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

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

Cheng Wang
University of Edinburgh
Edinburgh, United Kingdom
cheng.wang@ed.ac.uk

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

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

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

ABSTRACT

We present **CEMA**: **C**ausal **E**xplanations in **M**ulti-**A**gent systems; a framework for creating causal natural language explanations of an agent's decisions in dynamic sequential multi-agent systems to build more trustworthy autonomous agents. Unlike prior work that assumes a fixed causal structure, CEMA only requires a probabilistic model for forward-simulating the state of the system. Using such a model, CEMA simulates counterfactual worlds that identify the salient causes behind the agent's decisions. We evaluate CEMA on the task of motion planning for autonomous driving and test it in diverse simulated scenarios. We show that CEMA correctly and robustly identifies the causes behind the agent's decisions, even when a large number of other agents is present, and show via a user study that CEMA's explanations have a positive effect on participants' trust in autonomous vehicles and are rated as high as high-quality baseline explanations elicited from other participants. We release the collected explanations with annotations as the HEADD dataset.

KEYWORDS

Explainable AI; human-centric XAI; multi-agent systems; autonomous vehicles; causal explanations; dataset

ACM Reference Format:
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 *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

1 INTRODUCTION

Artificial Intelligence (AI) is subject to heightened social and regulatory scrutiny where trust, or a lack thereof, has proven a barrier to public adoption [22], especially in safety-critical systems such as autonomous driving (AD) [18]. This is in part attributed to the inherent lack of transparency of current black box deep learning-based systems [3]. In response, explainable AI (XAI) has gained popularity.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

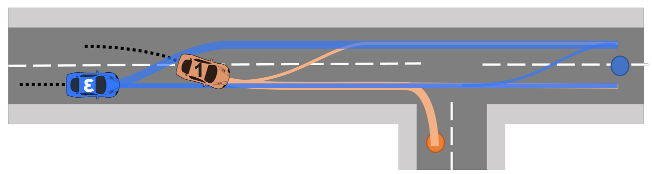


Figure 1: The autonomous vehicle (ε) is heading to the blue goal. It decided to change lanes after the other vehicle (1) cut in front of it and began to slow down. A passenger asks: *Why did you change lanes?* "To decrease the time to reach the goal." [teleological] *Why was changing lanes faster?* "Because the other vehicle is slower than us and is decelerating." [mechanistic] – Actual explanations by CEMA with explanation types in brackets. Blue/orange lines illustrate forward simulations using the probabilistic forward model.

Most XAI methods focus on explanations for supervised learning using tabular or image data [8]. However, these explanations are often purely numeric, and alone have little utility for non-experts who lack domain knowledge to understand the system's internal representations [13]. To address this, XAI is increasingly drawing inspiration from philosophy and the social sciences [28] which has created what we call the subfield of *social XAI*.

An essential part of social XAI is the ability to generate causal explanations. There are several methods for this task [37] and some were proposed for causally explaining sequential decision-making in single-agent systems [9, 38]. However, complex and dynamic multi-agent systems, such as the case with AD, involve tightly coupled interactions among agents where the decisions of any one agent may be difficult to explain even for humans, and there have been few works in XAI addressing this problem. An additional important feature of social XAI is the ability to communicate the extracted causes in the form of intelligible and easy to understand natural language explanations (NLE) as part of a conversational process. A conversation lets users target the pertinent or unclear actions of the agent, while a social XAI system can adjust the user's mental model without excessive cognitive overhead, thereby contributing to more trustworthy interactions with people [11].

To advance the social explainability of multi-agent systems, we introduce a new method called **CEMA**, which stands for **C**ausal **E**xplanations in **M**ulti-**A**gent systems. As illustrated in Figure 1, CEMA is a social XAI method that generates intelligible causal NLEs about an *ego agent's* decisions in sequential multi-agent systems both in terms of the ego's intrinsic motivations (i.e., teleological explanation) and the actions of other agents in the ego's neighborhood (i.e., mechanistic explanation). At the core of CEMA is a novel causal selection algorithm based on the Counterfactual Effect Size Model [33], which builds on a large body of research into how people select causes for explanations. Instead of creating a specific fixed causal structure, CEMA only relies on a probabilistic model for forward-simulating the joint state of the system, which makes it generally applicable where such models are available. By creating counterfactual simulations of what has occurred, CEMA ranks the salient causes behind the ego's actions based on which causes are most correlated with the ego's actions across counterfactual worlds. Causal selection follows a three-step process:

(1) **Roll back** the current factual state of the system to a previous point in time, such that the actions of the ego that we would like to explain have not yet occurred;
(2) **Simulate** a set of counterfactual worlds from this past time point using a probabilistic forward model of the system;
(3) **Calculate** the counterfactual causal effect size by correlating the ego's actions with changes in its rewards and actions of other agents across counterfactuals.

We evaluate CEMA on AD using diverse simulated driving scenarios from the literature with expert explanations [1], we show that CEMA correctly selects causes of the ego's decisions that are congruent with the expert explanations, even when a large number of agents are present. We show that CEMA is robust to changes in the number of counterfactual simulations and the accuracy of the predictive forward model. We also perform a user study to measure the perceived quality and effects of CEMA's explanations on people. First, we collect a set of high-quality human-written explanations as our baseline. We then show that CEMA's explanations are rated on average at least as high as this baseline while positively affecting participants' trust in AD. In summary, our contributions are:[1]

- CEMA: a framework to generate intelligible causal explanations of the decisions of an ego agent in dynamic multi-agent systems based on the Counterfactual Effect Size Model [33];
- Evaluation of CEMA on motion planning for AD, showing its ability to robustly identify correct causes even when a large number of agents are present;
- HEADD: a dataset of Human Explanations for Autonomous Driving Decisions consisting of human-written explanations with a minimum of 5 unique annotations regarding the causal content and trustworthiness of the explanations [15];
- User study showing CEMA's explanations are ranked at least as high as human explanations and a positive effect of CEMA's explanations on trust in AD.

2 BACKGROUND AND RELATED WORK

Causality is a cornerstone of useful human-centric explanations. A common approach for causal selection is to first model the system in the form of a structural causal model (SCM) [31], but this has some drawbacks for complex and dynamically evolving systems. First, it is challenging to model all causal factors in the system, such as the state, action, or reward influences, while keeping the SCM interpretable and useful for end users. Second, the SCM may grow to intractable sizes depending on the desired coverage of causal factors and the complexity of the system. Third, due to the temporal and non-stationary nature of dynamic systems, an SCM may frequently need to be recomputed to adapt to changes. Thus, existing work has applied SCMs only in simpler single-agent systems where, e.g., the agent is trained with a specific algorithm [27, 29].

