A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning

Balint Gyevnar[1,*], **Massimiliano Tamborski**[1], **Cheng Wang**[1], **Christopher G. Lucas**[1], **Shay B. Cohen**[1], **Stefano V. Albrecht**[1,2]
[1]School of Informatics, University of Edinburgh
[2]Five AI Ltd., UK
{balint.gyevnar,cheng.wang,c.lucas,s.albrecht}@ed.ac.uk,
m.tamborski@sms.ed.ac.uk, scohen@inf.ed.ac.uk

Abstract

Inscrutable AI systems are difficult to trust, especially if they operate in safety-critical settings like autonomous driving. Therefore, there is a need to build transparent and queryable systems to increase trust levels. We propose a transparent, human-centric explanation generation method for autonomous vehicle motion planning and prediction based on an existing white-box system called IGP2. Our method integrates Bayesian networks with context-free generative rules and can give causal natural language explanations for the high-level driving behaviour of autonomous vehicles. Preliminary testing on simulated scenarios shows that our method captures the causes behind the actions of autonomous vehicles and generates intelligible explanations with varying complexity.

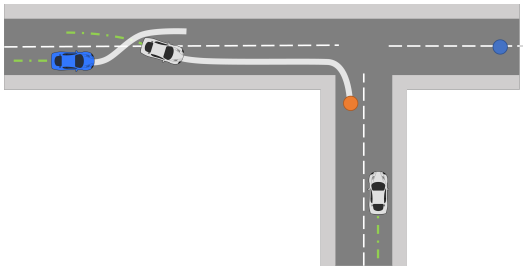

Figure 1: The ego vehicle (in blue) is heading straight to the blue goal but then changes lanes to the left. A passenger may inquire *"Why did you change lanes instead of just driving straight?"*. Our system uses the motion planning and prediction calculations of the AV to give a causally justified contrastive explanation: if the ego had gone straight, then it would have likely reached its goal slower because the vehicle in front is probably changing lanes right and then exits right.

1 Introduction

Autonomous vehicles (AVs) are predicted to improve, among other things, traffic efficiency and transport safety, reducing road fatalities possibly by as much as 90% [Wang *et al.*, 2022]. AVs are also predicted to decrease pollution and make transportation more accessible for passengers with disabilities. However, the current complex, highly-integrated, and often opaque systems of AVs are not easily (or at all) understood by most humans. This opaqueness often manifests in reluctance to accept the technology due to concerns that the vehicle might fail in unexpected situations [Hussain and Zeadally, 2019]. This have fostered continued distrust and scepticism of AVs in the public eye [Kim and Kelley-Baker, 2021].

We need to build trust in passengers if we want to overcome these psychological barriers and achieve wider acceptance for AVs. Crucial to the development of such trust, but neglected since the rise of black-box algorithms is the principle of explicability. This principle broadly means that the purposes, capabilities, and methods of the AV system must be **transparent**, that is, understandable and queryable by its passengers. While this principle is generally important for any AI system, it is especially important for AVs as they operate in safety-critical settings and their decisions have far-reaching consequences on human lives. There is a scientific consensus that we can increase transparency and build trust in AVs through the adoption of human-centric explainable AI (XAI) [Hussain and Zeadally, 2019; Omeiza *et al.*, 2021b; Atakishiyev *et al.*, 2021]. Humans prefer **causal** explanations [Miller, 2019], so our explanations must be causally justifiable in terms of the processes that generated our actions. We also want explanations to be **intelligible** for non-expert people to minimise cognitive overhead and help build general knowledge about AVs, reducing scepticism. Finally, explanations must be faithful to the decision generating process to ensure they are not misleading. We call this property the **soundness** of an explanation generation system.

We propose a human-centric *explanation generation method* called eXplainable Autonomous Vehicle Intelligence (XAVI), focusing on high-level *motion planning and prediction*. XAVI is based on an existing transparent, inherently interpretable motion planning and prediction system called IGP2 [Albrecht *et al.*, 2021]. An example of the output of our system is shown in Figure 1. Our method models the cause-effect relations behind the decisions of IGP2 as a Bayesian network allowing us to draw probabilistic causal judgements about the plans of the AV. These judgements inform a context-free grammar that is used to generate intelligible *contrastive explanations* (see Section 2), which compare the factual observed behaviour

*Contact author.

with some counterfactual behaviour. Preliminary testing of XAVI on driving scenarios with baseline explanations by the authors of IGP2 demonstrates that our method correctly captures some of the causality behind the actions of the AV and generates intelligible natural language explanations with varying complexity. We end with a discussion outlining important future work for explaining AV behaviour. We also release the code for XAVI on GitHub[1].

2 Background

Most existing methods of XAI are model-agnostic approaches that focus on classification/regression tasks relying on black-box and surrogate models. Local surrogate models usually calculate some feature importance ordering given a *single input* instance for a given black box model [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Montavon *et al.*, 2017]. However, instance-based explanations may ignore the overall workings of the black box and, when interpreted incorrectly, may build an incorrect understanding of our system in humans. Global surrogate models instead use black box models as a teacher to train simpler, interpretable white box systems [Bastani *et al.*, 2017; Lakkaraju *et al.*, 2017]. These can model the overall decision process of the teacher, though usually at the cost of soundness. In general, surrogate approaches have to introduce a new layer of abstraction that does not allow or distort the causal understanding of the decision process of the underlying black box and may introduce unwanted biases to the explanations. In addition, the output of these methods is usually expert oriented and difficult to understand for non-experts.

These shortcomings motivated several recent work that popularise a human-centric approach to explanation generation based on causality and intelligibility [Miller, 2019; Dazeley *et al.*, 2021; Ehsan and Riedl, 2020]. In the case of classical AI planning, XAI-PLAN [Borgo *et al.*, 2018] answers contrastive questions of the form "Why do X instead of Y?", while WHY-PLAN [Korpan and Epstein, 2018] generates natural language explanations based on model reconciliation, which compares the generated plan of the system to a user-given alternative plan. These methods represent a shift towards a more human-centric approach, however the main issue with classical AI planning methods is their reliance on fixed domain descriptions which make their use in dynamic and complex environments such as autonomous driving difficult.

