysis for a χ^2 goodness-of-fit test using the standard G*Power tool [24] with an effect size of 0.3, α error of 0.05, and 22 degrees of freedom: 14 for the independent variable Scenario and 8 for the Explanation (on average each scenario has 103 non-descriptive explanations from which we sample 13: $\lceil 103/13 \rceil = 8$). This gave a $1 - \beta$ error probability of 0.966.

Using the linguistic data from the preprocessing of Study 1, we performed analyses to understand how linguistic complexity correlates with the perceived qualities of explanations. We found that for each measure of complexity, our result did not change significantly depending on which measure we picked, therefore, in the following section we only report results using the number of tokens in the sentence.

3.4 Statistical Analysis

Data were analyzed in R v4.3.3 using RStudio version 2024.04.2+764. Package *lme4* [5] was used to fit mixed-effect regression models following recommendations of [64], as is popular in behavioral sciences for performing multi-level modeling. Mixed-effect models allow one to quantify the *main* or *fixed* effects of interest (e.g., explanatory mode) while also accounting for variation in the effects across participants or stimuli. When reporting the results, the key metric is the slope β of the relevant fixed effect and the corresponding confidence interval. We report the Akaike Information Criterion (AIC) which provides a relative measure of model fit (lower AIC indicates better fit) that penalizes more complex models.

Table 2: Example explanations with quality rankings (by Study 2 participants) for each explanation type, for scenario #8 (see Figure 2). Quality ratings are shown in parentheses except for the descriptive mode as we did not collect ratings for those since the Descriptive prompt was not specifically designed to elicit causal reasoning (see Section 3.3.2).

Explanation Type	Example Human-Generated Explanation
Descriptive	'The blue car drove to the corner and turned right. It then stopped for two pedestrians at the crosswalk. Once the pedestrians were across it continued on its way.' 'The blue car took a right and waited at the crossing for pedestrians to cross before moving forward.' 'Turned right. Waited for a passenger to pass and then continued driving.'
Teleological	'The blue car took these actions to obey traffic laws and maintain the safety of the pedestrians crossing the street.' (4.8) 'This was the logical way to stay on the main road.' (3.33) 'It just followed the directions.' (2.06)
Mechanistic	'The blue car was influenced by the two pedestrians waiting to cross as it slowed to a complete stop allowing them to cross. The right turn was also 90 degrees which required the car to slow in order to make a successful turn.' (4.78) 'The markings on the road showed that the blue car needed to turn right in order to stay on the same road.' (3.4) 'It was influenced a bit by the zebra crossing.' (2)
Counterfactual	'If there were no pedestrians, then the car could have just immediately sped up to the speed limit instead of stopping in front of the crosswalk.' (4.33) 'The blue car could have increased its speed and not waited for the pedestrians but it could have resulted in an accident.' (3) 'The car could have just kept going maybe hitting the people walking.' (2.13)

4 Results

Table 2 shows examples of participant-generated explanations for each explanatory mode prompt. As a manipulation check, we find that explanations generated to a Teleological prompt do exhibit more teleological features than explanations generated in response to other prompts, $p < .001$ for linear mixed model with random effects at the scenario, explanatory mode prompt, and participant

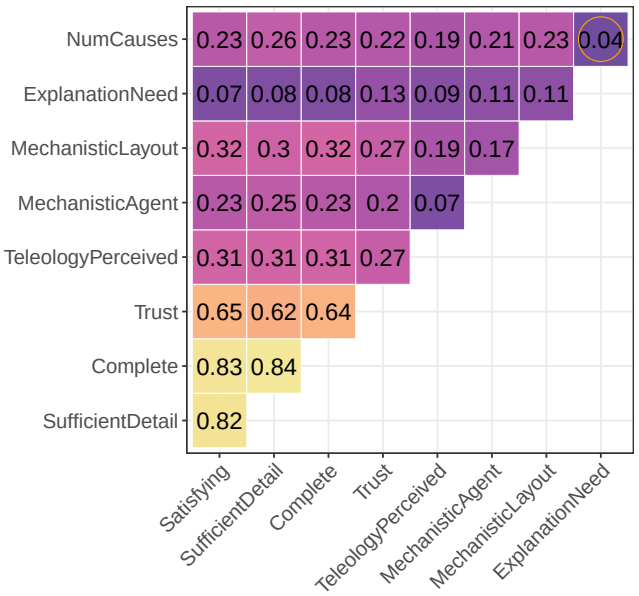


Figure 5: Zero-order correlation between ratings of Study 2. Correlation coefficients circled in orange are non-significant after Bonferroni correction. The Y-axis labels are the dependent variables of Study 2 (cf., Section 3.3.2); The most correlated variables are: (i) *Satisfying*: whether the participants found the explanation satisfying; (ii) *Complete*: perceived level of explanation completeness; (iii) *SufficientDetail*: perceived level of detail in the explanation.

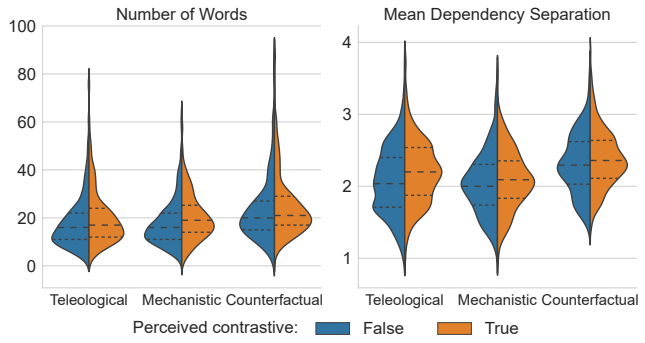


Figure 6: The distribution of the number of words and the mean dependency parse separation grouped by the explanatory mode prompts of Study 1 and colored by whether the explanation was perceived as contrastive by participants of Study 2. Lines inside the violin plots show the median and the interquartile ranges.

levels. We also find that they are less likely to mention the actions of other agents, $p < .001$. In contrast, the type of explanatory mode prompt has no effect on the tendency of explanations to mention aspects of the road layout or traffic laws, $p = .44$.

Figure 5 displays the zero-order correlation matrix among judgments made by participants. Because ratings of Satisfyingness, Completeness, and Sufficient Detail were highly correlated with each other (all $r > .8$, all $p < .001$), we created a composite 'Quality' variable by averaging them. This variable will be the main target in our analyses.