In addition, AI models have grown complex enough that generating explanations by "opening the black box", i.e., relying on an understanding of the intrinsic causal properties of the trained model, is often infeasible [40]. Instead, we can rely on the *counterfactual model of causation*, which is a well-understood formulation of causation in philosophical literature [19, 24]. Counterfactual cases uncover causes in relation to the factual case by highlighting events whose absence resulted in the counterfactual case rather than the factual case. Implementing the counterfactual model of causation for complex multi-agent systems is challenging in practice. We rely on Quillien and Lucas [33]'s Counterfactual Effect Size Model which is an empirically validated model to operationalize causal selection based on two assumptions about how humans themselves might select causes for explanations. First, people cognitively simulate counterfactual worlds by sampling from a distribution over possible alternative worlds that are grounded in, i.e., not too different from the factual world. Second, people approximate causal effect sizes by correlating variables (i.e., potential causes) in the world with the presence of an outcome across counterfactual simulations. This means that if we have a probabilistic model for forward-simulating a multi-agent system then we can rank and select the most important causes behind the ego agent's actions by simulating counterfactuals.

Furthermore, how a cause is used for the explanation determines its *explanatory mode*. We consider Aristotle's system as it stood the test of time and is still frequently used in the modern discourse of philosophy of explanations [26]. Aristotle argued for four modes: mechanistic, teleological, material, and formal [16]. The *mechanistic* mode gives an explanation describing the mechanisms of the cause of a change, while the *teleological* mode explains to what end or goal a change has occurred. For example in Figure 1, "other vehicle slowing down" is a mechanistic cause while "reaching goal faster" is a teleological cause behind the decision of the blue autonomous vehicle to change lanes. The material and formal modes stay constant in the systems we study, so we do not consider them.

An increasing body of literature studies the generation of explanations for sequential decision-making. However, most methods focus on deterministic planning in well-defined domains [9]. Prior work in explainable reinforcement learning does address single and multi-agent settings in dynamic systems [32], but causal methods are sparser. Madumal et al. [27] is the first to take a causal approach by building an SCM for the action-influence of agents in model-free RL, while Nashed et al. [29] generates explanation by mapping the algorithmic process of solving a Markov Decision Process into an SCM. Others use surrogate interpretable representations of agents' policies with, e.g., decision trees [36] and programs [39]. We are not aware of methods for social XAI in multi-agent systems.

[1] CEMA available at: https://github.com/uoe-agents/cema
HEADD available at: https://datashare.ed.ac.uk/handle/10283/8714

We use AD for evaluation, where probabilistic models for forward simulating the system are widely available [5]. Goal recognition methods predict other agents' future states [6, 7], while motion planning generates optimal behavior for agents [1, 17]. Social XAI also received some attention in AD. For example, Zhang et al. [42] found that explanations in terms of purely high-level tactical causes (e.g., lane change, turn) had little effect on drivers' trust, therefore, more fine-grained insights are required, e.g., in terms of relative position or acceleration. However, prior methods for social XAI in AD do not consider the sequential nature of decision-making [30], rely on a complex neural model which is impossible to certify for safety [23], or only provide high-level explanations [14].

3 CEMA: CAUSAL EXPLANATIONS IN MULTI-AGENT SYSTEMS

We assume that CEMA functions in goal-based sequential multi-agent systems with partial observability, and follow the system definition of Albrecht et al. [1]. Let $\mathcal{I}$ be the set of indexed agents in the environment. At timestep $t \in \mathbb{N}$, each agent $i \in \mathcal{I}$ is in local state $s_t^i \in \mathcal{S}^i$ and receives a local observation $o_t^i \in \mathcal{O}^i$ that probabilistically depends on s_t^i through $p(o_t^i \mid s_t^i)$. In addition, agent i selects an action $a_t^i \in \mathcal{A}^i$ in reaction to observations through $p(a_t^i \mid o_{1:t}^i)$, where the notation $o_{a:b}^i$ denotes a tuple for the sequence $(o_a^i, \dots, o_b^i)$. The joint state of all agents is denoted $s_t \in \mathcal{S}$ where $\mathcal{S} = \times_i \mathcal{S}^i$ and similarly for $o_t \in \mathcal{O}$ and $a_t \in \mathcal{A}$. Further, we assume that agent i is aiming to reach a goal $G^i \subset \mathcal{S}^i$ defined as any partial local state description, such as destination coordinates. The goal G^i may not be observable to other agents. If a state sequence $s_{1:t}$ achieves G^i for agent i, it receives reward $R^i(s_{1:t}) \in \mathbb{R}^d$ which is a d-dimensional vector of reward values where each element in R^i is indexed by a label from a set $\mathcal{R}$ of reward components, such as the time taken to reach the destination. We define the problem of explaining the actions of a particular ego agent $\varepsilon \in \mathcal{I}$ as creating the explanatory function $f \colon (\mathcal{O}^\varepsilon)^* \times (\mathcal{A}^\varepsilon)^* \to \mathcal{H}$ that maps a sequence of local observations and actions to an explanation from a set of possible explanations $\mathcal{H}$. For example, one could define $\mathcal{H} \subset \mathcal{A}^*$, so that an explanation is a partial sequence of actions. We use $\hat{s}_{a:b}$ to indicate that the sequence may contain counterfactual states. We write $s_{x:y} \prec s_{a:b}$ if $s_{x:y}$ is a subsequence of $s_{a:b}$.

We also assume the existence of a probabilistic model that can be used to stochastically forward simulate the system. These are readily available in existing multi-agent literature, for example, in the form of planners or trained reinforcement learning policies [2, 21]. Such probabilistic models define a conditional probability distribution over subsequent joint states of the system given previous observations and actions. We denote this model with $p(\hat{S}_{t+1:n} \mid o_{1:t}^\varepsilon, a_{1:t}^\varepsilon)$, where n is the last timestep. In the case when the local state is fully observable to the ego agent (such as in our evaluation), this model can be replaced with $p(\hat{S}_{t+1:n} \mid s_{1:t}^\varepsilon, a_{1:t}^\varepsilon)$, dropping $a_{1:t}^\varepsilon$ for notational simplicity. Note, that the goals of other agents remain unobservable even under this assumption.

3.1 Social XAI Framework

The process of CEMA (Figure 2) begins with the user asking a question about an ego agent ε and an action they would like explained.