Furthermore, while the motivation for building trust and transparency for AVs is well understood, few works have proposed methods that use AV domain knowledge to inform their explanation generation system. Previous methods used deep learning to generate textual explanations for AVs based on a data set of annotated recordings with textual explanations called BDD-X [Kim *et al.*, 2018; Ben-Younes *et al.*, 2022]. Additionally, Omeiza *et al.* (2021a) proposed an explanation generation system based on decision trees taught by a black box and using language templates. These methods generate intelligible explanations, but the generating processes are surrogate models which are neither causal nor transparent.

Recently, Albrecht *et al.* (2021) proposed an inherently interpretable integrated planning and prediction system called IGP2. This method relies on intuitive high-level driving actions and uses rational inverse planning to predict the trajectories of other vehicles, which are then used to inform motion planning with Monte Carlo Tree Search (MCTS). In this work, we rely on IGP2 as it is a white-box model, whose internal representations can be directly accessed while its decisions can be readily interpreted through rationality principles. Direct access to internal representations means access to the MCTS tree search which naturally allows for causal interpretation. We directly leverage this inherent causality to build our method.

3 IGP2: Interpretable Goal-Based Prediction and Planning for Autonomous Driving

In the following, we give a brief introduction to the notation and methods of IGP2. Let $\mathcal{I}$ be the set of traffic participants in the local neighbourhood of the ego vehicle denoted $\varepsilon \in \mathcal{I}$. At time step t each traffic participant $i \in \mathcal{I}$ is in a local state $s_t^i \in \mathcal{S}^i$ which includes its pose (position and heading), velocity, and acceleration. The joint state of all vehicles is denoted $s_t \in \mathcal{S} = \times_i \mathcal{S}^i$, and the state sequence $(s_a, \ldots, s_b)$ is written $s_{a:b}$. A trajectory is defined as a state sequence, where two adjacent states have a time step difference of one. IGP2 is goal-oriented, so it assumes that each traffic participant is trying to reach one of a finite number of possible goals $g^i \in \mathcal{G}^i$. Trajectories can be used to calculate rewards r^i for a vehicle, where r^i is the weighted linear sum of reward components corresponding to aspects of the trajectory. The set of reward components is denoted by $\mathcal{C}$ and consists of *time-to-reach-goal*, *jerk*, *angular acceleration*, *curvature*, *collision*, and *termination* (received when IGP2 runs out of computational budget). Some reward components are mutually exclusive. For example, if we receive a (negative) "reward" for collision, then we cannot receive a reward for anything else.

The planning problem of IGP2 is to find an optimal policy for the ego vehicle ε that selects actions given a joint state to reach its goal g^ε while optimising its reward r^ε. Instead of planning over low-level controls, IGP2 defines higher-level manoeuvres with applicability and termination conditions, and (if applicable) a local trajectory $\hat{s}_{1:n}^i$ for the vehicle to follow. IGP2 uses the following manoeuvres: *lane-follow*, *lane-change-{left,right}*, *turn-{left,right}*, *give-way*, and *stop*. These manoeuvres are then further chained together into macro actions denoted here with $\omega \in \hat{\Omega}$, which are common sequences of manoeuvres parameterised by the macro actions. The set of all macro-actions is $\hat{\Omega} = \{$*Continue*, *Change-{left,right}*, *Exit*, *Continue-next-exit*, *Stop*$\}$. IGP2 searches for the optimal plan over these macro actions.

Finding the optimal plan for the ego vehicle has two major phases. First in the *goal and trajectory recognition phase* (referred to as goal recognition from here on), IGP2 calculates a distribution over goals $G^i \subseteq \mathcal{G}^i$ given the already observed trajectory of vehicle i denoted $p(G^i | s_{1:t}^i)$. To each goal we then generate a distribution over possible trajectories $S_{1:n}^i \subseteq \mathcal{S}_{1:n}^i$ given by $p(S_{1:n}^i | G^i)$. In the *planning phase*, goal recognition is used to inform a Monte Carlo Tree Search (MCTS) algorithm over macro actions, which finds the optimal sequence of macro actions (i.e. plan) by simulating many possible plans over K iterations to see how each plan interacts with the other

[1] https://github.com/uoe-agents/xavi-ai4ad

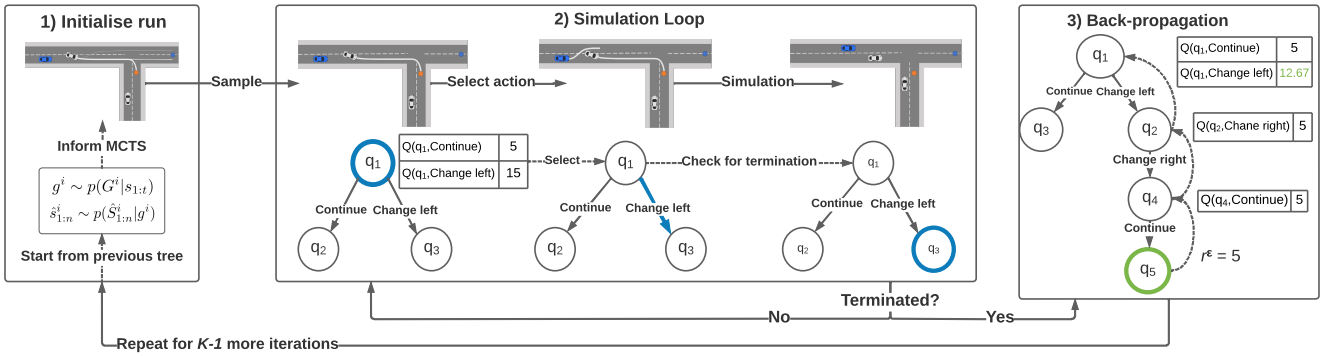

Figure 2: MCTS at work: *(Step 1)* Before each simulation, we sample and fix the trajectories of each non-ego vehicle. *(Step 2)* From the current state (in blue) we select our next macro action based on Q-values. In this example, we selected *Change-left*, which is then forward simulated while the other traffic participants follow their fixed trajectories. During simulation we check for termination conditions. *(Step 3)* If the ego reached a goal, or some other termination condition was met (e.g. collision), the ego receives a reward r^ϵ which is back-propagated through the trace of macro actions that reached the termination state to update the Q-values of each action. *(Step 4)* We repeat the process until K iterations are reached resulting in a search tree with maximal depth d_{max}.

