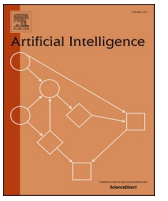
Artificial Intelligence 343 (2025) 104323

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



The value of real-time automated explanations in stochastic planning

Claudia V. Goldman [a,†,*], Ronit Bustin [b,†], Wenyuan Qi [c], Zhengyu Xing [c], Rachel McPhearson-White [d], Sally Rogers [d]

[a] Hebrew University Business School, Hebrew University of Jerusalem, Jerusalem 9190501, Israel
[b] Toga Networks, a Huawei Company, Tel Aviv, Israel
[c] General Motors (China) Investment Co., Ltd., Shanghai, China
[d] General Motors, Warren Technical Center, Warren, MI, USA

ARTICLE INFO

Keywords:
Explainable AI
Decision-Making
Human-Computer interaction

ABSTRACT

Recently, we are witnessing an increase in computation power and memory, leading to strong AI algorithms becoming applicable in areas affecting our daily lives. We focus on AI planning solutions for complex, real-life decision-making problems under uncertainty, such as autonomous driving. Human trust in such AI-based systems is essential for their acceptance and market penetration. Moreover, users need to establish appropriate levels of trust to benefit the most from these systems. Previous studies have motivated this work, showing that users can benefit from receiving (handcrafted) information about the reasoning of a stochastic AI planner, for example, controlling automated driving maneuvers. Our solution to automating these hand-crafted notifications with explainable AI algorithms, XAI, includes studying: (1) what explanations can be generated from an AI planning system, applied to a real-world problem, in real-time? What is that content that can be processed from a planner's reasoning that can help users understand and trust the system controlling a behavior they are experiencing? (2) when can this information be displayed? and (3) how shall we display this information to an end user? The value of these computed XAI notifications has been assessed through an online user study with 800 participants, experiencing simulated automated driving scenarios. Our results show that real time XAI notifications decrease significantly subjective misunderstanding of participants compared to those that received only a dynamic HMI display. Also, our XAI solution significantly increases the level of understanding of participants with prior ADAS experience and of participants that lack such experience but have non-negative prior trust to ADAS features. The level of trust significantly increases when XAI was provided to a more restricted set of the participants, including those over 60 years old, with prior ADAS experience and non-negative prior trust attitude to automated features.

1. Introduction

Automation has benefited human society by releasing human operators from performing repetitive and highly demanding physical

* Corresponding author.
E-mail address: claudia.goldman@mail.huji.ac.il (C.V. Goldman).
† Claudia V. Goldman and Ronit Bustin were affiliated with General Motors, Technical Center Israel, when this research was performed.

https://doi.org/10.1016/j.artint.2025.104323
Received 9 March 2024; Received in revised form 21 February 2025; Accepted 4 March 2025
Available online 8 March 2025

tasks. Artificial Intelligence (AI) [69] is taking these processes to newer levels of development, innovation, and production [33], entering our daily lives both at the physical as well as at the digital layers (e.g., autonomous vehicles [52] and mobile phones [22]).

As long as we make progress on the technological side, we should be aware of the human side, representing those end-users interacting with these systems (i.e., end-users of AI systems are not machine operators anymore, but they could be anyone of any age experiencing almost any context). From an AI perspective, considering these users means that the AI algorithms need to consider the trust levels that will be established by their users. It is already well known in the Human Factors domain, that trust in automation is essential for users to accept and use these automated systems [73,31]. Furthermore, to establish this trust, the automated actions need to be communicated [46,66], although determining the appropriate amount of information and protocols needed are not trivial questions. As a result of not establishing an appropriate level of trust, interacting with these automated systems can lead to misuse, disuse, or over-use [60], i.e., misinterpreting the system, disengaging from using it or over-trusting it even when it errs.

We are interested in stochastic AI planners applied to real-life problems (e.g., automated driving). Such AI planning systems compute policies of behavior for stochastic, sequential decision-making problems under uncertainty [69,79]. These policies of behavior result from processing a sequential temporal process, over a large set of states, while facing uncertainty for example about other entities affecting this process. This approach to planning raises three concerns of particular interest when developing these systems to interact with humans: (1) Due to the uncertainty in the environment and in the process, the resulting behavior might not be deterministic, whose logic is easy to track back, (2) the AI planning choices of actions for execution are dependent on long term effects computed from projecting into the future what might happen in real-time and (3) usually in real-life problems, the state space is huge, sometimes incomprehensibly large. Therefore, AI planners solving for a policy of behavior might opt for approximations necessary to predict next states, predict priors to actions and their associated values. Expecting users to establish an appropriate level of trust in such AI planners is not trivial [34,78,85]. These three aspects raise the concern that users that interact with such AI-controlled systems might not understand **why** the system is behaving in certain ways [26]. For example, when a user is riding an autonomous vehicle, this vehicle might take certain maneuvers that might not be expected by that user. They might have expected the vehicle to behave differently (e.g., wait less time before merging, making a lane change instead of staying on the same lane while following a slow vehicle). An automated explanation, notifying the user about why the vehicle behaved in that way, or explaining what the vehicle is about to do next might help the user overcome the surprise incurred by an unexpected behavior. Automated notifications might increase users' trust levels, higher understanding and better alignment with the autonomous AI controlled vehicle.