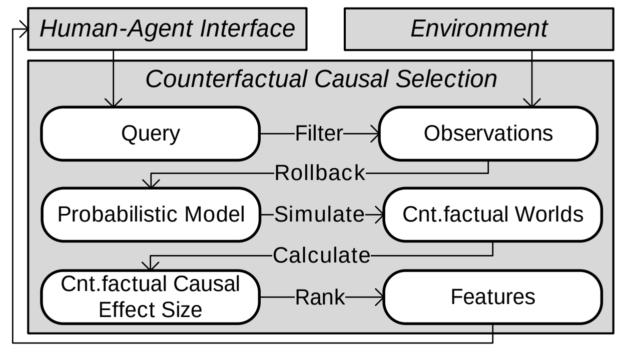


Figure 2: First, irrelevant observations are filtered out based on the query. Second, CEMA rolls back the filtered observations to a previous timestep so that the queried action is erased. From then, CEMA simulates counterfactual worlds to calculate the counterfactual causal effect size for the queried actions, which are used to rank the features of the system.

The question is parsed by an external human-agent interface into a machine-readable *query*, denoted q, encoding a description of the state sequence $\hat{s}_{u:v}$ that corresponds to the ego's queried action. Here, u is the start timestep and v is the final timestep of the queried action. Irrelevant states may be then filtered out from the observed states $s_{1:t}$ based on $\hat{s}_{u:v}$. For example, if $\hat{s}_{u:v}$ refers to an action in the past ($v < t$), then we can ignore states after timestep v. The queried action $\hat{s}_{u:v}$ need not be a subsequence of $s_{1:t}$, instead it can also be a hypothetical sequence that appears, e.g., in a counterfactual world. This allows the user to ask contrastive questions, for example of the form *"Why did you not do Y instead of X?"* The filtered observations and the query are then passed to the counterfactual causal selection module discussed in detail in Section 3.2.

As the focus of CEMA is to generate intelligible explanations for end users, in this framework explanations are composed from a set of *features* $\mathcal{F}$ which describe semantically meaningful properties of a state and/or action sequence. For discreet $\mathcal{S}$ and $\mathcal{A}$ with inherent interpretations, the set of features might simply equal $\mathcal{S} \cup \mathcal{A}$. For continuous spaces, such as in AD, $\mathcal{F}$ might include a discretized summary of actions, such as average acceleration or distance to the leading vehicle. The set of reward components $\mathcal{R} \subset \mathcal{F}$ are also considered features. For example in autonomous driving, these might be time to destination or presence of collisions. CEMA does not assume anything about the actual meaning or properties of features except that there is some feature function $\phi : \mathcal{S}^* \times \mathcal{A}^* \to \mathcal{F}$ converting a state and action sequence to features. Given the above, for CEMA we define the set of all explanations as $\mathcal{H} = (\mathcal{F} \times \mathbb{R})^*$, so that the output of the counterfactual causal selection process is a subset of features $\mathcal{F}$ with corresponding ranking by counterfactual causal effect size. Finally, the explanation is converted into an NLE and returned to the user via the human-agent interface.

3.2 Counterfactual Causal Selection

The counterfactual causal selection process has three main steps. First, it rolls back time before the start timestep u of the queried action, erasing the queried action (Algorithm 1). Second, this rollback

Algorithm 1 Counterfactual dataset simulation

Input: Parsed query q; observed joint state sequence $s_{1:t}$.
Output: Counterfactual dataset $\mathcal{D} = \{(\hat{s}^{(k)}_{\tau+1:n}, y^{(k)}, r^{(k)})\}_{k=1}^{K}$.
$\mathcal{D} \leftarrow \emptyset$.
$\tau \leftarrow$ Determine from $s_{1:t}$ assuring that $q.\hat{s}_{u:v}$ is erased.
for K iterations **do**
    Get $\hat{s}_{\tau+1:n} \sim p(\hat{S}_{\tau+1:n} \mid s_{1:\tau})$ via forward simulation.
    Determine reward for ego $r \leftarrow R^{\varepsilon}(\hat{s}_{\tau+1:n})$.
    Presence of query $y \leftarrow 1$ if $q.\hat{s}_{u:v} \prec \hat{s}_{\tau+1:n}$ else 0.
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\hat{s}_{\tau+1:n}, y, r)\}$.
end for

Algorithm 2 Calculate counterfactual causal effect size

Input: Counterfactual dataset $\mathcal{D}$.
Output: Mechanistic ($\mathcal{F}^m$) or teleological ($\mathcal{R}^t$) explanation.
Mechanistic explanation
$\mathcal{F}^m \leftarrow []$.
for interval end-point $p_j \in P$ **do**
    $\mathcal{D}_j \leftarrow$ Slice $\hat{s}^{(k)}_{\tau+1:n} \in \mathcal{D}$ from p_{j-1} to p_j giving $\hat{s}^{(k)}_{p_{j-1}:p_j}$.
    $\mathcal{X}, \mathcal{Y} \leftarrow$ convert $\mathcal{D}_j$ to features $\phi(\hat{s}^{(k)}_{p_{j-1}:p_j})$ and targets $y^{(k)}$.
    $\mathcal{M} \leftarrow$ Fit an interpretable classifier to $\mathcal{X}$ predicting $\mathcal{Y}$.
    $w, I \leftarrow$ Feature importance attributions w of $\mathcal{M}$ indexed in descending order by I.
    Append $\mathcal{F}^m_j = \{(\mathcal{F}_i, w_i) \mid i \in I\}$ to $\mathcal{F}^m$.
end for
Teleological explanation
$\mathcal{X} \leftarrow$ Filter $\mathcal{D}$ by $y^{(k)} = 1$ for match with query.
$\mathcal{Y} \leftarrow \mathcal{D} \setminus \mathcal{X}$, all samples not matching the query.
$w, I \leftarrow \mathbb{E}_{\mathcal{X}}[r] - \mathbb{E}_{\mathcal{Y}}[r]$ indexed by I in absolute desc. order.
$\mathcal{R}^t \leftarrow \{(\mathcal{R}_i, w_i) \mid i \in I\}$.

allows CEMA to simulate counterfactual alternatives to the queried action (Algorithm 1). Third, the counterfactual simulations inform us about which features of the system are most important for the queried action to occur and we use this information to calculate the counterfactual causal effect size for both the teleological and the mechanistic explanatory mode presented in Section 2 (Algorithm 2).

Algorithm 1 starts by rolling back the joint state sequence $s_{1:t}$ to a timestep τ, such that $\tau \leq u$, resulting in a truncated sequence $s_{1:\tau}$ that assures that the queried action $\hat{s}_{u:v}$ is erased from $s_{1:t}$. The value of τ can be a fixed distance from u or it can be determined to, for example, correspond to the start of a distinct qualitative change in the ego's behavior prior to u. The algorithm then performs K number of forward simulations of the system from time τ according to the probabilistic model $p(\hat{S}_{\tau+1:n} \mid s_{1:\tau})$. For each simulation, we obtain a sequence of future joint states of the system denoted $\hat{s}_{\tau+1:n}$, determine the reward $r \in \mathbb{R}^d$ for the ego, and whether the queried action $\hat{s}_{u:v}$ of the ego was present in the simulation ($y \in \{0, 1\}$). This process gives a dataset of simulations denoted $\mathcal{D}$.