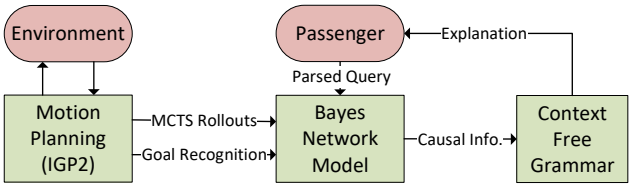

Figure 3: The XAVI system. IGP2 interacts with its environment and generates an optimal plan using MCTS and predictions from goal recognition. The accumulated information about how these components arrived at the optimal plan is used to build a Bayesian network (BN) model. We compare this model to a contrastive query from the passenger and extract causal relationships, which are fed to a context-free grammar that generates natural language explanations.

traffic participants. More details of MCTS can be found in the caption of Figure 2. During planning we track and accumulate all relevant information about how decisions are made which we then use to initialise our explanation generation system.

4 eXplainable Artificial Vehicle Intelligence

Though IGP2 is transparent and can be readily interpreted, it requires expert domain-knowledge to interpret its results and present intuitive explanations that are intelligible for the non-expert passenger. It is therefore desirable to automate the interpretation and explanation procedure in the human-centric way we outlined in Section 1.

We therefore present our *explanation generation system* called eXplainable Artificial Vehicle Intelligence (XAVI). The overall architecture of XAVI can be seen in Figure 3. The core idea of XAVI is to map the accumulated information about goal recognition and the steps of a complete MCTS planning run to random variables, which we then use to construct a Bayesian network (BN) model that encodes the properties of that particular MCTS planning run. This allows us to derive probabilistic causal information about alternative (i.e. counterfactual) sequences of macro actions and their possible effects on rewards and outcome.

Counterfactuals are a crucial part of XAVI, as the generated explanations *contrast* the factual, optimal plan with a counterfactual plan in which the ego would have followed a different sequence of macro actions. Contrastive explanations are studied in philosophical literature where most argue that all *why*-questions are (implicitly) contrastive [Miller, 2019]. This means that our generated explanations have a form similar to: "*If we had done <CF> [instead of <F>], then <EFFECTS> would have happened, because <CAUSES>*.". Here *<F>* and *<CF>* are the factual and counterfactual macro actions respectively, while *<EFFECTS>* are the changes to reward components and outcome in the counterfactual scenario. *<CAUSES>* describe relevant features of the trajectories of traffic participants (including the ego) that have caused the changes in *<EFFECTS>*. Note, we omit explicitly mentioning the factual *<F>* in our explanations since we assume that the passenger observed the ego's actions and is aware of what actions the ego had taken. We also assume, that the passenger's query is in a parsed format that allows us to directly extract counterfactual causal information from our Bayesian network model.

4.1 Bayes Network Model

Random Variables

The first step to create the Bayesian network model is to map MCTS planning steps to random variables. MCTS starts by sampling goals and trajectories for each non-ego vehicle i. Let the vector of random variables corresponding to goal sampling (we are not sampling for the ego) be $\mathbf{G} = [G^1, \ldots, G^{|\mathcal{I}|-1}]$ and the vector of trajectories be $\mathbf{S} = [S^1, \ldots, S^{|\mathcal{I}|-1}]$. The values of $G^i \in \mathcal{G}^i$ and $S^i \in \mathcal{S}^i_{1:n}$ are from the set of possible goals and trajectories for vehicle i. For example, setting $G^i = g^i$ means that we sample goal g^i for i.

Next, for every macro action selection step in the MCTS search tree, that is for each depth $1 \leq d \leq d_{max}$ in the tree, we define a random variable Ω_d with support of $\hat{\Omega}_d \subseteq \hat{\Omega}$ which is the set of all applicable macro actions at depth d.

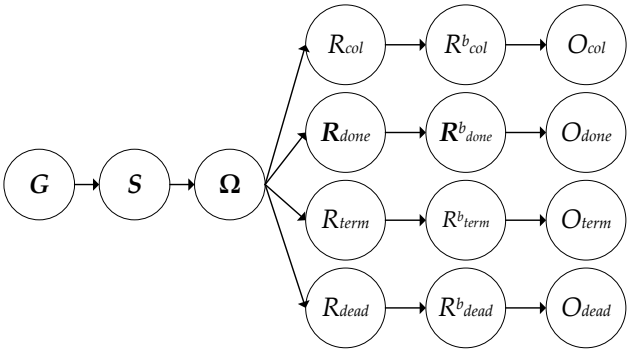

Figure 4: The underlying DAG of the Bayes network used for factoring the joint over all random variables. The chain-rule for $\mathbf{\Omega}$ is not shown explicitly. The associated reward components for O_{done} are $\mathbf{R}_{done} \in \{time, jerk, angular\text{-}acceleration, curvature\}$.

Each Ω_d may also take the value of the empty set $\varnothing$, which means that no action was selected at depth d. We collect each of these random variables into a single vector denoted $\mathbf{\Omega} = [\Omega_1, \ldots, \Omega_{d_{max}}]$. This means that specifying a trace in the search tree corresponds to assigning an action to each Ω_d which we can represent as a vector $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_{d_{max}}]$.