4.1 Observations about the generated explanations of Study 1

Although a rigorous thematic analysis (e.g., coding) is outside the scope of this work, while initially exploring the explanations generated in Study 1, we made three qualitative observations which we present below. We use scenario #8 to illustrate the points here, however, the observations recur across all scenarios, which we confirm quantitatively.

Participants mix different explanatory modes in one explanation: While we wrote the prompts of Study 1 to target a single explanatory mode, many participants gave explanations that involved multiple modes. For example, in scenario #8: 'The blue car saw the crossing and slowed down after taking a right turn. It stopped for the pedestrians to cross.' Here, 'seeing the crossing' is a mechanistic cause, while 'stopping to let the pedestrians cross' is a teleological cause. To confirm this observation quantitatively, we filtered for all explanations according to participants' judgments in Study 2 along the dependent variables TeleologyPerceived, MechanisticAgent, MechanisticLayout, and Contrastive (see Section 3.3.2). We used a minimum average ranking of 3.0 for the first three variables and majority voting for the last variable to determine whether the corresponding explanatory mode was present in the explanation. We found that 719 of the 986 (73%) non-descriptive explanations contained at least two explanatory modes.

Descriptive prompts often elicit teleological reasoning: Many participants included teleological information when instructed purely to describe the behavior of the ego vehicle. We filtered the Descriptive explanations for the presence of purposive keywords[5] and found 138 of the 361 explanations (38%) in the Descriptive category contained references to the ego vehicle's intentions and goals. For example, in scenario #8, one participant described the ego as follows: 'The blue car took a right and waited at the crossing for pedestrians to cross before moving forward.' Here, 'waiting for pedestrians to cross' is using the teleological explanatory mode. Descriptive explanations often include teleological (and only teleological) content, which highlights that people have a natural tendency to take an intentional stance even when they simply describe behavior.

Counterfactual explanations tend to be more complex linguistically: Figure 6 shows the distribution of linguistic features of the responses generated in Study 1, as per the analysis in Section 3.2.4. While the average number of words is similar across the Teleological and Mechanistic explanatory modes, $p = .19, U = 1589983$ in an independent Mann-Whitney U test, the Counterfactual prompt elicited longer sentences with a heavier tail, $p \ll .001, U = 3833441$. The average dependency separation is also higher for responses given to the Counterfactual prompt, $p \ll .001, U = 4013531$, meaning that there tend to be more words between two dependent parts

[5]Including 'wait to' (78), 'allow to' (26), 'to let' (26), 'to exit' (17), 'in order to' (13), 'need to' (11), and 'have to' (10), as well as their past tense conjugation.

of the sentence. Additionally, explanations that were perceived as contrastive (i.e., given in response to the Counterfactual prompt) by participants of Study 2 have higher complexity measures, especially for the Teleological ($p < .001, U = 440951$) and Mechanistic ($p < .001, U = 394835.5$) prompt types. One way to explain the higher linguistic complexity of counterfactual explanations is that they are typically expressed with conditionals, and this tends to result in long sentences (e.g. 'If I had done X then Y would have happened'). Counterfactual reasoning is also argued to require sophisticated computations and more effort [11, 78], which may also result in more complex sentences.

4.2 H1: Mechanistic and Teleological prompts lead to more satisfying explanations than Counterfactual prompts

On average, explanations generated in response to a Mechanistic or Teleological prompt in Study 1 were perceived as better by participants in Study 2 than explanations generated for a Counterfactual prompt; see Figure 7. This effect was statistically significant, as assessed in a linear mixed model with random slopes at the scenario level, and random intercepts at the scenario, explanation, and participant levels: relative to Counterfactual explanations, both Mechanistic ($\beta = .10$, [95% CI: .05, .14]) and Teleological prompts ($\beta = .09$, [95% CI: .05, .13]) elicited higher Quality.

To verify that this result is not due to participants putting in less effort when writing certain types of explanation, we addressed possible sequence effects by performing analyses in which we control for linguistic measures of complexity (see Section 3.2.4), as a proxy for effort. We found that the effect of explanatory mode is robust to all these controls. We also inspected explanations and found that response quality was high in the experiment. We replicated this section's 'satisfaction' analysis while including the linguistic complexity measure as a covariate to the linear mixed model reported above. We find that Mechanistic and Teleological explanations were still perceived as significantly better than Counterfactual explanations in every model, all 95% CI lower-bounds > .04.

In contrast, there was only weak evidence for an effect of explanatory mode on Trust. Using a similar linear mixed modeling approach as above, we find that relative to Counterfactual explanations, Mechanistic ($\beta = .04$, [95% CI: .01, .08]) and Teleological explanations ($\beta = .04$, [95% CI: -.01, .09]) are perceived as only slightly more trustworthy.

All linguistic measures of complexity had a positive effect on perceived Quality, all ps $< .001$ for linear mixed models with random effects at the scenario, explanation prompt, and participant levels. Figure 7 shows, for example, that longer explanations (as indexed by number of words) are rated as better. Interestingly, there was an interaction between number of words and explanatory mode: the number of words had a larger effect on perceived quality for Mechanistic and Teleological explanations relative to Counterfactual explanations, as shown by a linear mixed model with random intercepts at the scenario, explanation and participant levels, interaction effect: $p < .001$; while Counterfactual explanations tended to be more complex with on average the longest length (24.61 words; cf., Section 4.1). Taken together, these results suggest participants find counterfactual explanations less satisfying regardless of length.

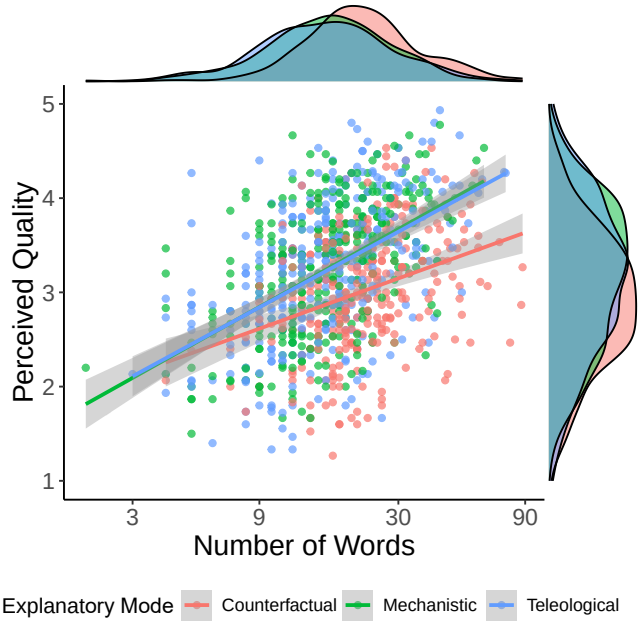


Figure 7: Perceived explanation quality as a function of the number of words in the explanation and the explanatory mode. Each dot corresponds to the average quality rating of one explanation (averaging the ratings of 5 to 7 participants). Lines are linear fits with shaded areas showing 95% CIs.