Algorithm 2 has two parts, one for each mode of explanation.

Mechanistic explanations are formulated in terms of the actions of other agents in the neighborhood of the ego vehicle. Actions of the other agents can have different causal effects on the ego at different times, so we first increase the granularity of explanations by cutting sequences into $|P|$ slices defined by their end-points $P = (p_1, \ldots, p_{|P|})$ with $p_0 = \tau + 1$ assumed implicitly. Each slice is then converted to a set of features using the feature function ϕ. Following Quillien and Lucas [33], features that co-occur more frequently with the queried action across counterfactuals should be ranked higher as a salient cause by humans. Therefore, for each slice $p_j \in P$ of a counterfactual simulation, Algorithm 2 measures the counterfactual causal effect size of features on the presence of the queried action y by correlating features with the presence of the action across the simulated counterfactuals. For this, an interpretable classifier $\mathcal{M}$ (e.g., logistic regression) is used to predict the presence of the queried action y from the features. The counterfactual causal effect sizes are given by importance attributions for features from $\mathcal{M}$, giving a mechanistic selection and ranking of features $\mathcal{F}^m_j \in \mathcal{H}$.

Teleological explanations are formulated in terms of the intrinsic reward components of the ego agent. For this explanatory mode, the counterfactual simulations inform us how the rewards of the ego, as measured by the reward vector $r \in \mathbb{R}^d$, change depending on the presence y of the queried action of the ego. For binary y, this means that Algorithm 2 splits $\mathcal{D}$ into two disjoint sets: one where the queried action was observed ($y = 1$) and one where it was not ($y = 0$). Following the average treatment effect for randomized controlled trials [4] we take the difference between the expected reward vectors of each set, then order the elements of the difference decreasingly by absolute value, giving a teleological ordering of reward components $\mathcal{R}^t \in \mathcal{H}$ by the causal effect of y.

4 APPLICATION TO MOTION PLANNING

We give a full demonstration of CEMA's capabilities by applying it to the problem of motion planning for AD which is a challenging reasoning task due to the tightly coupled interactions of many agents in a dynamically evolving system [35]. Specifically, we use CEMA to automatically explain the decisions of the Interpretable Goal-based Prediction and Planning (IGP2) system for AD [1]. We give a summary of IGP2 to the extent necessary for the following sections, but for full details please refer to the original paper.

The local state s^i of a vehicle i contains its pose (position and heading), velocity, and acceleration A sequence of temporally adjacent local states is called a trajectory. Local observation o^i contain the local states of nearby traffic participants. Actions a^i set low-level controls such as acceleration and steering, while goals G^i are spatial destinations. Reward components $\mathcal{R}$ are longitudinal and lateral acceleration, presence of collisions, time to reach a destination, and goal completion. IGP2 uses a hierarchy of systems rather than an end-to-end architecture. It defines a set of action sequence templates called *maneuvers* with dynamically generated trajectories for vehicles to follow, including `lane-follow`, `lane-change-{left,right}`, `turn-{left,right}`, `give-way`, and `stop`. Common sequences of maneuvers are then further chained into high-level *macro actions*: `Continue`, `Change-{Left,Right}`, `Exit`, and `Stop`.

IGP2 uses macro actions to predict for each non-ego vehicle i a joint distribution over possible goals and future trajectories given the observed joint local states $s_{1:t}$. Monte Carlo Tree Search (MCTS) is then used to forward simulate the world and obtain driving trajectories for the ego vehicle. In every MCTS simulation, the previously predicted joint goal and trajectory distribution is used to

Table 1: Binary features $\mathcal{F}$ to describe the fundamental motions and high-level actions of vehicles (including ego). For continuous values, the mean value is calculated along the length of the trajectory and thresholded with small value δ.

Feature	Calculation	Explanation
Acceleration	$a^i > \delta_a$	Accelerate
	$a^i < -\delta_a$	Decelerate
	$a^i \in [-\delta_a, \delta_a]$	Maintain velocity
Relative	$v^i - v^\varepsilon > \delta_v$	Faster than ego
speed	$v^i - v^\varepsilon < -\delta_v$	Slower than ego
	$v^i - v^\varepsilon \in [-\delta_v, \delta_v]$	Same speed as ego
Stop	$v^i \in [0, \delta_s]$	Does it stop
Maneuver	One-hot encode	Longest maneuver
Macro Action	One-hot encode	Longest macro action

randomly sample a goal and corresponding trajectory for each non-ego vehicle. MCTS generates a trajectory for the ego in a simulation by sequentially choosing macro actions based on backpropagated preference values (i.e., Q-values) until the ego reaches its goal.

4.1 Implementing CEMA

We define our set of features $\mathcal{F}$ in Table 1, which were chosen to describe both fundamental motions and high-level maneuvers of all vehicles including the ego. Features average along the length of the trajectory and may encounter the issue that at one timestep they have a positive causal effect, while at a later timestep, they have a negative causal effect, resulting in aggregate zero causal effect. The slicing operation in Algorithm 2 assures that this issue is avoided.

To focus on causal selection and avoid the ambiguities of natural language, we hand-code each query q to contain a description of the queried subsequence $\hat{s}_{u:v}$ given as a subset of features from $\mathcal{F}$. For natural language generation, we use a deterministic realization engine called SimpleNLG [12], which generates a grammatically correct English sentence from a content specification, e.g., subject and verb. This a better fit than neural generation algorithms, due to a lack of annotated data and hallucinations in neural models.

Since IGP2 can assign to some (reachable) goals and trajectories near-zero probabilities, we use additive smoothing – detailed in Appendix A.3 – with parameter α to make sure every goal and trajectory can be sampled for the non-ego vehicles. We then generate two datasets with Algorithm 1. For teleological explanations, we set $\tau = u$, rolling back time just before the queried action of the ego. This is because teleological explanations are determined by the MCTS reward components which only depend on the ego's present and future actions. For mechanistic explanations, we set τ to the start time of the last action prior to u, erasing both the queried action of the ego and the action that came before it. For slicing the trajectories in Algorithm 2, we set $P \leftarrow (u, n)$ which slices the trajectory $\hat{s}_{\tau+1:n}$ into a past $\hat{s}_{\tau+1:u}$ and present-future $\hat{s}_{u:n}$ subsequence in reference to the start of the ego's queried action.