For each reward component $c \in \mathcal{C}$ that MCTS uses to calculate r^{ε} we can define a continuous random variable $R_c \in \mathbb{R}$ that gives the value for that particular reward component, or is $\varnothing$ if the reward component is not present. For each R_c, let $R_c^b \in \{0, 1\}$ be a binary variable that indicates the existence of reward component c. That is, if $R_c^b = 1$ then $R_c \neq \varnothing$. Let the vectors that collect the random variables for each component be $\mathbf{R} = [R_c]_{c \in \mathcal{C}}$ and similarly for $\mathbf{R}^b$.

Finally, we define outcome variables. It is important to note, that IGP2 does not explicitly represent various types of outcomes, so these variables do not correspond to any actual steps in MCTS. Instead, outcome variables are used here to conveniently describe the state of the ego vehicle at the termination of a simulation. There are four outcome types given by the set $\mathcal{O}$: *done* (g^{ε} was reached), *collision*, *termination* (reached d_{max} in MCTS without reaching g^{ε}), and *dead* (for any outcomes not covered by the previous three). For each outcome type $k \in \mathcal{O}$ we define a corresponding binary variable $O_k \in \{0, 1\}$ which indicates whether that outcome was reached at the termination of a MCTS simulation. The vector of outcome variables is denoted with $\mathbf{O}$.

Joint Factorisation

We now define the directed acyclic graph (DAG) to factorise the joint over random variables defined in the previous section and describe the probability distributions of each factor.

Consider the DAG shown in Figure 4. Goals for non-ego vehicles are sampled independently according to their distributions from goal recognition while the trajectories depend on the sampled goals. The joint distribution of these variables over each vehicle except the ego (hence iterating over $i \in \mathcal{I} \backslash \varepsilon$) are given below, which simply state that the probabilities of goals and trajectories of vehicles are mutually independent:

$$p(\mathbf{G}|\mathbf{s}_{1:t}) = \prod_{i \in \mathcal{I} \backslash \varepsilon} p(G^i | s^i_{1:t}), \tag{1}$$

$$p(\mathbf{S}|\mathbf{G}) = \prod_{i \in \mathcal{I} \backslash \varepsilon} p(\hat{s}^i_{1:n} | G^i). \tag{2}$$

Because trajectories of other traffic participants affect what macro actions are selected in MCTS, the random variables $\mathbf{\Omega}$ are conditioned on $\mathbf{S}$. Furthermore, the joint distribution of macro action selections $\mathbf{\Omega}$ is given by the chain rule, which corresponds to the product of macro action selection probabilities in the MCTS tree along a search trace:

$$p(\mathbf{\Omega}|\mathbf{S}) = p(\Omega_1|\mathbf{S}) \prod_{d=2}^{d_{max}} p(\Omega_d | \mathbf{\Omega}_{1:d-1}, \mathbf{S}). \tag{3}$$

The definition of $p(\Omega_d | \mathbf{\Omega}_{1:d-1}, \mathbf{S})$ corresponds to the probabilities of selecting a macro action at depth d from the unique state s reached by following $\mathbf{\Omega}_{1:d-1}$. For each value $\omega_d \in \Omega_d$ we estimate $p(\omega_d | \mathbf{\Omega}_{1:d-1}, \mathbf{S})$ from the K simulations of MCTS as the number of times ω_d was selected in state s over the total number of times any action was selected in state s (i.e. the total number of visits of state s).

Reward components in $\mathbf{R}$ depend on the driven trajectory of the ego, and therefore the selected sequence of macro actions given by $\mathbf{\Omega}$. However components are otherwise calculated independently from one another. The joint distribution for $\mathbf{R}$ is the product of distributions of each reward component:

$$p(\mathbf{R}|\mathbf{\Omega}) = \prod_{c \in \mathcal{C}} p(R_c | \mathbf{\Omega}). \tag{4}$$

If s now denotes the state reached by following $\mathbf{\Omega}$, we estimate $p(R_c|\mathbf{\Omega})$ from the K simulations as a normal distribution with sample mean $\mu_c(s)$ and sample variance $\sigma^2(s)$ calculated from the values observed for R_c in state s.

The existence indicator variables depend only on their corresponding reward component and their joint distribution otherwise assumes mutual independence:

$$p(\mathbf{R}^b|\mathbf{R}) = \prod_{c \in \mathcal{C}} p(R_c^b | R_c), \tag{5}$$

where $p(R_c^b = 1 | R_c) = 1$ iff $R_c \neq \varnothing$, so $R_c^b = 1$ only when we have observed some non-empty value for R_c.

Finally, the outcome variables depend only on the existence of certain reward components as given in Figure 4. For each outcome variable $k \in \mathcal{O}$ let $\mathbf{R}_k^b \subseteq \mathbf{R}^b$ be the vector of random variables of binary reward components that k depends on. The joint distribution of outcomes is mutually independent:

$$p(\mathbf{O}|\mathbf{R}^b) = \prod_{k \in \mathcal{O}} p(O_k | \mathbf{R}_k^b), \tag{6}$$

where $p(O_k = 1 | \mathbf{R}_k^b) = 1$ iff every reward component $R_c^b \in \mathbf{R}_t^b$ is not $\varnothing$. That is, the outcome k is true iff all corresponding reward components have been observed at the termination of the MCTS simulation.

Finally, by multiplying the left-hand side of Equations 1-6 we get the joint distribution over all random variables. The

binary random variables $\mathbf{R}^b$ are primarily used to simplify the calculation of the outcome probability distribution over $\mathbf{O}$, so for most calculation we marginalise $\mathbf{R}^b$ out, giving the joint we work with: $p(\mathbf{G}, \mathbf{S}, \mathbf{\Omega}, \mathbf{R}, \mathbf{O})$.