We recognize a potential gap in communication between an AI-controlled system and its users. In the area of automated driving, these gaps might lead to discomfort and **misunderstanding** of automated maneuvers leading to low levels of trust and acceptance (see for example [28] on users' attitude towards autonomous vehicles revealing low levels of prior trust and fear of self-driving technology [57]. In this domain, where five levels of driving automation are distinguished [71], gaps between the mental model of the vehicle's user and the reasoning model of the automated vehicle can lead to higher frequencies of manual take overs (in Level 2 and Level 3 driving) or to low levels of acceptance of advanced driving technologies (such as in Level 4 or Level 5 driving)).

In this paper, we are interested in bridging this gap by computing real-time notifications automatically that explain the behaviors performed by the AI-learned policy to end-users, experiencing them. It has been shown that agents explaining their AI based decisions contribute to the establishment of trust between the system and its users [13,29]. This poses both a computational and a human-usability related questions: (1) Given a human experiencing an automated behavior of an AI controlled system, and the inner reasoning process of that AI stochastic planner, what information should be extracted and processed by the explainable AI (XAI) algorithm (in a computational manner, what information can be computed from such reasoning process)? and (2) what additional processing this information might need to go through to become human-understandable and be of value to the system users? Answering the first question is not straightforward due to the complexity of the reasoning process and since users might have different preferences for different types of explanation and different contexts [10,41,54,56,82]. Moreover, these explanations need to be timed in real-time, according to the system behavior and the users' needs. This timing contrasts with explanations that aim at summarizing the policy of behavior of a planning system, which can be provided offline or post interaction, i.e., when the summary of the policy is available [32]. The second question is challenging as well, since developing the explainability capability is not enough to show the value of these explanations to their users.

Our contributions to the area of explainable-AI are summarized as follows:

1) Applicability to a real-life problem (scalability and applicability): We chose to implement a stochastic AI planner with the AlphaZero algorithm [6,75], that can cope with the complexity and huge states' space of a real-life domain such as autonomous driving. Our implementation handles an extremely large space state, representing 8 features related to the Ego vehicle (e.g., laneID, lane-change-status, lateral velocity, longitudinal velocity, latitude position, longitude-position, heading, distance to destination and 7 features for each one of the closest 20 vehicles around the Ego (e.g., lane-change-status, in-radius, relative velocities (latitudinal and longitudinal), relative position (latitudinal and longitudinal), and relative heading). Most of these features have continuous values, making this space in practice incomprehensibly large. Our framework copes with this complexity. Other, more classical approaches to stochastic planning would not be applicable to such complex problems where a transition function between states is not explicitly provided, and full coverage of the state space cannot be guaranteed during a training stage. Our XAI solution is applicable to planning problems, that can be solved by common model-based and model-free solutions.
2) Human-understandable Explanations and Types of Explanations: Our solution comprises a set of XAI algorithms for computing three different types of real-time explanations, considering (1) what the planning system is reasoning about for the probable futures, (2) what alternative actions were considered by the planner and (3) why the planner chose a certain maneuver for execution according to its reward function. The solution also includes three timing schedules and an assessment of the presentation modality.

We also defined the notion of dominance of actions and dominance of value components based on a divergence measure among probability distributions to limit the use of thresholds. During the development process of this comprehensive XAI solution (content, timing and modality of presentation), our focus remains on the algorithmic aspect as well as the computation of human-understandable explanations out of the AlphaZero policy of behavior.

3) Large-scale user study assessment of explanations value: The XAI solution presented in this paper has been assessed by a very large user study (800 participants) which includes users of very different age groups (30–89 years old) and backgrounds (technical and non-technical). We measured these users' levels of subjective understanding and trust through simulated automated driving scenes with continuous and large state spaces. We analyzed the data collected, and identified when the XAI notifications, shown together with a dynamic HMI display, increase trust and subjective understanding of the participants, compared to those cases where participants received only a dynamic HMI display. XAI notifications were found to significantly decrease misunderstanding. Also, participants with prior ADAS experience reached higher levels of subjective understanding than those provided with HMI only displays. Even when participants lacked ADAS experience, XAI increased their level of subjective understanding when they had non-negative prior trust attitude to automated features. Trust was a more challenging goal. Nevertheless, we observed that participants over 60 years old with prior ADAS experience and non-negative prior trust attitude did reach higher levels of trust compared to those that received only and HMI solution.

The XAI solution, presented in this work, serves to make complex AI solutions accessible to human users by augmenting such systems with transparency. The resulting transparent planning systems will help their users understand better the automated behaviors. This is an important stage in the process of accepting advanced technologies, applied to real-life challenges.

The paper is structured as follows: Section 2 overviews the work done in the area of explainable AI. Section 3, then, discusses what we mean by an automated explanation computed from the reasoning process of a stochastic AI planner. The three proposed XAI algorithms are then described in Section 4, followed by an empirical evaluation in Section 5. We conclude and discuss further directions for future research in Section 6.

2. Explainable AI (XAI) for planning - related work

The problem of understanding a human or machine choice of a single decision, made optimally under uncertainty can be quite challenging. For example, a rational agent might solve this problem by choosing the action with the most likely desired output, after having considered a possible risk associated with this choice; humans on the other hand [36] might take irrational actions due to their human character. If we take the role of observers of such a single action decision making, it could be pretty difficult to explain to ourselves why that particular action was the optimal choice made. If we observe the rational agent, we might need a model of this agent's reasoning. In the case of observing a human, we might need to consider the irrationality of their behavior (or alternatively a psychological model that might formalize such behavioral choices). In this paper, we are interested in *sequential* decision-making processes, that develop over time, such as autonomous driving. The complexity of understanding a sequence of actions as observers, increases, compared to that of understanding the choice of a single optimal action. In our case, humans interact with an automated system, an AI planner, that outputs automated driving maneuvers during a simulated ride. Since humans and intelligent agents behave according to different models of reasoning, the question whether humans, experiencing these simulated autonomous rides, will understand the system behaviors and will trust these interactions is not trivial to answer.