We use feature weights from logistic regression with K-fold cross-validation to determine feature importance values. We found logistic regression to work best as it is simple, inherently interpretable, and all features are binary so their scale does not affect the importance values.

5 COMPUTATIONAL EVALUATION

We evaluate CEMA on the four scenarios (S1–S4) used by Albrecht et al. [1]. The scenarios are shown in Figure 3 with expert explanations of the ego's behavior by Albrecht et al. [1] In line with our focus on social XAI, we test CEMA on many user queries regarding different ego agents and behaviors, and the generated outputs of CEMA are presented through five simulated conversations (Table 2), highlighting CEMA's ability to correctly identify the causes behind each queried action. For all queries, we simulate $K = 100$ counterfactual worlds with a smoothing weight $\alpha = 0.1$. Further details about the experimental setup are given in Appendix B. We focus on S1 for presentation, but all results are confirmed across all scenarios and all presented in Appendix C. We show that:

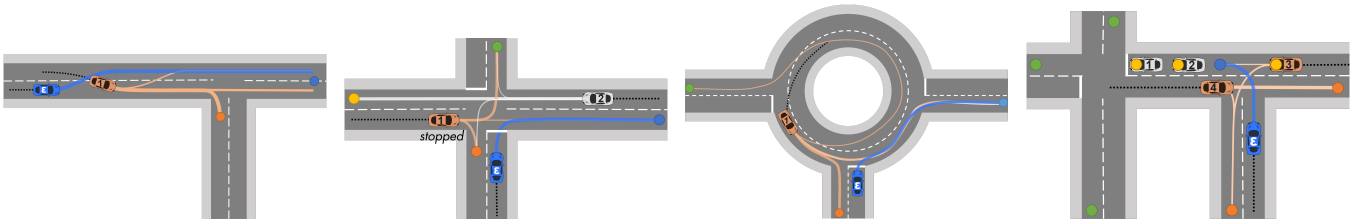
(1) CEMA correctly finds and ranks the relevant causes of the ego's actions that are congruent with expert explanations;
(2) It correctly identifies the relevant causes even in the presence of a large number of agents;
(3) The causal selection process is robust to changes in the sampling size K and the accuracy of the probabilistic model.

5.1 Correctness of Causal Selection

As shown in Table 2, CEMA correctly selects causes which are congruent with the expert explanations of Albrecht et al. [1].

In conversation S1-A, the causes behind the factual lane change of the ego are queried. The top plot in Figure 4 shows that CEMA correctly finds that a decrease in time-to-goal is the most significant teleological cause. As the bottom plot in Figure 4 shows, CEMA correctly identifies that the non-ego slowing down is a mechanistic cause of the ego's lane change. CEMA also determines that this slowing down is due to the non-ego vehicle decelerating in order to turn right. The middle plot of Figure 4 confirms that the initially faster non-ego vehicle cutting in front of the ego is also a mechanistic cause of the ego's lane change. This shows the importance of slicing the trajectories into segments as CEMA produces more fine-grained causes that focus on action in a particular time interval.

In conversations S1-B to S4, we also see that CEMA correctly identifies causes for contrastive questions – for example, *"Why aren't you going straight?"* – in which the user asks about an alternative action (i.e., foil) that the ego could have done as opposed to the factual observed actions (i.e., fact). Leveraging the counterfactual simulations, CEMA contrasts the simulations containing the foil to simulations containing the fact and derives the appropriate teleological causes. CEMA delivers consistent explanations even when queries target the same action but are phrased differently. For example, *"Why will you change lanes?"* is a direct question, while *"Why aren't you going straight?"* is contrastive, yet they both refer to the same changing lane action of the ego and CEMA finds consistent causes for both queries. In S4, CEMA correctly finds that the stopping of non-ego 3 is the most relevant cause behind the ego's early merging behavior and it also finds other intuitive causes. For example, the vehicle at the front of the waiting line of cars is



(S1) The non-ego in front of the ego changes lanes and begins to slow down. This is indicative of its intention to turn right at the junction. To avoid being slowed down, the ego decides to change lanes as it is heading straight.

(S2) The ego is turning right but must give way. It observes the vehicle on the left stopping. This is only rational if it is trying to turn left and is giving way for the oncoming vehicle. The ego can use this to enter the road earlier.

(S3) The ego observes the non-ego changing lanes to the right. This is only rational if the non-ego is leaving the roundabout at the next exit. The ego can therefore enter the roundabout faster without waiting to give way.

(S4) Non-ego 3 is slowing down to stop. Once non-ego 4 drives past as indicated by its maintained high speed, the stopping of non-ego 3 stays rational only if it is to allow the ego to merge without waiting for non-ego 4 to pass.

Figure 3: The four scenarios used for evaluation based on Albrecht et al. [1]. Colored circles are goals. Solid lines are predicted trajectories of non-egos with thickness corresponding to predicted probability. Black dotted lines are observations.

stopped. Would this vehicle move, the waiting line of cars would begin moving and non-ego 3 could not allow the ego to merge.

CEMA can also correctly find the relevant causes even when a large number of agents are present. For this, we greatly increase the number of agents in all scenarios and rerun CEMA. For example, we extend S1, adding two extra lanes to the east-west road and increasing the number of agents to 20. This gave 180 features, most of which had no causal influence on the ego, but CEMA could still identify the most important causes as in the original scenario.

5.2 Robustness of Causal Selection

We demonstrate robustness to changes in **(a)** the sampling size K, and **(b)** in the accuracy of the probabilistic simulation model, to show that correct explanations are generated even when sampling is limited by resources and that our system works with prediction algorithms of varying performance. For size robustness, we randomly sample a dataset of $K \in \{5, 10, \ldots, 100\}$ sequences 50 times and calculate the causal attributions for each dataset. For robustness, we interpolate between the true predicted and uniformly distributed behaviors by increasing the smoothing strength α on a log scale.

The top plot in Figure 5 shows the evolution of causal attributions as we increase K in S1. We see that CEMA becomes increasingly confident in its attributions as K increases, while confidence intervals remain tight. Even with few samples, CEMA identifies causes correctly. The bottom plot of Figure 5 shows how causal attributions change as α increases which corresponds to increasing uncertainty in behavior predictions. We see that feature importance values are little affected by changes in the sample distributions as they fluctuate around the same values. Similar patterns are observed across scenarios, which demonstrates that CEMA is robust to changes in both the sampling size and the accuracy of external predictions.