Note on Complexity
The size of the conditional probability distributions (CPDs) for $p(\omega_d|\mathbf{\Omega}_{1:d-1}, \mathbf{S})$ can, in the worst case, grow exponentially with the depth d according to $\mathcal{O}(|\hat{\Omega}|^d)$. However, there are two reasons why this is not a prohibitive issue. First, the search trees of IGP2 are very shallow, usually $d_{max} \leq 4$, and secondly since the MCTS tree is sparse, most values of the CPDs are zero. So instead of storing the full CPDs explicitly, we can associate each CPD to the state it is applicable in (given by $\mathbf{\Omega}_{1:d-1}$) and calculate the needed probabilities on-the-fly.

4.2 Extracting Causal Information

From the joint distribution we can infer various conditional distributions that allow us to draw causal judgements about counterfactual scenarios. Let us assume that MCTS selected the factual, optimal plan for the ego denoted with $\boldsymbol{\omega}_F = [\omega_1, \ldots, \omega_{d_{max}}]$. Further assume, that the passenger query describes a (possibly incomplete) set of counterfactual macro actions $\boldsymbol{\omega}_{CF} = [\omega_{j_1}, \ldots, \omega_{j_n}]$ corresponding to the random variables $\Omega_{j_1}, \ldots, \Omega_{j_n}$ indexed by the set $\mathcal{J} = \{j_1, \ldots, j_n\}$.

First, we calculate the outcome distribution of $\mathbf{O}$ given the counterfactual, that is the distribution:

$$p(\mathbf{O}|\boldsymbol{\omega}_{CF}). \tag{7}$$

This allows us to determine how the outcome of MCTS would have changed if ego had followed the counterfactual actions.

Second, we want to determine how the reward components differ from the factual to the counterfactual scenario. This would allow us to order the components by the amount that they were affected by the switch to $\boldsymbol{\omega}_{CF}$, and we can use this ordering to populate the *<EFFECTS>* variable in our explanation. Formally we can do this, by calculating:

$$\Delta\mathbf{R} = \mathbb{E}[\mathbf{R}|\boldsymbol{\omega}_F] - \mathbb{E}[\mathbf{R}|\boldsymbol{\omega}_{CF}]. \tag{8}$$

We can sort $\Delta\mathbf{R}$ in decreasing order by the absolute value of its elements to get the required ordering.

Finally, we would like to determine how much the trajectories of each individual non-ego participant affect the macro action selections of the ego. We can use this information to determine which traffic participants are most relevant to mention in our explanations and in what order. We can derive this ordering by comparing how the marginal distribution of macro actions $p(\mathbf{\Omega})$ changes when conditioned on different trajectories of non-egos. Since $p(\mathbf{\Omega})$ already encodes the optimal sequence of macro actions taking into account the trajectories of other participants, we are trying to find the conditional distribution that changes the marginal the least. Formally, for a vehicle i and for each of its possible goals $g \in G^i$ and trajectories $s \in S^i_{1:n}$ we calculate the Kullback-Leibler divergence between the marginal and the conditional of $\mathbf{\Omega}$:

$$\begin{aligned} D^i_{g,s} &= D_{KL}[p(\mathbf{\Omega})||p(\mathbf{\Omega}|G^i = g, S^i = s)] \\ &= \sum_{\boldsymbol{\omega} \in \mathbf{\Omega}} p(\boldsymbol{\omega}) \log_2 \left(\frac{p(\boldsymbol{\omega})}{p(\boldsymbol{\omega}|g, s)} \right). \end{aligned} \tag{9}$$

If $D^i_{g,s}$ is the same for all goals and trajectories, that implies that the actions of vehicle i does not affect the actions of the ego, so we will ignore vehicle i. Otherwise, we can sort all $D^i_{g,s}$ increasingly giving as an ordering on the importance of vehicles, goals, and trajectories. Note, that if vehicle i has only a single predicted goal and trajectory then $D^i_{g,s} = 0$. In this case we cannot use this measure to determine whether the vehicle i interacted with the ego or not. A more robust method to replace this measure would be to repeat the MCTS planning with each vehicle i removed from the simulations and looking at whether the actions of the ego have changed. This may however be computationally quite expensive to do.

4.3 Generating Natural Language Explanations

To generate intelligible explanations from the information derived in the previous section, we define a set of (recursive) generative rules given by a context-free grammar. We feed the extracted information to this grammar, which will generate a unique sentence. Since the raw generated sentences may be somewhat unnatural, we apply a post-processing step, where commonly occurring complex expressions are converted to simpler phrases (e.g. with higher time to goal $\rightarrow$ slower).

The complete set of generative rules is given in the appendix in Figure 6. To instantiate this grammar we pass the following information to it:

- s: Information about the counterfactual scenario containing three fields: the counterfactual macro actions $s.\boldsymbol{\omega}$, as well as the most likely outcome $s.o$ and its probability $s.p$ as given by Equation 7.
- $\mathbf{e}$: A list of effects on reward components of switching to the counterfactual. Each element $e \in \mathbf{e}$ contains two fields: the difference in reward $e.\delta$ as given by Equation 8, and the name of the reward component $e.r$ corresponding to the difference.
- $\mathbf{c}$: A list of causes that resulted in the effects we observed. Each cause $c \in \mathbf{c}$ has three fields: the non-ego traffic participant $c.i$ the cause is related to, the trajectory (and the macro actions that generated it) $c.\boldsymbol{\omega}$ the non-ego is taking as calculated using Equation 9, and the probability $c.p$ of the ego taking that trajectory.

For example, assume that we give to the CFG the following: $s = \{\boldsymbol{\omega} : [\textit{Continue}], o : \textit{done}, p : 0.75\}$. We also have $\mathbf{e} = [\{\delta : -5, r : \text{time}\}]$, and finally we got $\mathbf{c} = [\{i : 1, \boldsymbol{\omega} : [\textit{Change-right}], p : 0.6\}]$. Then the generated explanation before post-processing would be: *"If ego had continued ahead then it would have likely reached its goal with lower time to goal because vehicle 1 would have probably changed right."*.