A common device used in vehicles to communicate information with the drivers is the Human Machine Interface (HMI) that can provide static or dynamic information [43], representing the road, the host vehicle, its direction of driving and some of the traffic around it. For an interesting review covering the historical evolution of challenges and solutions in automotive human factors see [2] and [11]. In this paper, we aim at evaluating the information that can be computed by XAI algorithms, explaining an AI stochastic planner's actions, while users experience simulated autonomous driving scenarios. To the best of our knowledge such additional information, processed from an AI reasoning methodology, has not been considered so far in real, existing, HMI displays.

Users' expectations from such automated behaviors might result from their own usage of the system, at times in a manual manner. For example, human drivers might develop expectations from automated vehicles based on their own driving style and their driving experience in different traffic situations. Craig et al. [15] found that drivers, in general, might expect autonomous vehicles to drive less aggressively than themselves, unless they trust a-priori this technology. Then they would expect the vehicles to drive as humans. It is not obvious that any kind of explanation about such complicated reasoning processes can help users to understand and trust better the AI-system. Goldman and Bustin [26] reported on two very large user studies (over around 2500 participants) that tested the need for specific explanations compared to very general explanations and a single graphical human machine interface. General and specific explanations were handcrafted according to the researchers' experience on automated driving maneuvers and were presented in several simulated driving animated scenes. *Specific* information was found to reduce both misunderstanding and the feeling of discomfort levels. The different types of information provided, included a dynamic HMI interface and three hand-crafted types (1) Plan-Next: what is going to happen next? (2) Risk: information referring to the risk or reward that affected the driving maneuver and (3) combination of Plan-Next and Risk information. Additional research motivating the study of different types of information can be found in [5,54]. The need for optimizing the information provided to drivers, while interacting with an AI controlled vehicle, was also studied in [10,41]. In Beggiato et al. [5] six HMI experts were interviewed for what information users might need in a driving experience. This information was **manually** created based on these recommendations and evaluated in a user study with 20 participants. This information included HMI information and information about the next maneuver (What is the vehicle going to do next). Koo et al. [41] evaluated verbalized messages that explained "how" and "why" of a simulated semi-autonomous vehicle. Messages

providing "why" information were preferred by drivers over messages providing "how" information. Chang et al. [10] found that more information does not necessarily mean higher trust or higher understanding. Their study showed four graphical displays to users showing what the vehicle senses. No automated explanation about the reasoning process controlling the behavior was computed. Miller [54] covers contrastive explanations as a type of information that humans prefer in humans-to-humans' interactions.

A broad variety of solutions for offline and online planners have been studied for problems of different sizes and complexities (e.g., [69]). However, the more classical approaches are not applicable to real-life problems such as automated driving, which covers an incomprehensibly large state space, needs to cope with high levels of uncertainty and the value function covers multiple parameters related to different domains such as progress, driver comfort and safety. Similar to the work done by [30], we implemented the AlphaZero planning algorithm [6,75]. The reasoning of this planner is the focus of our explainable AI work. In this paper, we assume a set of discrete actions. This is a stepping stone approach into this complex problem to provide insights that can help in designing and shaping the follow up research on XAI for humans interacting with automated behaviors in stochastic environments. Our XAI solution for the case when these actions are continuous will be reported separately.

AlphaZero is a model-based reinforcement-learning (RL) algorithm. The basic assumption is the availability of a generator that can predict the progress of the environment. During training, the target is like other RL algorithms. We want to obtain a system that for any given state, it will provide the optimal action (or rank the possible actions) to execute. AlphaZero trains a neural network (NN) that provides this prediction; however, since the state space is incomprehensibly large the planner cannot cover all states during training, and it uses its learned prediction only as a prior. During real-time operation, AlphaZero uses a time-budget (computation-budget) to expand the possible futures in the form of a Monte Carlo Tree Search (MCTS) in which it takes the outcome of the NN as a prior in its calculations. The purpose is to fine-tune its decision corresponding to the specific state encountered. The algorithms presented in the following sections refer to explanations computed in real-time for the real-time evaluations of the MCTS. Appendix B presents the technical details of this implementation.

Research on explainable AI algorithms [17,24] covers two sub domains. The first is dedicated to explaining black boxes [63]. For example, explaining learning models, trained for perception tasks [25,39], smart manufacturing [27], sales forecasting [49], healthcare [61], finance [7] and the sciences [65]. For a recent comprehensive review see Saranya and Subhashini [72] where among others, solutions included the analysis of influence of certain input features on the outputs (through perturbations) or processing additional information from the network or developing surrogate models in the form of simpler learning models. Outputs might include data structures such as decision trees [50] or graphical representations of the explanations (see for example SHAP values [3,67, 74]). Black boxes that learn about the vehicle perception might help in explaining what the system is currently observing and can be used in the final explanation provided to the user.

We are interested in the second sub-domain of XAI, which explains planning systems (XAIP). Planning solutions can be architected as white, gray, black boxes, or combinations of these. Our focus is on providing real-time explanations for the AlphaZero and MCTS algorithms.