6 USER STUDY

So far, we have focused on the technical details of CEMA. Ultimately, however, the primary target of CEMA is non-expert end users, so we must evaluate the quality of CEMA's explanations and their various effects on humans with actual participants via a user study. We aim to answer the following research questions:

(1) How do people perceive the quality of CEMA's explanations as compared to a human baseline?
(2) What are the effects of CEMA's explanations on people's trust in autonomous vehicles?

We used Prolific to recruit participants from the USA whose first language is English. As most people have not had first-hand experience with autonomous vehicles (AV), we engaged them via animated videos of the scenarios. We design two surveys and summarize our methodology below with full details in Appendix D.

In the first survey (N=54; Male=25, Female=29), participants were asked to describe and explain in their own words the behavior of the AV in all four scenarios. This gave 408 explanations across scenarios, of which we excluded 26 vacuous responses (e.g., *"I don't know", "None", etc.*), and annotated the remaining explanations with a different set of participants regarding their causal content, overall quality and complexity, and trustworthiness. We release an extended version of this annotated dataset of natural language explanations, called the *Human Explanations for Autonomous Driving Decisions (HEADD)* dataset, containing 14 scenarios with several agents and environmental elements, including occlusions, pedestrians, and 1308 explanations. We collected explanations as we are not aware of any reproducible and publicly available methods for AD that would allow for a meaningful comparison to CEMA's explanations.[2] Comparing against a human baseline is also a better fit for CEMA as its explanations are intended to have low cognitive overhead and be easy to understand. In contrast, more complex expert explanations would likely be less effective for end users [11].

In the second survey (N=200; M=99, F=101), we designed two tasks, one for each research question. First, to measure the quality of explanations, we asked participants to rate a random sample of 10 explanations from a set of 30 explanations (5 from CEMA and 25 from HEADD with the highest quality ratings) for each scenario on a 5-point Likert scale. With 50% chance, we highlighted

[2]We explored ChatGPT as a baseline, but it was inadequate as its responses were very inconsistent and only sometimes correct (see Appendix D.4 for details).

Table 2: Actual responses of CEMA to queries (*in italic*). (S1-A) The passenger sees on an onboard display, that the ego is planning to change to the left lane. They find this unexpected and inquire. (S1-B) The passenger observes the ego changing lanes and asks for the reasons behind the maneuver while it is ongoing. (S2) The passenger observed that the ego had entered the junction without stopping to give way, which could be dangerous. (S3) The passenger sees that the ego will turn right at the roundabout without stopping despite the oncoming vehicle. (S4) Once non-ego 4 has passed, the ego merges onto the main road which makes the passenger wonder why the ego did not give way to non-ego 3.

Scenario 1 (S1-A)	**Scenario 1 (S1-B)**
Why will you change lanes? It will decrease the time to goal. *Why does it decrease the time to the goal?* Because vehicle 1 will be slower than us. *Why will it be slower?* It will decelerate and turn right. *What if it hadn't changed lanes before?* We would've gone straight.	*Why aren't you going straight?* It would increase the time to the goal. *Why is it slower to go straight?* Because vehicle 1 is slowing down. *Any other reasons?* It is slower than us. *What will you do after this?* We will go straight and accelerate.
Scenario 2 (S2)	**Scenario 3 (S3)**
Why did you not stop to give way? It would've increased the time to the goal. *Couldn't it have caused a collision?* A collision would not have occurred. *How did you know?* Vehicle 1 was turning left and stopped. *What if it went straight?* We would've given way and slowed down.	*What will you do at the roundabout?* We will turn right and accelerate. *Would not stopping lead to a collision?* Not stopping doesn't cause a collision, but stopping increases the time to reach the goal. *How do you know we won't collide with the oncoming car?* It has been changing lanes right and is turning right.
Scenario 4 (S4)	
Why are you not stopping to give way? Stopping and giving way would increase our time to reach the goal. *Is it safe to turn left early?* Accelerating and turning left does not cause a collision. *Why not?* Because vehicle 3 stops. *What if vehicle 3 went straight?* We would slow down and give way.	

in bold the explanations from CEMA. Second, to measure trust, we use the 9 trust scales proposed by Hoffman et al. [20] adapted to the AD domain. We use a between-subjects design: half the participants are shown the trust scales prior to ranking explanations,

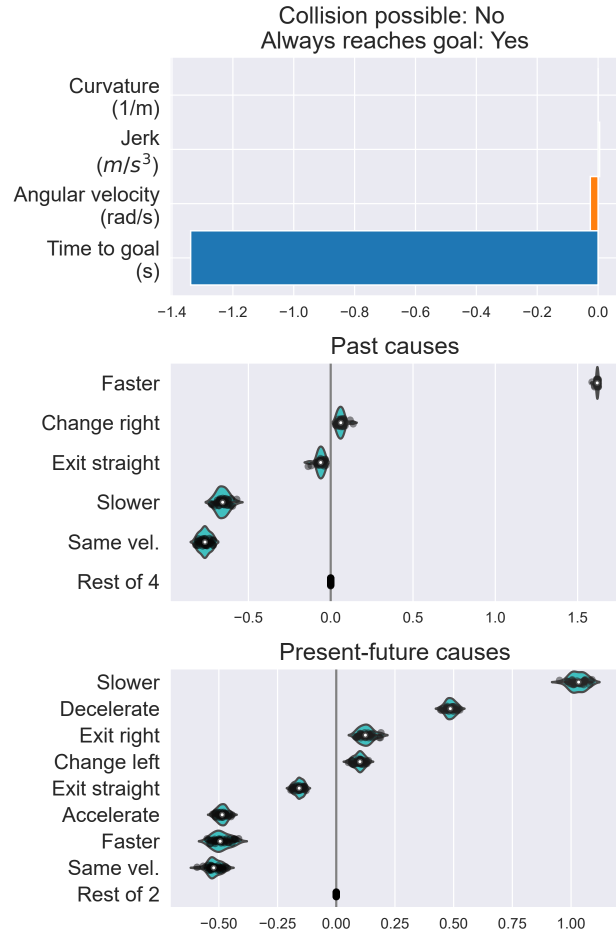


Figure 4: [Top] Signed differences between expected reward components correctly identify time-to-goal as the most significant teleological cause. [Mid/Bot] Feature importance attributions for the slice before and during/after the queried subsequence correctly rank mechanistic causes. Violin plots show 5-fold cross-validation repeated 7 times.

and the other half after having ranked explanations. We also asked participants about their driving experience and previous exposure to AVs using the SAE automation scale [34]. We hypothesize that **H1:** the explanations generated by CEMA are scored on average as highly as the human baseline explanations; **H2** participants who saw explanations from CEMA have on average higher levels of trust than those who have not. We analyze our data by fitting linear mixed-effects models for each hypothesis. We report the estimated means ($\hat{\beta}$) and standard errors (σ) for each variable and use the Wald test [41] to determine whether the effects of a variable are statistically significant on the outcome.