Intuitively, short explanations tend to be lower effort explanations that are perceived as worse regardless of the original explanation prompt, but participants who put extra effort generated better explanations in response to the Teleological and Mechanistic prompts.

Figure 8 shows that the patterns discussed above are relatively robust across scenarios. Removing the scenario-level random slopes from the linear mixed model we used to test the effect of explanation type did not decrease model fit (full model, AIC = 15070, without random slopes, AIC = 15060). On the other hand, the effect of linguistic complexity appears to vary slightly depending on the scenario: removing the scenario-level random slopes from a linear mixed model predicting perceived Quality from number of words results in a slightly lower model fit (full model, AIC = 14900, without random slopes, AIC = 14908, $p = .003$).

4.3 H2: Teleological features are the main predictor of perceived quality and trust

Participants rated explanations along various feature dimensions: for example, whether an explanation mentioned the agent's goals, how many causes it described, etc. We ran linear mixed models to assess how well these features predicted participants' judgments of the Quality and Trust of explanations. Figure 9 shows the standardized coefficients from two linear mixed models, respectively predicting participants' judgments of Quality and Trust, with random intercepts at the scenario, explanation, and participant levels.

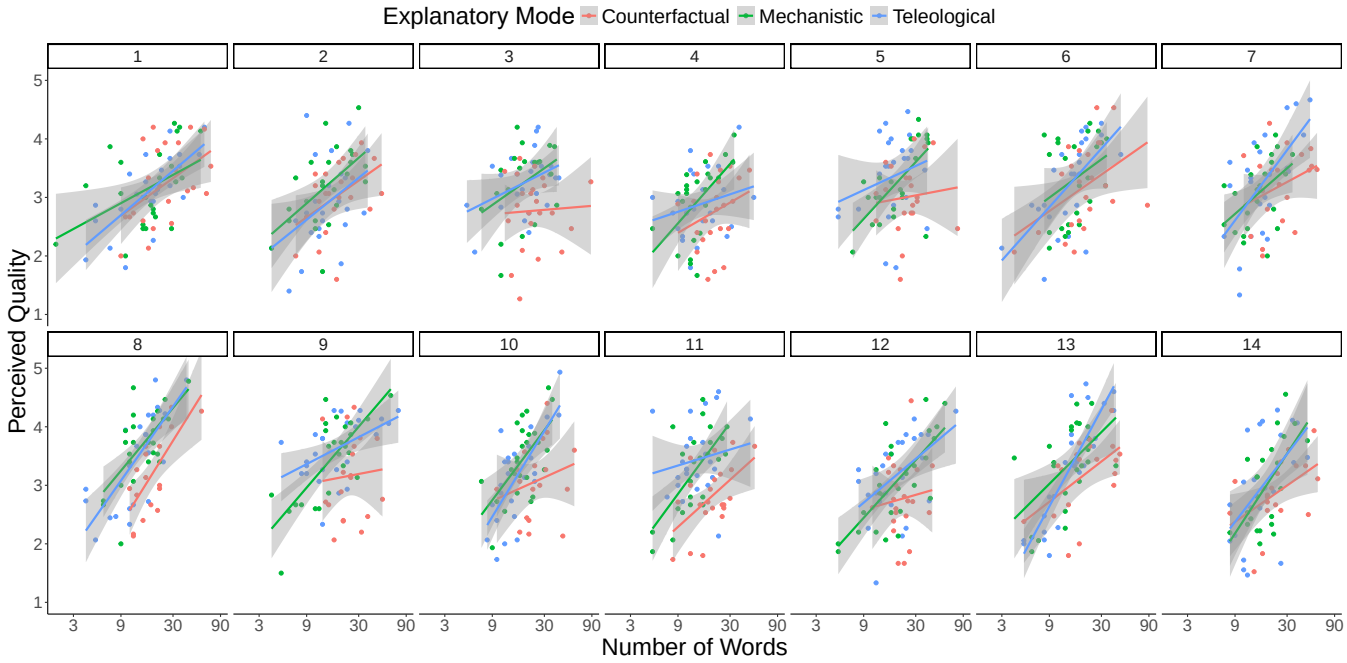


Figure 8: Perceived explanation quality as a function of number of words and explanatory mode, for each of the 14 scenarios. Each dot corresponds to the average quality rating of one explanation (computed by averaging the ratings of about 5 participants). Lines are linear fits with shaded areas showing 95% CIs.

Overall, perceived Teleology was the best predictor of perceived Quality and Trust: the explanations that participants judged as mentioning the goals, desires, or intentions of the agent were perceived as better and more trustworthy. To more formally establish that Teleology is the best predictor of perceived Quality, we computed the AICs of linear mixed models where we removed either Teleology, MechanisticLayout, or MechanisticAgent as predictors. The model without Teleology had a substantially worse fit (AIC=14262) than the models without MechanisticLayout (AIC=14139) and without MechanisticAgent (AIC=14108). A similar approach yields the same result for perceived Trust (model without Teleology, AIC = 14712, without MechanisticLayout, AIC=14633, without MechanisticAgent, AIC=14577). The number of causes mentioned, the extent to which the explanation mentioned the actions of other agents, and the extent to which it mentioned road layout or traffic laws, also reliably predicted both perceived Quality and Trust.