The literature on XAIP has covered explanations for AI planning solutions spanning deterministic planners, through model-based planners to Reinforcement learning planners [69]. Several directions of work have opted to compromise optimality for interpretability by creating explainable plans [9,18,21]. Previous work on XAI explaining AI planners is not applicable in terms of scalability, suitability to very large and complex real-life domains, and understandability of their explanations to any kind of user. For example, Kulkarni et al. [44] and Sreedharan et al. [77] work on model reconciliation assumed planning is deterministic. Decision trees studied by Lim et al. [47] can be simpler to explain when they are small, but these do not necessarily scale with larger trees as in MCTS representing real life scenarios. Khan et al. [37] studied explanations for explicit model-based planning problems, based on the frequency of reaching a state with extreme reward values. Russell et al. [68] compute global and local explanations by creating a decision tree and applying LIME [63] to express the important features affecting the predicted reward values. Their approach was demonstrated in a very simple domain, whereas the more complex version of the same domain was found difficult to be explained with the same approach. Partially Observable MDPs (POMDPs) were explained in Wang et al. [81] by all its formal components on a non-real world setup. In particular, the authors compute explanations based on the relative likelihoods of events while Miller [54] claims that humans are not good at understanding probabilities or likelihood in numerical form.

Work on explaining RL based planners has been reviewed very recently in Milani et al. [53]. According to the taxonomy they created, our XAI work here corresponds to the category of Learning Process and Markov Decision Process (LPM), whereas the other two categories are Feature Importance (FI) in terms of the input state features that affect the choice of action and Policy-Level (PL) (e.g., summaries of polices [20,32] that can be presented once the policy is available). Our explanations are computed based on the planner's reasoning process and its components in real-time and they are also provided in real-time to the end-users. Van der Waa et al. [80] considered explanations for a simple domain by presenting the user with information about possible consequences of acting in a state. Closer to our direction of work is the XAI work by Anderson et al. [4]. They assumed that the reward function is composed of separate components which are processed to produce explanations. These are shown in a graphical bar representation. For reducing the load of the person receiving the explanation, the minimal sufficient explanation is computed by including those positive and negative reward factors that were critical to the action choice. Similarly, Juozapaitis et al. [35] explained the RL planner's choice of action by analyzing the reward components that were either critical positive or negative in the choice of that action compared to others. Explaining the reward components is in line with providing the "why" (handcrafted) information as was done in Koo et al. [41]. One of our XAI algorithms, VAL, referring to the reward components is based on dominant factors that do not require a threshold to be set as is the case in [4,35]. Lin et al. [48] studied contrastive explanations [54] through additional features' functions to add semantic understandable features to the explanations. The planner's choices for actions are explained by contrasting the semantic properties predicted for each possible action. Similar to [35], if there are too many options, minimal sufficient explanations are computed. The value of these

explanations was estimated in a computational manner in three strategic computer games, without assessing this value to the end users. In this paper, we implemented an XAI algorithm for computing contrastive explanations based on the planner's reasoning process. A third type of explanations, similar to the one studied here as Plan-next, focuses on the possible futures the planner is reasoning about (e.g., see [5,64,84]. However, in Beggiato et al. [5], the explanations were handcrafted. In Yau et al. [84] explanations are not human understandable; they were applied in small scale environments and no user study was performed to evaluate them. Finally, in Rietz et al. [64], explanations combine decomposing reward functions with hierarchical reinforcement learning, these were evaluated in a simple 2D navigation environment and no user study was performed. We also defined the notion of dominance of actions and dominance of value components based on a divergence measure among probability distributions to limit the use of thresholds.

Causal explanations were studied to explain the causal effects that state factors had on the actions chosen by a stochastic planner [19,51,58]. For example, Elizalde et al. [19] referred to the importance of state factors that caused higher changes in the value of an action. Nashed et al. [58], developed a theoretical sound framework that computed causal explanations based on Structural Causal Models (SCM). Explanations were computed by applying causal inference to this structure on different factors such as state features, reward values and transition functions. Explanations could include subsets of these factors (while SHAP [74] does not consider such dependencies among these factors). In its current form, this solution is not yet applicable to real life domains with incomprehensibly large state spaces as we handled in the automated driving domain. SCMs were also studied in [51] for finding causal relations between variables, values, and actions in an RL process. However, as shown in Madumal et al. [51], the trust levels did not improve by using this type of explanations. Deep RL XAI solutions [1,14] were evaluated on non-real-world domains (i.e., not at scale with real life domains represented in 2D) and they were tested by experts only (e.g., students or users with technical background in the area of the system). Those experts are not representative of possible laymen, for example the end-users of an autonomous vehicle. Thus, a user study evaluating whether there is a beneficial value of these explanations is missing in these works.

Feature attribution (e.g., [12]) computes explanations by associating an output of the system to the input features which were influential in *predicting* that choice. In our reported work, we have not studied such type of explanations. Our system is not predicting an action that might be executed but, instead, it is choosing the actual action that will be executed. A separate study related to simulated automated driving has been reported where state features are associated with an explanation for the planner' behavior [58].

In comparison to this existing XAI literature, our work is unique along these three dimensions: scalability and applicability, large scale user study testing and human accessibility of the computed information. We focus on the computation and evaluation of three types of explanations answering why a simulated autonomous driving maneuver was performed in simulated realistic scenarios. We provide a computational solution for plan-next, contrastive and value explanations and evaluate their effect on a large set of varied users. These types were shown to be beneficial in a simulated driving environment, but they were either hand crafted or evaluated in small pools of participants, usually with technical backgrounds. The next section presents how information can be processed from a tree data structure holding the reasoning of a stochastic planner.