5 Experiments

The criteria for human-centric AI set out in Section 1 necessitate our system to be transparent, causal, and intelligible. Our system is transparent by design, as neither IGP2 nor XAVI rely on any components that are black boxes or otherwise uninterpretable. We would then like to understand how well XAVI can capture the causal relationships when tested in realistic driving scenarios, and would also like to assess the intelligibility of our generated explanations.

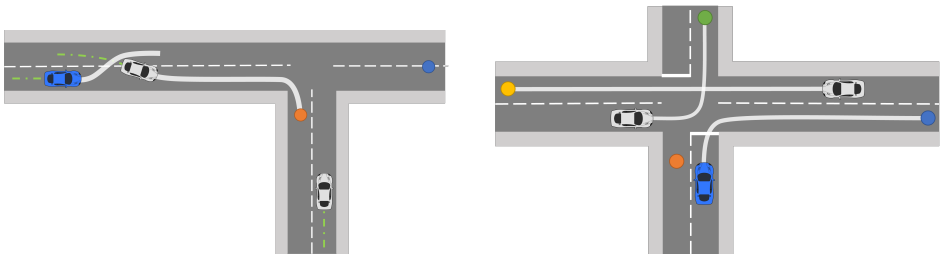

Figure 5: *(Left; S1)*. The ego vehicle (in blue) starts out in the right lane with its goal being to reach the end of the road it is currently on. The vehicle in front of the ego (vehicle 1) starts in the left lane. At one point vehicle 1 cuts in front of the ego by changing lanes right and then begins slowing down. This behaviour is only rational if vehicle 1 intends to turn right at the junction ahead. To avoid being slowed down, the ego changes to the left lane. The factual, optimal actions of the ego in this scenario are therefore $\boldsymbol{\omega}_F^1 = [\textit{Change-left}, \textit{Continue}]$. *(Right; S2)* The ego is trying to turn right and approaches the junction. Ego sees the vehicle on its left on a priority road (vehicle 1) slow down for a stop. Considering that there is an oncoming vehicle from the right coming at high speed (vehicle 2) the action of vehicle 1 is only rational if its goal is to turn left and it is stopping to give way. Ego can use the time while vehicle 1 is stopped to turn right earlier, instead of waiting until vehicle 1 passes. The factual actions of ego is: $\boldsymbol{\omega}_F^2 = [\textit{Exit-right}, \textit{Continue}]$.

$\boldsymbol{\omega}_{CF}^1$	**Generated Explanation**
Continue	***"If ego had gone straight then it would have..."***
	"likely reached the goal slower because vehicle 1 probably changes right then exits right."
	"likely collided with vehicle 1 because vehicle 1 probably changes right then exits right."
Exit-right	***"If ego had turned right then it would have..."***
	"not reached the goal."
	"collided with vehicle 1 because vehicle 1 probably changes right and exits right."

Table 1: *(Scenario S1)* Explanations with one cause and one effect. Our system can successfully determine the cause and effect of the lane change and the effect of exiting right. Note, that the system also captures the possibility of collisions when the ego and vehicle 1 start close to one another, as in this case ego cannot break quickly enough to avoid vehicle 1 cutting in front of it.

$\boldsymbol{\omega}_{CF}^2$	**Generated Explanation**
Exit-straight	***"If ego had gone straight then it would have..."***
	"not reached the goal."
	"not reached the goal because vehicle 1 likely turns left."
Exit-left	***"If ego had turned left then it would have..."***
	"not reached the goal."
	"not reached the goal because vehicle 1 likely turns left."

Table 2: *(Scenario S2)* Explanations with at most one cause and one effect. Both counterfactuals result in non-completion of the ego's goal, which is correctly captured as well as the rational action of vehicle 1 turning left. Note, the vehicle described in the causes affects the motion of the ego but is not directly responsible for the *counterfactual* outcome of the ego not reaching its goal.

$\boldsymbol{\omega}_{CF}^1$	**Generated Explanation**
Continue	***"If ego had gone straight then it would have..."***
	"likely reached the goal slower because vehicle 1 probably changes right then exits right."
	"likely reached the goal slower because vehicle 1 probably changes right then exits right and vehicle 2 exits right."
	"likely reached the goal slower and with more jerk because vehicle 1 probably changes right then exits right."
	"likely reached the goal slower and with more jerk and with less angular acceleration because vehicle 1 probably changes right then exits right."

Table 3: *(Scenario S1)* Varying number of causes and effects for the counterfactual $\boldsymbol{\omega}_{CF}^1 = [\textit{Continue}]$. Including more information in the explanation improves its informativity, however longer explanations become more difficult to comprehend.

For this, we perform a preliminary evaluation of the XAVI system in two simulated driving scenarios powered by the high-fidelity, open-source CARLA [Dosovitskiy *et al.*, 2017] simulation environment. The scenarios used here are scenarios S1 and S2 from the evaluation section of IGP2. Albrecht *et al.* give intuitive and rational explanations about the behaviour of the ego vehicle for each scenario presented in the IGP2 paper. In particular, details of scenarios S1 and S2 and the explanations of the observed behaviours are presented in Figure 5. We rely on these explanations as ground truth to see how well our generated explanations *match the causal attributions* of the ground truth explanations. We also vary the number of effects and causes passed to the explanation generation grammar to assess how the intelligibility of generated explanations changes with the complexity of the explanations.

To increase the diversity of our explanations, we perform ten simulations for each scenario where we randomly initialise the positions of the vehicles around their pre-defined starting points in a 10 meters longitudinal range. We also randomly initialise all vehicles' velocities in the range $[5, 10]$ m/s. In all scenarios, the counterfactual action specifies the value for the first macro action selection random variable Ω_1. So for example, the query *"Why did you change left instead of continuing straight?"* corresponds to $\Omega_1 = \boldsymbol{\omega}_{CF} = [Continue]$.