Importantly, perceived Teleology (how much an explanation mentioned the agent's desires, goals, and intentions, as judged by participants in Study 2) and the Teleology prompt (whether participants in Study 1 were explicitly instructed to write Teleological explanations) had independent effects on participants' Quality judgments: each variable has a significant effect when controlling for the other (see Figure 9). Even for explanations generated in response to a Counterfactual or Mechanistic prompt, those that mentioned more teleological features were judged as better and more trustworthy. We did not find a difference in the effect of perceived Teleology across explanation types: adding random slopes at the explanation

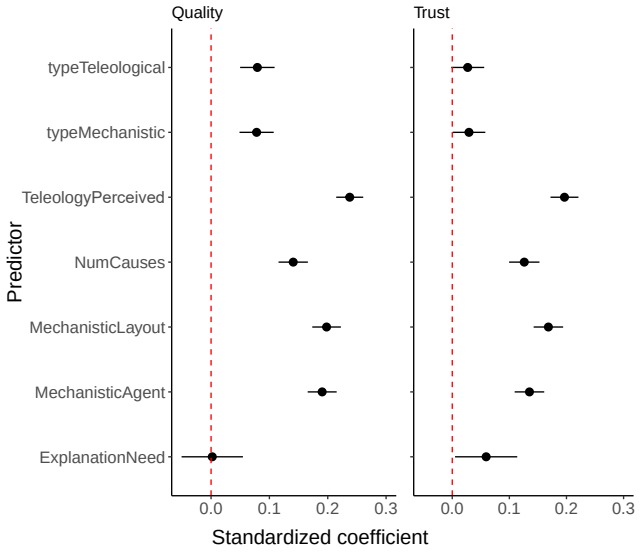


Figure 9: Standardized coefficients from linear mixed models predicting perceived Quality (left) and perceived Trust (right). The first two predictors (type) represent the experimental manipulation in Study 1 (the advantage of the Teleological and Mechanical prompts relative to the Counterfactual prompt, which is taken as a baseline; cf., Section 3.2.2), while the other predictors represent the effect of perceived features of explanations (cf., Section 3.3.2). Error bars represent 95% CIs.

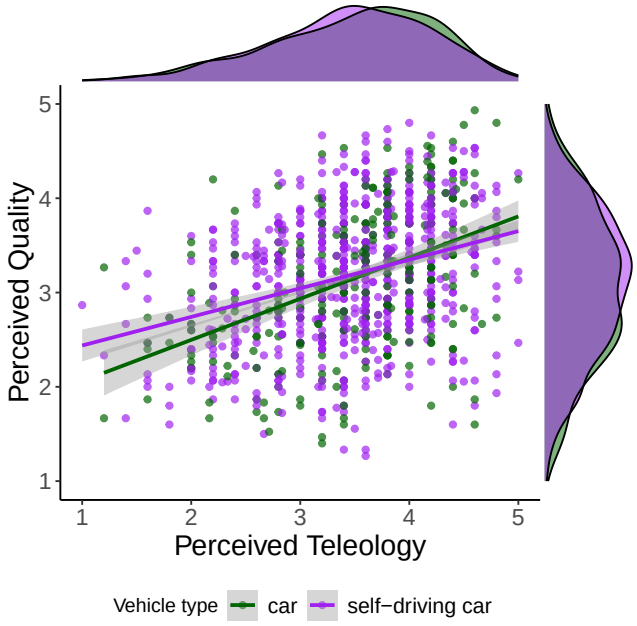


Figure 10: Perceived Quality as a function of Perceived Teleology, and whether the blue vehicle was identified as an autonomous vehicle. Each point represents one explanation. Lines are linear fits with 95% confidence intervals.

type level did not improve the fit of linear mixed models predicting perceived Quality (p = .62) or perceived Trust (p = .27).

4.4 H3: Neither perceived teleology nor quality ratings are affected by autonomous vs. human driver status

We manipulated across conditions whether participants were told the blue car was an autonomous vehicle or was driven by a human driver (independent variable 'AV' in Study 1). We then ran linear mixed models as above to assess whether the fact the person who generated the explanation in Study 1 was explaining the actions of a human or an autonomous vehicle had any effect on either Quality ratings or perceived Teleology (how much an explanation mentioned the agent's desires, goals and intentions, as judged by participants in Study 2). We found no significant difference in either ratings between the two conditions: including AV as a predictor variable contributed no significant improvement in the fit of the linear mixed models predicting perceived Quality (p = .44, χ^2 test for goodness of fit) or perceived degree of Teleology (p = .4). Bayes Factors comparing the full linear mixed models to models omitting AV suggest evidence in favor of the null hypothesis (BF = .02 and BF = .02, respectively). Furthermore, there was no improvement in model fit from including the interaction between AV and perceived Teleology (p = .94, BF = .09), indicating that the effect of perceived Teleology on perceived Quality is the same regardless of whether participants think the car is an autonomous vehicle; see Figure 10.

5 Discussion

In this paper, we introduced a rich dataset of human evaluations of explanations, called the Human Explanations for Autonomous Driving Decisions (HEADD) dataset. These explanations were human-generated and targeted the behavior of autonomous vehicles in short video clips. We hope the HEADD dataset will be a valuable resource to help researchers understand how laypeople generate and interpret explanations. Based on this dataset, our main result is that when people explain the behavior of self-driving vehicles, they often take an intentional stance, conceiving the vehicle as an agent with goals and beliefs. Specifically, we find that H1 and H2 are accepted and H3 is rejected:

- **H1:** Explanations to a Teleological prompt are judged as more satisfying than explanations generated in response to a Counterfactual prompt, which discourages teleological content;
- **H2:** Explanations that are perceived as having teleological content are judged as more satisfying—perceived teleology is the most important predictor of explanation satisfaction;
- **H3:** Whether people are explaining the behavior of human drivers or autonomous vehicles has no effect on the perceived quality of the explanations or the perceived teleology.

The result for H3 suggests that people have no qualms about referring to autonomous vehicles as having beliefs, desires, and intentions. The intentional stance is not just leveled at people but can be a convenient abstraction to help us quickly conceptualize and refer to the outcome of any complex system that fulfills a function that could be expressed in a simplified way as 'goal-like behavior'. This is evident in the way we talk, as witnessed by utterances like 'the car doesn't want to start today', or 'my laptop won't talk to the printer'. Even if people may not actually attribute mental content to machines, they still find it convenient to reason as if machines had mental states [14, 74, 100].

Participants' preference for teleological explanation (H1, H2) highlights the usefulness of concepts from cognitive science for XAI [10]. Cognitive scientists emphasize the fact that most of our knowledge is organized around domain-specific intuitive theories [28, 35]. Explanations that do not reflect the intuitive theory within which people understand a system might not be the most effective for conveying understanding. For autonomous driving, explanations that highlight reasons, objectives, and information states of the vehicles (i.e. that take the intentional stance) are likely to be more explanatory or satisfying to humans than those that focus on how the vehicles work. Notably, when people adopt the intentional stance, they might not favor simple counterfactual-based explanations, in contrast to previous findings in simpler contexts [12, 92]. A preference for the intentional stance also provides support for the design of decision-making systems that explicitly utilize a goal-oriented model (e.g., [1, 33]), as the decisions of these systems would be more amenable to explanation that follows human intuition.