3. Explaining the reasoning process of a stochastic planner

In general, the automated explanation process depends on the information available to the AI planner while reasoning on its decision-making solution. That reasoning depends further on what is known to the planner (e.g., the states and available actions in the context of the problem it is trying to solve) and it also depends on a stochastic model that represents the dynamics of the planner domain (the transitions between states when an action is executed by the agent applying the planner's policy). Depending on the implementation of the planning algorithm, this stochastic model can be either explicit (i.e., the probabilistic transitions between states

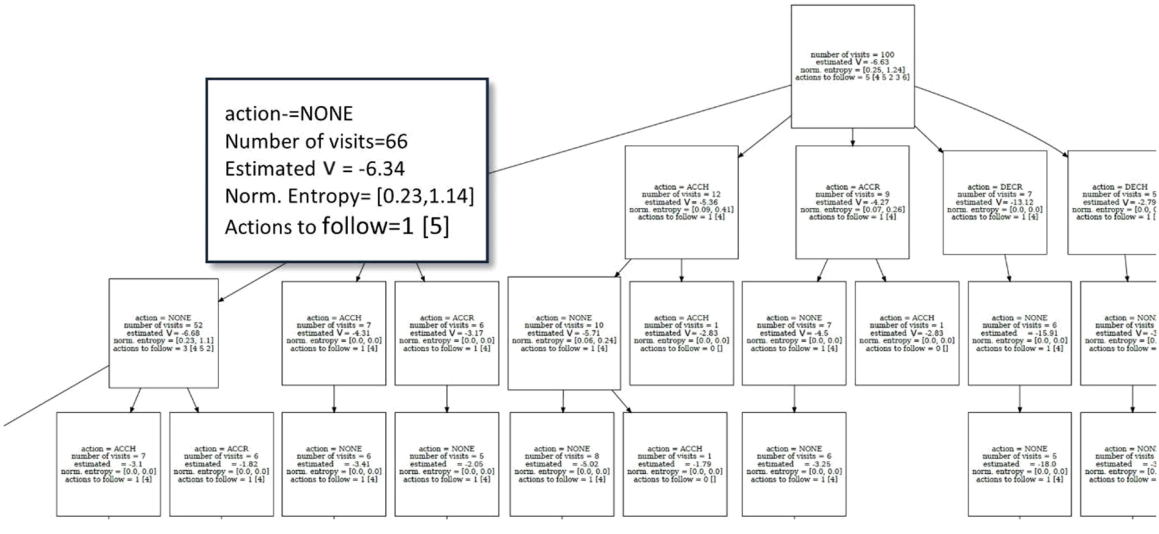


Fig. 1. An example of a Monte Carlo Tree Search data structure, showing the node's information.

are known by the planner because these transitions were given – as would be the case for a Markov Decision Process) or implicit (i.e., the transitions between states are not known to the planner; these could be estimated or predicted from data that the planner has collected by exploring the state space (e.g., Q-learning [79]) or by learning a predictive model from data provided or by running a physical model of the environment of the problem – as would be the case for a driving and traffic model based on rules of physics). As we discussed in Section 2, the most studied approaches to planning are not applicable for many real-life problems. In such complex cases, models are not provided explicitly (usually there are many unknowns ahead of time), and models cannot even be predicted in full since the state space is incomprehensibly large (in the sense that the transition model cannot be estimated for *all* states in the huge state space). In this paper we focus on the AlphaZero planning algorithm [6,30,75] and specifically the MCTSs constructed in real-time evaluation, as the algorithm fine-tunes its prior learned insights and fits them to the encountered state. The XAI approach explains the information accumulated in a tree data structure during the planning process (see Fig. 1). These MCTS trees are, in fact, our best "windows" to the policy of behavior learned by the planning algorithm. As mentioned in Section 2, this planning algorithm is considered model based because it assumes a generation model (G) that allows the construction of the tree (G can be an explicit or an implicit model in real time). In our implementation, the model resulted from a physics-based driving model. Given that we are in a specific state and perform a specific action, G gives us a good estimate of the next state (the imaginary state); the model G *generates* the next state. However, in contrast to classical model-based planning algorithms, where the transition function is provided prior to the solution of the problem, this is not a full transition function. The model can be complex and probabilistic. Furthermore, if we perform the same action again it could give a different outcome. Therefore, the iterative construction of the tree is our closest estimate of this transition function.

Fig. 1 shows a tree, resulting from an MCTS, created during the application of the AlphaZero planning algorithm. At the root, we have the current state, encountered by the automated system (e.g., the current scene of a driving scenario). The children of this root state are all those states reachable from the root by performing any one of the actions available to the planner (e.g., in our case we consider discrete driving maneuvers: accelerate-harsh, accelerate-regular, decelerate-harsh, decelerate-regular, left-lane-change, right-lane-change, none). At each node we have the following information (as stored by the MCTS algorithm):

- The action being evaluated at this node,
- The number of visits the algorithm has reached this state already, that is, the number of times the algorithm chose (during the selection step, see Appendix B) to go down the tree through this node,
- The accumulated value is an approximation of the expected accumulated value of that state (up to a factor of number of visits),
- We have added information computed about the uncertainty of the subtrees below every node (their entropy) and the dominance of actions at each node. We will refer to the dominance computation when we describe the relevant XAI algorithms. Our work on the entropy related measures is reported separately in [8].