For scenario S1, we test two counterfactual actions: in one the ego continues straight behind vehicle 1 until it reaches its goal, so that $\boldsymbol{\omega}^1_{CF} = [$*Continue*$]$; in the other, the ego turns right at the junction, so $\boldsymbol{\omega}^1_{CF} = [$*Exit-right*$]$[2].

For scenario S2, we test the following two counterfactual actions: $\boldsymbol{\omega}^2_{CF} = [$*Exit-left*$]$, and $\boldsymbol{\omega}^2_{CF} = [$*Exit-straight*$]$. Note, that *"Exit-straight"* is used here to differentiate the action from regular *"Continue"* as the former macro action encodes giving way at a junction while the latter does not.

The generated explanations which include at most a single cause and a single effect are shown in Tables 1 and 2. Our system is able to correctly identify the effects of switching to the counterfactuals while also explaining which actions of the other vehicles (if any) caused those effects. In scenario S1, XAVI also revealed a possible collision outcome when the ego and vehicle 1 are spawned close to one another. Note, this outcome did not occur in the original IGP2 paper due to differences between random initialisations of vehicle positions.

Explanations with more than one cause or effect for scenario S1 are shown in Table 3. We do not have a similar table for scenario S2 as all relevant causal information can be captured by at most one cause. This is because the actions of vehicle 2 do not affect the actions of the ego directly in any way. For scenario S1, we can see that the shorter explanations can already capture the most crucial causal information of the ground truth, but more effects can be uncovered by XAVI. However, more detailed explanations increase the complexity of explanations which may make them harder to understand.

[2]The macro action *Exit-right* is a sequence of three manoeuvres: it encodes lane following until the junction, giving right-of-way, and turning. This means that ego will follow behind vehicle 1 also when executing *Exit-right*.

6 Discussion

Our results show that XAVI successfully captures some of the causal relationships as compared with the ground truth explanations from IGP2, while also being able to discover other, unexpected outcomes. The system is then able to generate intelligible explanations of varying complexity.

However, there are limitations to our work that need to be addressed in future work. One limitation of the method is its inability to explain the causes behind actions of traffic participants in terms of properties that are lower level than macro actions, e.g. features of raw trajectories. We would like to be able to justify our actions with causes that are finer in detail than the very high-level macro actions we currently have. High-level macro actions may encode different behaviours depending on how the other traffic participants are acting, therefore formulating causes in terms of macro actions may mask crucial differences between different runs of simulation. Indeed, the given causes for the ego's actions in Tables 1 and 3 were the same between counterfactuals, but it is clear that for different counterfactuals different causes relating to the particular motion of vehicles would be more relevant. For example, in Table 1 for the counterfactual *Continue* where ego reaches its goal slower, we should mention that vehicle 1 is *slowing down* for a turn instead of just mentioning that it is exiting right. On the other hand, for the same counterfactual where ego collides with vehicle 1 the more relevant cause for the collision is the actual fact that vehicle 1 unexpectedly changes right. To enable this lower-level extraction of causes would mean that we need to find a way to compare and filter features of trajectories based on their causal relationships to other variables, which is a difficult task given the complexity of driving environments.

Automatic explanation generation methods are by their very nature *post-hoc*, that is they work after our decisions were made. A common concern with any such post-hoc method is that they may not be sound, that is, faithful to the workings of the system they are explaining. Without a formal proof of soundness, we cannot fully claim that XAVI is totally faithful to IGP2. For example, XAVI does not represent each time step of the simulations explicitly or reason about how Q-values are updated. However, given the variables XAVI *does* reason about, we argue that our model is constructed to follow the steps of MCTS exactly without changing, removing, or adding extra information over a completed IGP2 planning run.

Another aspect to consider relates to the queries of passengers. While contrastive explanations work well for *why*-questions, there are many other types of questions users may ask (e.g. *"What?"*, *"How?"*), and we should support these lines of queries in the future. Parsing passengers' questions is also a non-trivial task. How could we know which macro actions a passenger is referring to in their question? What if those macro actions are not at all in our search tree? This last question also shows that we need a principled way to deal with missing data or cases where the system cannot give an explanation. What is more, giving explanations where algorithmic exceptions are present in a non-misleading and consistent way is especially important, as these explanations reveal shortcomings of our systems, that may reduce trust levels.

One aspect of human-centric AI, which we did not mention in this work is the benefit of being dialogue-oriented. Miller (2019) strongly argues that human-centric systems should be able to hold conversations with their human partners and allow opportunities for users to pose follow-up questions. This is beneficial for the users because they can converse with our system as long as their curiosity or information-need is not satisfied. Moreover, the system itself benefits from being able to hold conversations, as it can put the system on equal social status with humans, which is fundamental for developing trust [Large *et al.*, 2017]. We may also use follow-up questions to assess the passengers' understanding of our system, and deliver relevant explanations that are specifically designed to match the individual needs of each passenger.

Our evaluation of XAVI is preliminary, though the results are encouraging. However, we need to test our system on many more interesting scenarios so that we can generate a more varied set of explanations if we want to be certain that XAVI does indeed work properly and is useful for passengers. Besides the scenarios by Albrecht *et al.* (2021), we can base further evaluation on the scenarios presented by Wiegand *et al.* (2020) which were specifically collected to evaluate XAI in autonomous driving scenarios. In the future, it will also be important to run a user study on how the generated explanations affect trust and knowledge levels in humans, as our ultimate goal is to achieve trustworthy autonomous driving. Moreover, this will help to quantitatively assess the intelligibility of generated explanations and compare XAVI to other explanation generation systems.