In addition, our study contributes to a nascent literature in psychology that investigates 'naturally occurring' explanations [87, 99]. While the psychological literature on explanations traditionally uses well-controlled stimuli, asking participants to evaluate a handful of experimenter-generated explanations, recent studies have asked participants to evaluate explanations collected from online forums [99] or collected from a crowd-sourcing platform [87]. While

these studies have focused on explanations that target general facts (e.g., 'Why does thunder make noise?'), we contribute to this literature by exploring how people explain specific events (e.g., 'Why did the car stop in this particular situation?'). More generally, our video stimuli depict scenes that are sufficiently rich to be interesting, but also simple enough that explainers can plausibly identify the reason for the agent's behavior. We also replicate some results from these previous studies, finding, for example, that more complex explanations (as indexed by the number of causes they mention) are more satisfying.

Taken together with existing research on human-centered XAI, our findings also have a number of implications for XAI in autonomous driving and, more broadly, for autonomous systems, which we discuss in the following subsections.

5.1 Explanatory modes aid the purposeful design and evaluation of causality in XAI

Our main recommendation is that the design and empirical evaluation of XAI that aims to be human-centered should consider explanatory modes as an important axis of analysis. Previous work has repeatedly shown that explanations that align with the cognitive processes of humans are also more easily interpreted [28, 35, 44, 67, 68, 91, 96]. Still, there are many ways in which people give explanations: even a simple question like 'Why does a pen have ink?' can be answered in multiple qualitatively different yet still causal ways (for example: 'because someone filled the pen with ink', or 'because the pen needs ink in order to perform its function') [67, Section 2.4]. As we saw, these explanatory modes have a significant impact on the perceived quality and satisfyingness of the explanation.

However, it is not only the evaluation of explanations that benefit from an increased understanding of the effects of explanatory modes but also their design. Most often the designers of XAI think along two axes: (i) the type of user query the explanation is supposed to answer [51]; and (ii) the algorithmic properties of the method that generates the explanation [81]. While these axes are undoubtedly helpful in characterizing the explanation generation method, they are less effective in capturing the requirements on the causal content of the explanation.

First, it is common practice in XAI to describe explanations as a function of user queries, capturing the causal content with various question-prototypes such as 'what if'-questions, 'why'-questions, or 'why not'-questions [51]. A clear benefit of this approach is that humans can specify what interests them the most, however, as mentioned above, even a simple 'why' question can be answered in multiple ways. Due to such ambiguities, explanatory modes could complement question prototypes to give both a human-centered and a more principled description of causal content. A simple first approach to test what combination works best would be to draw up a table with rows corresponding to question prototypes and columns to explanatory modes and create explanations for each table cell that fit the two properties. Designers can then select the best combination in a more structured way.

Second, design in XAI is often reduced to mere algorithm selection, such that properties like locality ('global' or 'local') and model-specificity ('specific' or 'agnostic') usurp all other considerations. This is made possible primarily by the prominence of out-of-the-box model interpretation methods, such as SHAP [60]. Yet, blind reliance on these tools not only harms the human-centered usability of the system but also places bounds on the causal content that can be achieved. Take the example of SHAP, a local (i.e., input-dependent) and model-agnostic interpretation tool, which was built to generate mechanistic explanations. A recent survey found that people repeatedly apply SHAP in autonomous driving [48] despite a range of mathematical [6, 47] and usability-related [47, 75, 79, 89] issues. Furthermore, there is no straightforward way to generate teleological explanations with it. Most papers never even consider whether mechanistic explanations are the most conducive to achieving a safer or more trustworthy driving experience.

We believe that incorporating explanatory modes into the design process from the start could help mitigate the downstream effects of blindly relying on existing algorithms. The framework of explanatory modes can complement the already existing algorithmic taxonomies of XAI by providing a high-level perspective on the design of causal content. Should the explanation be purely descriptive? Should it focus on the goals of the system? Or should it highlight the logic involved in making a decision? While the answer will depend on the use case at hand, the framework of explanatory modes gives a good first-principles approach to beginning human-centered explanation design.

5.2 Teleological explanations may be preferred

Our results indicate that for autonomous driving teleological explanations are most preferred by people. This suggests that effective XAI systems that target autonomous systems might benefit from giving teleological explanations priority. One potential reason why people might prefer teleological explanations to other explanatory modes is that teleological explanations are robust to variation in the environment (e.g., if the goal of the agent is to reach destination X then it will try to achieve this regardless of whether a road en route is closed; see Section 2.2). Teleological explanations, therefore, have the benefit of better describing the outcomes of a complex system, which is essential to a better understanding, especially when dealing with safety-critical systems, such as autonomous driving.

More broadly, using teleological explanations is particularly relevant to the fields of explainable reinforcement learning [30, 65] and explainable AI planning [101] where explanations can easily be formulated in terms of the goals or terminal states of agents. For example, existing work on model reconciliation in planning [85] and reward decomposition [e.g., 39, 53, 82] directly consider goals and purpose and may benefit from framing their generated explanations within the explanatory modes framework, especially making clear the distinction between mechanistic (e.g., what minimal changes to a plan reconcile the differences with another plan) and teleological explanations (e.g., what changes to the goal state reconcile two different plans).

However, even for supervised ML, where there is no clearly specified terminal state or goal for the system, the analysis of explanatory modes can be relevant, because the system as a whole must have a purpose. For example, people may be well within their

rights to expect explanations about why a particular credit application was refused, not just in terms of the feature attributions within an arbitrary logistic regression model (i.e., mechanistic explanation), but rather in terms of the monetary goals of the bank. Our framework of explanatory modes as given in Section 1.1 allows the analysis of explanations based on such higher-level goals because it is based on cognitive psychology and does not assume a particular computational methodology.

5.3 Explanatory modes may be valuable for analyzing explainability pitfalls

Based on the long-standing focus of XAI on mechanistic explanations, designers might think that people will prefer machines that are described in mechanistic terms (using non-agentic expressions such as 'programming' and 'directed'). However, our findings indicate that people like teleological explanations equally (using mentalistic expressions such as 'intend to' and 'want to') regardless of whether the agent was framed as a person or a self-driving car. This finding suggests that there is no penalty for using teleology to explain the behavior of artificial agents.