For **H1**, we found that CEMA's explanations were rated significantly higher when its explanations were not highlighted and were not significantly worse when they were highlighted. On average, explanation ratings ($\hat{\beta}_0$=3.31, σ=0.08) were marginally lower for human-written explanations ($\hat{\beta}$=−0.16, σ=0.08, p=0.21), and ratings

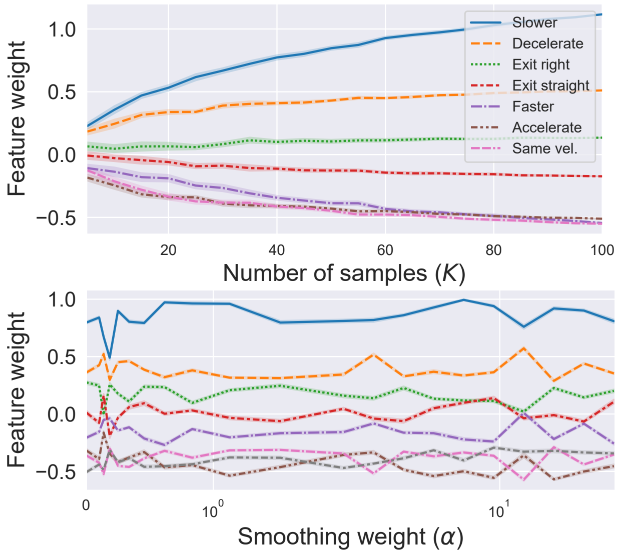


Figure 5: Changes to causal attributions with [Top] different sample sizes and [Bot] different smoothing weights for present-future mechanistic causes in conversation S1-A. Shaded regions are bootstrapped 95% confidence intervals.

were significantly lower for human explanations when CEMA's explanations were not highlighted to participants ($\hat{\beta}$=−0.22, σ=0.08, $p < 0.05$). Variations across scenarios were negligible (SD=0.07). We also found that people tend to rank CEMA's explanations higher when they had exposure to AVs previously ($\hat{\beta}$=0.1, σ=0.06, p = 0.09).

For **H2**, we found that, on average, participants' trust ratings ($\hat{\beta}_0$=1.53, σ=0.5) were significantly higher after seeing explanations ($\hat{\beta}$=0.11, σ=0.05, $p < 0.05$), which aligns with expectations from literature [30]. Participants' trust also increased significantly when they rated CEMA's explanations higher ($\hat{\beta}$=0.35, σ=0.15, $p \approx 0$) or when they had previous exposure to AVs ($\hat{\beta}$=0.33, σ=0.05, $p \approx 0$), but trust remained largely unchanged by human explanations ($\hat{\beta}$=0.12, σ=0.15, p=0.85). Trust ratings were not significantly affected by whether CEMA's explanations were highlighted ($\hat{\beta}$=0.02, σ=0.05, p = 0.66) and there were no significant interaction effects between the average ratings of explanations and highlighting ($\hat{\beta}$=−0.04, σ=0.04, p = 0.36). The estimated trust levels varied across the 9 trust scales (SD=0.53) but not the observed tendencies. Our results suggest that people who had some exposure to AVs or had a preference for CEMA's explanations were more likely to trust AVs in general, regardless of whether they knew which explanations came from CEMA. Taken together with the result for **H1**, this suggests that CEMA's explanations may be more effective at improving people's trust in AVs than non-expert human explanations.

7 DISCUSSION AND FUTURE WORK

Our primary goal with CEMA is to advance the field of social XAI applied to dynamic multi-agent systems. A crucial component of intelligible explanations is the use of semantically meaningful features [11]. Importantly, the challenge of designing useful features is not unique to CEMA but is a necessary step for any automated explanation generation system in social XAI. With CEMA, we assumed that there is a feature function ϕ which performs the translation from the raw representations of state and action spaces to the more abstract semantic feature space. This translation from state to feature space is domain-dependent and should be considered a crucial step during the deployment of social XAI systems. However, CEMA is feature-agnostic so that counterfactual causal selection does not depend on ϕ or the interpretations of features.

CEMA also does not rely on a fixed causal graph to model dynamic multi-agent systems. Instead, it assumes that there is a probabilistic model, such as a stochastic planner, trained joint policy, or autoregressive model trained on observational data, which can be used to forward simulate the state of the system. Based on the work of Quillien and Lucas [33] and the counterfactual model of causation [19, 24], CEMA can derive causes to an ego agent's actions in any system where such a model is obtainable. The assumption here is that these models cover alternatives that are grounded in factual observations with a non-zero probability, and any reasonably expressive algorithm would fulfill these criteria.

The user study suggests that people may prefer explanations generated by CEMA, however, trust levels are still low. This may be – as several participants indicated in their feedback – because people prefer to see agents act more conservatively, without exploiting potentially riskier but more efficient actions. Explanations that justify efficient but less safe decisions then have to overcome the inherent wariness of people, which was indeed high among participants, though it somewhat decreased after seeing explanations.

We designed CEMA to be used in conversations with users, but we did not focus on natural language processing in this work. For example, we assume that queries unambiguously describe the timing of actions – allowing us to focus on causal selection – but actual natural language queries are fuzzy and imprecise. By building modern NLP components, we can strengthen the social and conversational aspects of CEMA. Future work will involve the integration of language parsing [25] and dialogue systems [10] leveraging modern neural language models to deliver explanations.

Our implementation of CEMA for AD improves on existing social XAI methods for AD in several aspects. In contrast to Omeiza et al. [30], we avoid using a surrogate model and generate causal explanations that take the temporal nature of driving into account. Compared to Gyevnar et al. [14], CEMA supports multiple modes of explanations with both high-level and low-level features.

To conclude, our goal is to address some of the transparency-related social concerns of AI. CEMA fills a gap in social XAI by enabling causal explanation generation in dynamic sequential multi-agent systems. As we expect to see autonomous agents proliferate in everyday environments, social explanations will be crucial for building user trust and for the acceptance of new technologies.

ACKNOWLEDGMENTS

The authors thank C. Brewitt, M. Tamborski, and Y. Ziser for their helpful comments on earlier drafts. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing (grant EP/S022481/1) and the Edinburgh Laboratory for Integrated Artificial Intelligence.

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)*.