The XAI challenge is then to process this information such that the resulting explanations will help non-technical end-users understand and trust the AI system making decisions, controlling its behaviors. The rest of the paper will focus on our solution to this XAI challenge, assuming a tree data structure is available. Nevertheless, we want to bring to the attention of the reader that our complete XAI solution can be adjusted to other planning algorithms, commonly studied for other domains under the umbrella of decision-making under uncertainty.

In classical model-based planning problems, we can describe the policy of action as a **state machine**. A tree structure can be formed directly from the information provided by the model (i.e., states, actions, reward, and transitions) and the state machine. The XAI approach presented in this paper can be applied directly on this tree structure: for any current state, the XAI algorithm processes the information the planner holds according to its model in this tree, and can output the same types of explanations similar to the way that the XAI algorithm is described for the tree created for the AlphaZero algorithm.

In model-free planning problems, applying reinforcement learning methods, on the other hand, we can estimate the transition probabilities throughout the exploration process and maintain an expected probability function for any transition between two states. We can build, then, a tree for **expected** trajectories resulting from chosen actions and predicted transition probabilities. Then, the types of information that can be computed by our XAI solution could be computed from these trees of projections into probable futures, shading light of what the model-free planner is expecting to happen or not to happen according to the information it has at the current state and the probabilities computed so far.

4. Explaining AI Planning Decisions Automatically

The need for an automated system (e.g., a vehicle) to behave transparently so as to increase trust of an end user has been motivated by the review of literature in Section 2. Making an automated system behavior transparent to a user does not mean that it needs to provide as much information as possible [10,41]. That is, there is a need to distinguish between types of information and evaluate different amounts and types of information provided to end users. In this section, we present our solution for computing automated notifications in real time for a complex stochastic planner: for any state (context) representing a simulated driving scene, our XAI algorithms compute contextually related, real time, automated notifications. Following the works reviewed in Section 2, we implemented three types of automated explanations: plan-next (similar to handcrafted explanations studied in [26] and in [5]), contrastive (motivated by the work of Miller [54] in human to human explanations) and value based explanations about why a behavior was chosen ("why" hand-crafted explanations were preferred by users over "how" explanations in [41] and studied also by [4]).

1) Plan-Next (PN): what is the most likely behavior that would be executed by the planner up to some horizon? What is the sequence of actions and states that compose that behavior reflecting the reasoning process of the stochastic planner?
2) Contrastive (CON): what would have happened if the planner would have chosen at the root state, the second-best option instead of the optimal action? Or what other alternatives did the planner consider? We compute this explanation once the choice of action a^* has already been made. We explain what alternative action could have been chosen instead, but the planner ended not choosing it eventually. So, we contrast two possible choices of action (one that was made and the other that was considered by the planner during its reasoning process, but, eventually, it was not chosen).
3) Value components (VAL): what were the value components that most impacted the choice of the action at the root state, given all the possible futures considered by the planner? Note that this type of explanation is different from feature attribution explanation which focuses on states' features as possible explanans. These value components represent the formal reasoning process of a stochastic planning system. Users experiencing a simulated driving ride are not aware about the reward function implemented in the code of the planner behind the vehicle's driving behavior. Explaining the most impactful value components in a human understandable way is aimed at making this automated behavior accessible to the end users experiencing it.

4.1. *Explaining what the vehicle is about to do next (Plan-Next)*

Results from a user study performed in 2021 on simulated automated driving scenarios (see [26]) showed that one third of the population evaluated (i.e., approximately 400 participants), identified the lack of information about what the vehicle is about to do next as a source of driving discomfort. We assume that the planning reasoning information is provided as a tree data structure (as shown in Fig. 1), created during the process of running an MCTS planner. Each MCTS can be viewed as a model of a stochastic state transition (see Appendix B), with the transition probabilities defined by the number of visits (normalized).

The general XAI algorithm for computing explanations of type Plan-Next is shown in Fig. 2, assuming T^s is a tree data structure rooted at current state s.

The algorithm has four main steps: choose a potential future (trajectory or subtree), shorten it by a criterion that determines its length, abstract it to the most significant actions and parse it semantically. The XAI algorithm needs first to determine the potential future that the explanation will be referring to, according to what the planner is predicting. The developer will code a criterion for choosing such path p. We implemented this algorithm focusing on a single trajectory, the one with highest probability to occur. Alternatively, other versions could explain the choice of the action at the root of the tree, by analyzing the information held by the subtree, developed from the node reached by following the action chosen at the root of the tree. That information could shed light, for example, on the average expected speed or the uncertainty perceived in any possible expected future that could develop from this node. In our case, the criterion to choose a path instructed the algorithm to extract the highest probability path, by going down the tree and performing inference, starting from the edge led by the action to be performed at the current state, at the root node.

In an MDP version of the planner, this path can be chosen by the actions that maximize the transition probabilities. Other criteria can be applied to choose this trajectory or subtree, for example, the trajectory attaining at least some minimal value, or a trajectory that expresses a behavioral pattern which is like a human pattern of behavior (e.g., a pattern of acceleration, or turning in a curve) or a trajectory up to a state where the transition probability is kept above some certainty threshold.