However, there might be issues with a teleological framing. First, it may potentially be exploited as part of a dark design pattern [13], wherein an actor frames the behavior of its agents in such a way that this downplays its responsibility in case of errors. For example, in a traffic accident, an explanation might portray the self-driving car's actions as if it had been taken by a human, using mentalistic expressions to try to direct attention away from its designers.

Second, while this manipulation may be done intentionally, we might accidentally manipulate users' perceptions through explainability pitfalls [23] (EPs). A teleological explanation in which the agent is framed as a human taking an intentional stance (e.g., 'The car wanted to avoid the pedestrian') would likely contribute to increased anthropomorphization and bring with that a range of unexpected biases [19]. Still, it is very difficult to predict how those biases could affect the downstream interactions and experiences of users. As Ehsan and Riedl [23] argue, the best designers can do is to set 'a high bar of accountability [...] to be 'pitfall aware' when designing XAI systems'. Unfortunately, as discussed in Section 5.1, XAI is often reduced to a blind reliance on interpretation tools, which directly ignores this recommendation. We believe that the analysis of explanatory modes is a valuable tool not only for expanding this myopic approach but also for charting the map of potential EPs in an XAI system, especially when it comes to making clear the differing effects of the mentalistic framing of agents in teleological explanations against the non-agentic framing of mechanistic ones.

6 Limitations and Future Work

The present study has some limitations. We chose to generate and evaluate explanations for the domain of autonomous driving. While this has the advantage of being accessible to a large demographic, it also partially limits the interpretation of our results to this domain. However, our insights from cognitive science are domain-independent, and future work should investigate both people's reliance on the intentional stance and the role of explanatory modes as an axis of analysis for better XAI.

Our experiment focuses on explanations of observed behavior rather than of a particular AI model, which brings with it the limitation that participants are not explaining in terms of the logic involved in the decision-making of the autonomous vehicle's actions. Although, it is arguably impossible to explain in terms of the vast subsymbolic mathematical transformations of modern deep learning systems. Still, it is unclear whether our 'behavioral explanations' would be preferred if people knew more about the AI model driving the autonomous vehicle. However, to create better explanations for people, one must still understand how they explain. As such, the framework of explanatory modes and the supporting psychological experiment should still be very relevant to XAI.

In Section 4.1, we found that people tended to mix different explanatory modes in the same explanation. Given that we elicited explanations as free-text responses, it is impossible to have a clean separation between the various explanatory modes. While this is an important consideration when designing an XAI system (i.e., should we mix explanatory modes in an automatically generated explanation?), to interpret our results, we do not need a clean separation. Our findings show that teleology, regardless of the presence of other explanatory modes, tends to improve the perceived quality of the explanations (cf., Figure 9). A cleaner separation would likely make our claims about human preferences less robust because it would not be too similar to how people really explain behavior.

Furthermore, we only recruited participants from the US, which could have introduced a cultural bias. As mentioned in Section 1, different people interpret explanations differently based on individual context, so someone with a different cultural background to the US may prefer different explanatory modes. In addition, considering existing work which has shown that, at least on the level of ethics, cultural backgrounds play a significant role in how autonomous vehicles are perceived [3, 7], we believe it is important to investigate cultural effects on explanatory modes in the future.

Additionally, because driving is a generally well-understood domain, we expect the participants of our experiments to have pre-existing knowledge of the domain. Our findings should be interpreted in light of this context because in other less accessible autonomous decision-making domains, such as network routing or warehouse management, other explanatory modes might be preferred. We discuss the interplay between domain complexity and explanatory modes in Sections 2.1 and 2.2.

Following previous studies, we collected *subjective* measures of explanation quality [36, 87, 99], such as how satisfying or trustworthy the listener considers the explanation. However, subjective measures can diverge from more objective measures (such as how much the explanation improves the listener's ability to predict the system) in subtle ways [9, 12, 30, 46, 90]. Although we found teleological explanations were perceived as superficially more satisfying, it remains an open question to what extent they can help the user better predict the behavior of an autonomous vehicle, or better infer the details of what happened (cf. [45, 69, 70]). Future research should also investigate the extent to which explanation preference varies across different contexts, beyond the range of scenarios considered here. It seems plausible that explanatory preferences might vary in function of many features of a situation, for example, whether an agent's goal is easy to infer, or how much the agent can see. Finally, despite our dichotomous framing, good

natural explanations contain several elements of the modalities discussed, in various combinations.

Acknowledgments

The authors thank Nina Markl and the anonymous reviewers for their helpful feedback on previous drafts. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UK Research and Innovation (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

[1] Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. 2021. Interpretable Goal-based Prediction and Planning for Autonomous Driving. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE Press, Xi'an, China, 1043–1049. doi:10.1109/ICRA48506.2021.9560849

[2] Aristotle. 1933. *Metaphysics*. Vol. 1. Harvard University Press, Cambridge, MA.

[3] Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. 2020. Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants. *Proceedings of the National Academy of Sciences* 117, 5 (Feb. 2020), 2332–2337. doi:10.1073/pnas.1911517117

[4] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 0064.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. doi:10.18637/jss.v067.i01

[6] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences* 121, 2 (2024), e2304406120. doi:10.1073/pnas.2304406120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2304406120

[7] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The Social Dilemma of Autonomous Vehicles. *Science* 352, 6293 (June 2016), 1573–1576. doi:10.1126/science.aaf2654

[8] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[9] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. doi:10.1145/3377325.3377498

[10] Ruth M.J. Byrne. 2023. Good explanations in explainable artificial intelligence (XAI): evidence from human explanatory reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23)*. International Joint Conferences on Artifical Intelligence (IJCAI), Macao, P.R.China, Article 733, 9 pages. doi:10.24963/ijcai.2023/733

[11] Ruth M. J. Byrne. 2005. *The Rational Imagination: How People Create Alternatives to Reality*. The MIT Press, Cambridge, MA, USA. doi:10.7551/mitpress/5756.001.0001

[12] Lenart Celar and Ruth MJ Byrne. 2023. How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition* 51, 7 (2023), 1–16. doi:10.3758/s13421-023-01407-5

[13] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In *IUI workshops (CEUR Workshop Proceedings, Vol. 2327)*. CEUR-WS.org, Los Angeles, CA, 6 pages.