7 Conclusion

In this paper, we present an explanation generation system called eXplainable Autonomous Vehicle Intelligence (XAVI). XAVI is designed to be fully transparent, causal, and intelligible thereby building towards a more human-centric explainability approach. It is based on mapping a Monte Carlo Tree Search-based motion planning and prediction system for autonomous vehicles to a Bayesian network that models causal relationships in the planning process. Preliminary evaluation of the system on a driving scenario shows that XAVI can accurately retrieve the causes behind and the effects of an autonomous vehicle's actions, and generate intelligible explanations based on causal information. We also discuss several possible next steps and issues that need to be addressed in future work, such as lower-level causes, conversation-enabled systems, the need for error-handling, and question parsing.

Acknowledgements

The authors would like to thank Cillian Brewitt and the anonymous reviewers for their helpful feedback. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

[Albrecht *et al.*, 2021] Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. Interpretable Goal-based Prediction and Planning for Autonomous Driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, March 2021.

[Atakishiyev *et al.*, 2021] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *arXiv:2112.11561 [cs]*, December 2021.

[Bastani *et al.*, 2017] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting Blackbox Models via Model Extraction. May 2017.

[Ben-Younes *et al.*, 2022] Hédi Ben-Younes, Éloi Zablocki, Patrick Pérez, and Matthieu Cord. Driving Behavior Explanation with Multi-level Fusion. *Pattern Recognition*, 123:108421, March 2022.

[Borgo *et al.*, 2018] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. Towards Providing Explanations for AI Planner Decisions. *arXiv:1810.06338 [cs]*, October 2018.

[Dazeley *et al.*, 2021] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525, October 2021.

[Dosovitskiy *et al.*, 2017] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16. PMLR, October 2017.

[Ehsan and Riedl, 2020] Upol Ehsan and Mark Riedl. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. February 2020.

[Hussain and Zeadally, 2019] Rasheed Hussain and Sherali Zeadally. Autonomous Cars: Research Results, Issues, and Future Challenges. *IEEE Communications Surveys Tutorials*, 21(2):1275–1313, 2019.

[Kim and Kelley-Baker, 2021] Woon Kim and Tara Kelley-Baker. Users' Trust in and Concerns about Automated Driving Systems. Technical report, AAA Foundation for Traffic Safety, April 2021.

[Kim *et al.*, 2018] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. *arXiv:1807.11546 [cs]*, July 2018.

[Korpan and Epstein, 2018] Raj Korpan and Susan L Epstein. Toward Natural Explanations for a Robot's Navigation Plans. In *Notes from the Explainable Robotic Systems Worshop*, page 3, Chicago, Illinois USA, March 2018.

[Lakkaraju *et al.*, 2017] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. July 2017.

[Large *et al.*, 2017] David R. Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics*, 63:53–61, September 2017.

[Lundberg and Lee, 2017] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*, November 2017.

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

[Montavon *et al.*, 2017] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, May 2017.

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

[Omeiza *et al.*, 2021b] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. Explanations in Autonomous Driving: A Survey. *arXiv:2103.05154 [cs]*, March 2021.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*, August 2016.

[Wang *et al.*, 2022] Hong Wang, Amir Khajepour, Dongpu Cao, and Teng Liu. Ethical Decision Making in Autonomous Vehicles: Challenges and Research Progress. *IEEE Intelligent Transportation Systems Magazine*, 14(1):6–17, January 2022.

[Wiegand *et al.*, 2020] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. "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, pages 1–11, New York, NY, USA, October 2020. Association for Computing Machinery.

$$
\begin{array}{ll}
S[s, \mathbf{e}, \mathbf{c}] & \rightarrow \text{ if } ACTION[\varepsilon, s.\boldsymbol{\omega}, \varnothing] \text{ then } EFFECTS[s.o, s.p, \mathbf{e}] \text{ because } CAUSES[\mathbf{c}] \text{ .} \\
ACTION[i, \boldsymbol{\omega}, p] & \rightarrow str(i)\ ADV[p]\ MACROS[\boldsymbol{\omega}] \\
MACROS[\boldsymbol{\omega}] & \rightarrow str(\boldsymbol{\omega})_{|\boldsymbol{\omega}|=1} \mid MACROS[\boldsymbol{\omega}_1] \text{ then } MACROS[\boldsymbol{\omega}_{2:}] \\
EFFECTS[o, p, \mathbf{e}] & \rightarrow \text{it would have } OUT[o, p]\ \ COMPS[\mathbf{e}] \\
COMPS[\mathbf{e}] & \rightarrow \epsilon_{\mathbf{e}=\varnothing} \ \mid\ COMP_{|\mathbf{e}|=1}[\mathbf{e}] \ \mid\ COMPS[\mathbf{e}_1] \text{ and } COMPS[\mathbf{e}_{2:}] \\
COMP[e] & \rightarrow \text{with } REL[e.\delta]\ \ str(e.r) \\
CAUSES[\mathbf{c}] & \rightarrow \epsilon_{\mathbf{e}=\varnothing} \ \mid\ CAUSE_{|\mathbf{c}|=1}[\mathbf{c}] \mid CAUSES[\mathbf{c}_1] \text{ and } CAUSES[\mathbf{c}_{2:}] \\
CAUSE[\mathbf{c}] & \rightarrow ACTION[c.i, c.\boldsymbol{\omega}, c.p] \\
OUT[o, p] & \rightarrow ADV[p]\ str(o) \\
REL[\delta] & \rightarrow \text{lower}_{\delta<0} \mid \text{higher}_{\delta>0} \mid \text{equal}_{\delta=0} \\
ADV[p] & \rightarrow \text{never}_{p=0} \mid \text{unlikely}_{0<p\leq 0.33} \mid \text{probably}_{0.33<p\leq 0.67} \mid \\
 & \quad \text{likely}_{0.67<p<1.0} \mid \text{certainly}_{p=1.0} \mid \epsilon_{p=\varnothing}
\end{array}
$$

Figure 6: The explanation generation grammar rules. The function $str(.)$ returns a pre-defined textual representation of its argument. Subscripts denote conditions for the rule to be applicable. Note, ϵ denotes the empty string while ε the ego vehicle.