Secondly, we might decide to cut this path p to obtain a shorter path that might be easier to process and to understand. The decision to cut the length of the path can be determined by a threshold of maximal length or it can be determined by a measure of how certain the planner is at different depths of the tree. For example, we might decide to keep adding nodes to the final path while the level of uncertainty of the planner at that node is kept below some upper bound. Whenever the uncertainty of the planning system is higher, we consider that potential sub-tree too uncertain and we do not include those nodes into the path that will be parsed as an explanation. In this paper, our results refer to the Plan-Next algorithm implementation, considering all nodes along the path that were developed in the tree (in the implementation of the MCTS planning system, maximal number of nodes (the budget) was set to 100).

Thirdly, in a discrete planner, such as the one implemented here, actions are taken from a finite set of actions. Therefore, at the end of this step, we obtain a sequence of actions. We apply a criterion to this sequence that allows us to make an abstraction of the predictive future behavior of the vehicle. This can be helpful when composing the actual notification to be displayed to the user (e.g., imagine a possible sequence returned by the algorithm up to this point including 10 equal atomic actions such as "accelerate"; a user-friendly way of presenting the information in this sequence is to consider the semantics of this sequence, for example by abstracting them: "Accelerating for some time" instead of [accelerate, …, accelerate]. Other examples might include names of maneuvers that can

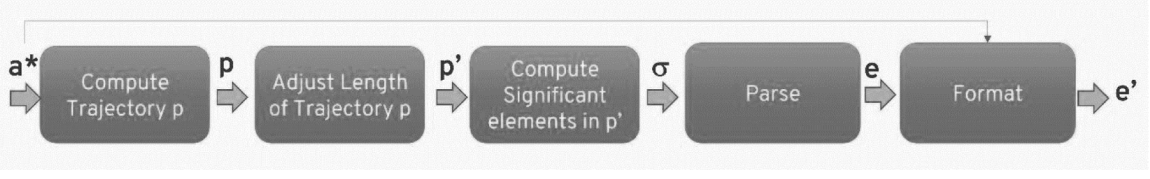


Fig. 2. XAI Algorithm #1: How to compute an explanation e' of type Plan-Next, for an action a* chosen for execution by AlphaZero Stochastic Planner.

be interpreted from a sequence of actions. See the table below for more examples). When the explanation is aimed at engineers for the purpose of debugging, we might still want to have all the details of the planner's predicted future. We considered the complete sequence of actions as a string to which a regular expression is applied (see examples in Table 1). The semantic translation of the sequence of actions determined by the algorithm into a human understandable expression is then a process of pattern matching. Alternatively, we could have chosen only a subset of the actions in the final path as those associated with a value higher than some threshold, or as being dominant actions (see Appendix A) and apply a regular expression to the string corresponding to this sequence.

Finally, the sequence (i.e., the regular expression) is parsed into a semantic expression as shown in Table 1. The table is shown for the actions' set implemented in our work, comprising {NONE, LEFT, RIGHT, ACC0, ACC1, DEC0, DEC1}, where ACC0 and DEC0 refer to regular accelerations or decelerations of $\pm$ 10kmh in 5 secs and ACC1/DEC1 refer to harsher accelerations or decelerations of $\pm$ 40kmh in 10 secs. LEFT and RIGHT refer to left or right lane change actions and NONE is the action taken when no changes are made to the speed or the lane. These seven actions were indexed by 0–6.

The algorithm ends by outputting a human understandable explanation. We apply a template to compose a grammatically correct and human-understandable sentence. In the case of a Plan-Next type of explanation, we apply: **[Semantic-Current-chosen- action- a^*] then [semantic-maneuver] .** See Fig. 3 for an example of an explanation of type Plan-Next output on the vehicle display in real-time for the tree information as shown in Fig. 4 for the action None for a simulated merge driving scenario.

4.2. Explaining the planner's alternative choices (Contrastive)

In this section, we refer to why-contrastive questions [55]. We are interested in explaining to the user why an action, different from the one chosen by the planner, was not chosen for execution. That would be the *alternative explananda.* Such contrastive explanations are common when humans explain to each other [54]. Drivers might wonder if the vehicle had considered actions that a person might have chosen, even when the vehicle performs a different action. Just knowing that the vehicle considered them in the same current state, can be beneficial for a user that otherwise might think that such alternative actions were not considered at all. A relevant example could be when a user might be expecting the vehicle to slow down when approaching an intersection because they are experiencing a speed that is subjectively higher than what they might have wished. In such situations, a driver can benefit from receiving information explaining what options the planner considered and why it did not choose them (for example, slowing down was considered as an option but speeding up or keeping the same speed was found to be more beneficial). The second XAI algorithm implemented is shown in Fig. 5.

The main difference between this algorithm and the one presented for Plan-Next explanations is in computing first the set of foils. These are the alternatives considered by the planner, but eventually not chosen; foils are potentially all actions different from action a^*. In the current implementation, we compute a single foil, but an extended version of this algorithm (depending on the context on which it is applied) could include computing a list of foils sorted out by some criteria measure. In some cases, it might be interesting to compare the action chosen to the least probable action or the riskiest action not chosen. Here, the foil action is the "second-best", that is the action associated with the highest number of visits that is different from a^*.

Once we determine the foil, we continue in a similar manner as we did in the Plan-Next XAI algorithm. We focus on the trajectory leading from the foil to some horizon, collecting those actions predicted with highest probability. Then, we parse this sequence of actions into a semantic expression. Finally, we apply a grammatical template, corresponding to the contrastive meaning of the explanation: **[Semantic-Current-chosen- action- a^*] because it will work better than [semantic-maneuver] .** See Fig. 6 for an example computed by the tree in Fig. 7 for a NONE action.