[14] Herbert H Clark and Kerstin Fischer. 2023. Social robots as depictions of social agents. *Behavioral and Brain Sciences* 46 (2023), e21.

[15] Carl F. Craver and David M. Kaplan. 2011. Towards a Mechanistic Philosophy of Neuroscience. In *Continuum Companion to the Philosophy of Science*, Steven French and Juha Saatsi (Eds.). Continuum, London, UK, 268.

[16] Daniel C Dennett. 1987. *The intentional stance*. MIT Press, Cambrdige, MA.

[17] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. 13 pages. arXiv:1702.08608 [cs, stat]

[18] Stephanie Droop and Neil R Bramley. 2023. Extending counterfactual reasoning models to capture unconstrained social explanations.

[19] Brian R Duffy. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems* 42, 3-4 (2003), 177–190.

[20] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies* 58, 6 (June 2003), 697–718. doi:10.1016/S1071-5819(03)00038-7

[21] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–32. doi:10.1145/3613904.3642474

[22] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence (Lecture Notes in Computer Science)*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, New York, NY, 449–466.

[23] Upol Ehsan and Mark O. Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6 (2024), 100971. doi:10.1016/j.patter.2024.100971

[24] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[25] Matija Franklin, Edmond Awad, and David Lagnado. 2021. Blaming automated vehicles in difficult situations. *iScience* 24, 4 (2021), 102252. doi:10.1016/j.isci.2021.102252

[26] Jorge Ignacio Fuentes. 2024. Computational systems as higher-order mechanisms. *Synthese* 203, 2 (2024), 1–26.

[27] György Gergely, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró. 1995. Taking the intentional stance at 12 months of age. *Cognition* 56, 2 (1995), 165–193.

[28] Tobias Gerstenberg and Joshua B Tenenbaum. 2017. Intuitive Theories. In *The Oxford Handbook of Causal Reasoning*. Oxford University Press, New York, NY, US. doi:10.1093/oxfordhb/9780199399550.001.0001

[29] Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. 2023. Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?. In *26th European Conference on Artificial Intelligence*. IOS Press, Krakow, Poland, 964–971.

[30] Balint Gyevnar and Mark Towers. 2025. Objective Metrics for Human-Subjects Evaluation in Explainable Reinforcement Learning. arXiv:2501.19256 [cs.AI] https://arxiv.org/abs/2501.19256

[31] Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2024. Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 771–779.

[32] Joseph Y Halpern. 2016. *Actual Causality*. MIT Press, Cambridge, MA.

[33] Josiah P. Hanna, Arrasy Rahman, Elliot Fosong, Francisco Eiras, Mihai Dobre, John Redford, Subramanian Ramamoorthy, and Stefano V. Albrecht. 2021. Interpretable Goal Recognition in the Presence of Occluded Factors for Autonomous Vehicles. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Prague, Czech Republic). IEEE Press, New York, NY, US, 7044–7051. doi:10.1109/IROS51168.2021.9635903

[34] Denis J Hilton and L McClure John. 2007. The course of events: counterfactuals, causal sequences, and explanation. In *The psychology of counterfactual thinking*. Routledge, London, UK, 56–72.

[35] Lawrence A Hirschfeld and Susan A Gelman. 1994. *Mapping the mind: Domain specificity in cognition and culture*. Cambridge University Press, Cambridge, UK.

[36] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 15 pages. doi:10.3389/fcomp.2023.1096257

[37] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/

[38] Sehrang Joo, Sami R Yousif, and Frank C Keil. 2022. Understanding "Why:" How implicit questions shape explanation preferences. *Cognitive Science* 46, 2 (2022), e13091.

[39] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. 2019. Explainable Reinforcement Learning via Reward Decomposition.

[40] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, 4466–4474.

[41] Deborah Kelemen. 1999. Why are rocks pointy? Children's preference for teleological explanations of th natural world. *Developmental psychology* 35, 6 (1999), 1440.

[42] Deborah Kelemen and Evelyn Rosset. 2009. The human function compunction: Teleological explanation in adults. *Cognition* 111, 1 (2009), 138–143.

[43] Deborah Kelemen, Joshua Rottman, and Rebecca Seston. 2013. Professional physical scientists display tenacious teleological tendencies: purpose-based

reasoning as a cognitive default. *Journal of experimental psychology: General* 142, 4 (2013), 1074.

[44] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581001

[45] Lara Kirfel, Thomas Icard, and Tobias Gerstenberg. 2022. Inference from explanation. *Journal of Experimental Psychology: General* 151, 7 (2022), 1481.

[46] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2125–2137. doi:10.1145/3531146.3534630

[47] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based Explanations as Feature Importance Measures. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 5491–5500.

[48] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V. Albrecht. 2024. Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review. *IEEE Transactions on Intelligent Transportation Systems* 25, 12 (Oct. 2024), 19342–19364. doi:10.1109/TITS.2024.3474469

[49] David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. 2013. Causal responsibility and counterfactuals. *Cognitive science* 37, 6 (2013), 1036–1073.

[50] David Lewis. 2004. Causation as influence. *The Journal of Philosophy* 97, 4 (2004), 182–197.

[51] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[52] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10 (Oct. 2022), 147–159. doi:10.1609/hcomp.v10i1.21995

[53] Zichuan Lin, Li Zhao, Derek Yang, Tao Qin, Tie-Yan Liu, and Guangwen Yang. 2019. Distributional Reward Decomposition for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.

[54] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[55] Tania Lombrozo. 2016. Explanatory preferences shape learning and inference. *Trends in cognitive sciences* 20, 10 (2016), 748–759.

[56] Tania Lombrozo and Susan Carey. 2006. Functional explanation and the function of explanation. *Cognition* 99, 2 (2006), 167–204.

[57] Tania Lombrozo and Daniel Wilkenfeld. 2019. Mechanistic versus functional understanding. *Varieties of understanding: New perspectives from philosophy, psychology, and theology* 209 (2019).

[58] Christopher G Lucas, Thomas L Griffiths, Fei Xu, Christine Fawcett, Alison Gopnik, Tamar Kushnir, Lori Markson, and Jane Hu. 2014. The child as econometrician: A rational model of preference understanding in children. *PloS one* 9, 3 (2014), e92160.

[59] Christopher G Lucas and Charles Kemp. 2015. An improved probabilistic account of counterfactual reasoning. *Psychological review* 122, 4 (2015), 700.