[2] Stefano V. Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66–95. https://doi.org/10.1016/j.artint.2018.01.002

[3] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Information Fusion* (April 2023), 101805. https://doi.org/10.1016/j.inffus.2023.101805

[4] Peter C. Austin. 2011. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research* 46, 3 (May 2011), 399–424. https://doi.org/10.1080/00273171.2011.568786

[5] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B. Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M. Paixão, Filipe Mutz, Lucas de Paula Veronese, Thiago Oliveira-Santos, and Alberto F. De Souza. 2021. Self-Driving Cars: A Survey. *Expert Systems with Applications* 165 (March 2021), 113816. https://doi.org/10.1016/j.eswa.2020.113816

[6] Cillian Brewitt, Balint Gyevnar, Samuel Garcin, and Stefano V. Albrecht. 2021. GRIT: Fast, Interpretable, and Verifiable Goal Recognition with Learned Decision Trees for Autonomous Driving. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1023–1030. https://doi.org/10.1109/IROS51168.2021.9636279

[7] Cillian Brewitt, Massimiliano Tamborski, Cheng Wang, and Stefano V. Albrecht. 2023. Verifiable Goal Recognition for Autonomous Driving with Occlusions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[8] Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research* 70 (May 2021), 245–317. https://doi.org/10.1613/jair.1.12228

[9] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2020. The Emerging Landscape of Explainable Automated Planning & Decision Making. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4803–4811. https://doi.org/10.24963/ijcai.2020/669 Survey track.

[10] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (Nov. 2017), 25–35. https://doi.org/10.1145/3166054.3166058

[11] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. 2021. Levels of Explainable Artificial Intelligence for Human-Aligned Conversational Explanations. *Artificial Intelligence* 299 (Oct. 2021), 103525. https://doi.org/10.1016/j.artint.2021.103525

[12] Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*. Association for Computational Linguistics, Athens, Greece, 90–93.

[13] Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. 2023. Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*. IOS Press, 964–971. https://doi.org/10.3233/FAIA230367

[14] Balint Gyevnar, Massimiliano Tamborski, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2022. A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning. In *Workshop on Artificial Intelligence for Autonomous Driving*. International Joint Conference on Artificial Intelligence. https://doi.org/10.48550/arXiv.2206.08783 arXiv:2206.08783 [cs]

[15] Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2024. HEADD: Human Explanations for Autonomous Driving Decisions. https://doi.org/10.7488/ds/7676

[16] R. J. Hankinson. 1998. *Cause and Explanation in Ancient Greek Thought*. Clarendon Press.

[17] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[18] Jacob Haspiel, Na Du, Jill Meyerson, Lionel P. Robert Jr., Dawn Tilbury, X. Jessie Yang, and Anuj K. Pradhan. 2018. Explanations and Expectations: Trust Building in Automated Vehicles. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery, New York, NY, USA, 119–120. https://doi.org/10.1145/3173386.3177057

[19] Denis J. Hilton. 1988. Logic and Causal Attribution. In *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*. New York University Press, New York, NY, US, 33–65.

[20] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608 [cs]* (Feb. 2019). arXiv:1812.04608 [cs]

[21] Renhao Huang, Hao Xue, Maurice Pagnucco, Flora Salim, and Yang Song. 2023. Multimodal Trajectory Prediction: A Survey. arXiv:2302.10463 [cs.RO]

[22] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy Artificial Intelligence: A Review. *Comput. Surveys* 55, 2 (Jan. 2022), 39:1–39:38. https://doi.org/10.1145/3491209

[23] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Computer Vision – ECCV 2018 (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 577–593. https://doi.org/10.1007/978-3-030-01216-8_35

[24] David Lewis. 1973. Causation. *Journal of Philosophy* 70, 17 (1973), 556–567. https://doi.org/10.2307/2025310

[25] Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022. A Survey of Discourse Parsing. *Frontiers of Computer Science* 16, 5 (Jan. 2022), 165329. https://doi.org/10.1007/s11704-021-0500-z

[26] Tania Lombrozo and Susan Carey. 2006. Functional Explanation and the Function of Explanation. *Cognition* 99, 2 (March 2006), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009

[27] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable Reinforcement Learning through a Causal Lens. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 03 (April 2020), 2493–2500. https://doi.org/10.1609/aaai.v34i03.5631

[28] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[29] Samer B. Nashed, Saaduddin Mahmud, Claudia V. Goldman, and Shlomo Zilberstein. 2023. Causal Explanations for Sequential Decision Making Under Uncertainty. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2307–2309.

[30] Daniel Omeiza, Helena Web, Marina Jirotka, and Lars Kunze. 2021. Towards Accountability: Providing Intelligible Explanations in Autonomous Driving. *Proceedings of the 32nd IEEE Intelligent Vehicles Symposium* (2021).

[31] Judea Pearl. 2009. *Causality* (second ed.). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511803161

[32] Yunpeng Qing, Shunyu Liu, Jie Song, and Mingli Song. 2022. A Survey on Explainable Reinforcement Learning: Concepts, Algorithms, Challenges. arXiv:2211.06665 [cs]

[33] Tadeg Quillien and Christopher G. Lucas. 2023. Counterfactuals and the Logic of Causal Selection. *Psychological Review* Advance online publication (2023). https://doi.org/10.1037/rev0000428

[34] SAE International. 2021. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Technical Report. United States.

[35] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. 2018. Planning and Decision-Making for Autonomous Vehicles. *Annual Review of Control, Robotics, and Autonomous Systems* 1, 1 (2018), 187–210. https://doi.org/10.1146/annurev-control-060117-105157

[36] Andrew Silva, Matthew Gombolay, Taylor Killian, Ivan Jimenez, and Sung-Hyun Son. 2020. Optimization Methods for Interpretable Differentiable Decision Trees Applied to Reinforcement Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, 1855–1865.

[37] 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. https://doi.org/10.1109/ACCESS.2021.3051315

[38] Stratis Tsirtsis, Abir De, and Manuel Rodriguez. 2021. Counterfactual Explanations in Sequential Decision Making Under Uncertainty. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 30127–30139.

[39] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. 2019. Programmatically Interpretable Reinforcement Learning. *arXiv:1804.02477 [cs, stat]* (April 2019). arXiv:1804.02477 [cs, stat]

[40] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)* 31, 2 (2017), 841–888.

[41] Abraham Wald. 1943. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Trans. Amer. Math. Soc.* 54, 3 (1943), 426–482. http://www.jstor.org/stable/1990256

[42] Yiwen Zhang, Wenjia Wang, Xinyan Zhou, Qi Wang, and Xiaohua Sun. 2022. Tactical-Level Explanation Is Not Enough: Effect of Explaining AV's Lane-Changing Decisions on Drivers' Decision-Making, Trust, and Emotional Experience. *International Journal of Human–Computer Interaction* 0, 0 (Aug. 2022), 1–17. https://doi.org/10.1080/10447318.2022.2098965