[60] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[61] Stefanie M. Faas, Johannes Kraus, Alexander Schoenhals, and Martin Baumann. 2021. Calibrating Pedestrians' Trust in Automated Vehicles: Does an Intent Display in an External HMI Support Trust Calibration and Safe Crossing Behavior?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3411764.3445738

[62] Prashan Madumal, Tim Miller, Frank Vetere, and Liz Sonenberg. 2018. Towards a Grounded Dialog Model for Explainable Artificial Intelligence. arXiv:1806.08055 [cs]

[63] David R Mandel and Darrin R Lehman. 1996. Counterfactual thinking and ascriptions of cause and preventability. *Journal of personality and social psychology* 71, 3 (1996), 450.

[64] Lotte Meteyard and Robert AI Davies. 2020. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language* 112 (2020), 104092.

[65] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2024. Explainable Reinforcement Learning: A Survey and Comparative Review. *Comput. Surveys* 56, 7 (July 2024), 1–36. doi:10.1145/3616864

[66] Andrew C Miller, Nicholas J Foti, and Emily B Fox. 2021. Breiman's two cultures: You don't have to choose sides. *Observational Studies* 7, 1 (2021), 161–169.

[67] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[68] Tim Miller. 2023. Explainable AI Is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support Using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342.

[69] Andrew Nam, Christopher Hughes, Thomas Icard, and Tobias Gerstenberg. 2023. Show and tell: Learning causal structures from observations and explanations. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

[70] Nicolas Navarre, Can Konuk, Neil Bramley, and Salvador Mascarenhas. 2024. Functional rule inference from causal selection explanations. In *roceedings of the 46th Annual Meeting of the Cognitive Science Society*. eScholarship, Merced, CA.

[71] Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. 2024. User Characteristics in Explainable AI: The Rabbit Hole of Personalization?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3613904.3642352

[72] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* 29, 3 (Sept. 2019), 441–459.

[73] Judea Pearl. 2000. *Causality*. Cambridge University Press, Cambridge, UK.

[74] Jairo Perez-Osorio and Agnieszka Wykowska. 2020. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology* 33, 3 (2020), 369–395.

[75] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 344–350. doi:10.1145/3375627.3375850

[76] Tadeg Quillien and Michael Barlev. 2022. Causal Judgment in the Wild: Evidence from the 2020 U.S. Presidential Election. *Cognitive Science* 46, 2 (2022), e13101. doi:10.1111/cogs.13101

[77] Tadeg Quillien and Tamsin C German. 2021. A simple definition of 'intentionally'. *Cognition* 214 (2021), 104806.

[78] Tadeg Quillien and Christopher G Lucas. 2023. Counterfactuals and the logic of causal selection. *Psychological Review* 131 (2023), 1208—-1234. Issue 5. doi:10.1037/rev0000428

[79] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. doi:10.1038/s42256-019-0048-x

[80] Waddah Saeed and Christian Omlin. 2023. Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. *Knowledge-Based Systems* 263 (March 2023), 110273. doi:10.1016/j.knosys.2023.110273

[81] Gesina Schwalbe and Bettina Finzel. 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* 38, 5 (Jan. 2023), 3043–3101. doi:10.1007/s10618-022-00867-8

[82] Yael Septon, Tobias Huber, Elisabeth André, and Ofra Amir. 2023. Integrating Policy Summaries with Reward Decomposition for Explaining Reinforcement Learning Agents. In *Advances in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection: 21st International Conference, PAAMS 2023, Guimarães, Portugal, July 12–14, 2023, Proceedings* (Guimaraes, Portugal). Springer-Verlag, Berlin, Heidelberg, 320–332. doi:10.1007/978-3-031-37616-0_27

[83] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining Machine Learning Models with Interactive Natural Language Conversations Using TalkToModel. *Nature Machine Intelligence* 5, 8 (Aug. 2023), 873–883.

[84] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2239–2250.

[85] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2021. Foundations of Explanations as Model Reconciliation. *Artificial Intelligence* 301 (Dec. 2021), 103558. doi:10.1016/j.artint.2021.103558

[86] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001.

[87] Justin Sulik, Jeroen van Paridon, and Gary Lupyan. 2023. Explanations in the wild. *Cognition* 237 (2023), 105464.

[88] J. Eric T. Taylor and Graham W. Taylor. 2021. Artificial Cognition: How Experimental Psychology Can Help Generate Explainable Artificial Intelligence. *Psychonomic Bulletin & Review* 28, 2 (April 2021), 454–475.

[89] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. doi:10.1145/3287560.3287566

[90] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial intelligence* 291 (2021), 103404.

[91] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. doi:10.1145/3290605.3300831

[92] Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. 2023. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 171–187. doi:10.1145/3581641.3584090

[93] Greta Warren, Eoin Delaney, Christophe Guéret, and Mark T. Keane. 2024. Explaining Multiple Instances Counterfactually: User Tests of Group-Counterfactuals for XAI. In *Case-Based Reasoning Research and Development - 32nd International Conference, ICCBR 2024, Merida, Mexico, July 1-4, 2024, Proceedings (Lecture Notes in Computer Science, Vol. 14775)*, Juan A. Recio-García, Mauricio Gabriel Orozco-del-Castillo, and Derek Bridge (Eds.). Springer, New York, NY, 206–222. doi:10.1007/978-3-031-63646-2_14

[94] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. 2020. "I'd like an Explanation for That!" Exploring Reactions to Unexpected Autonomous Driving. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3379503.3403554

[95] James Woodward. 2004. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, UK. doi:10.1093/0195155270.001.0001

[96] Scott Cheng-Hsin Yang, Nils Erik Tomas Folke, and Patrick Shafto. 2022. A Psychological Theory of Explainability. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 25007–25021.

[97] Chien Wen (Tina) Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien Tseng. 2023. Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3544548.3580945

[98] Carlos Zednik. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology* 34, 2 (June 2021), 265–288.

[99] Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. 2017. Evaluating everyday explanations. *Psychonomic bulletin & review* 24 (2017), 1488–1500.

[100] John Zerilli. 2022. Explaining machine learning decisions. *Philosophy of Science* 89, 1 (2022), 1–19.

[101] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. 2017. Plan explicability and predictability for robot task planning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (Singapore). IEEE, New York, NY, 1313–1320. doi:10.1109/ICRA.2017.7989155