4.3. Explaining why an action was chosen (Value Components)

The third XAI algorithm explains the semantics of the reward function, indicating why the vehicle has chosen to drive in a certain manner. In a regular RL setting, the objective of the intelligent agent is to learn a policy of actions that maximizes the expected utility

Table 1
A Parser Example for Regular Expressions Applied to Strings Composed of Discrete Actions.

Regular expression of the actions in chosen path	Return semantic maneuver
ACC0 [ACC0]*	Accelerating
DEC0 [DEC0]*	Slowing down
LEFT [LEFT]*	Looking for a faster lane
RIGHT [RIGHT]*	Keep driving on the right side
NONE [NONE]*	Maintaining speed in this lane
[ACC0]* ACC1 [ACC1]*	Accelerating quickly
[DEC0]* DEC1 [DEC1]*	Slowing down quickly
[ACC|NONE|DEC]* LEFT [ACC|NONE] or [ACC|NONE|DEC]* RIGHT [ACC|NONE]*	Passing
[ACC|NONE|DEC]* LEFT [LEFT]*	Looking for a faster lane
[ACC|NONE|DEC]* RIGHT [RIGHT]*	Keep driving on the right side
[ACC0]* [NONE|DEC]* ACC0 [NONE|ACC0]*	Accelerating
[NONE|DEC0|DEC1]*[ACC0]* ACC1 [ACC0|ACC1|NONE]*	Accelerating quickly
[ACC0|ACC1|NONE]* DEC0 [NONE|DEC0]*	Slowing down
[ACC0|ACC1|NONE]* [DEC0]* DEC1 [DEC1|DEC0|NONE]*	Slowing down quickly

Fig. 3. An Example of a Plan-next Explanation for a Merge Scenario.

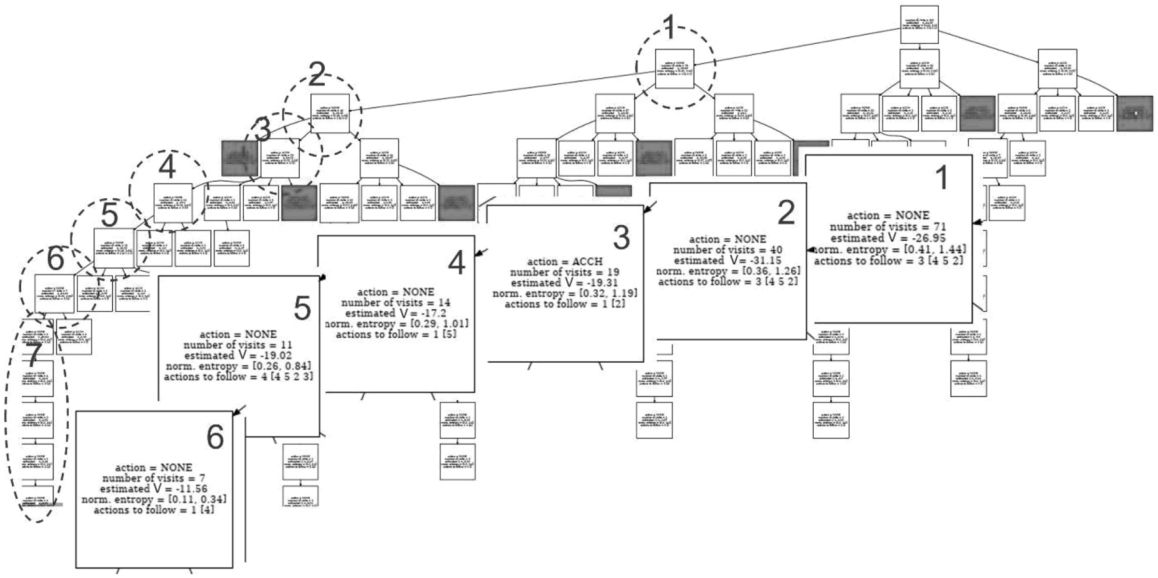


Fig. 4. The most likely trajectory in the MCTS data structure developed at time 20 seconds of the algorithm when the action chosen for execution at the root state is "None". The numbered nodes are those processed by the Plan-Next XAI algorithm to produce the explanation as shown in Fig. 3.

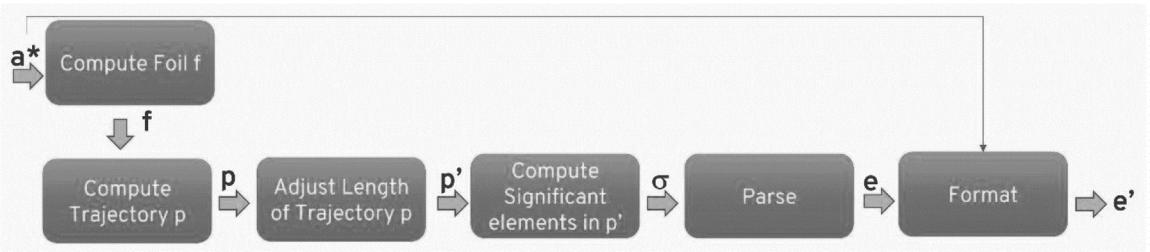


Fig. 5. XAI Algorithm #2: How to compute an explanation e' of type Contrastive, for an action a* chosen for execution by AlphaZero Stochastic Planner.

Fig. 6. An Example of a Contrastive Explanation for a Merge Scenario.

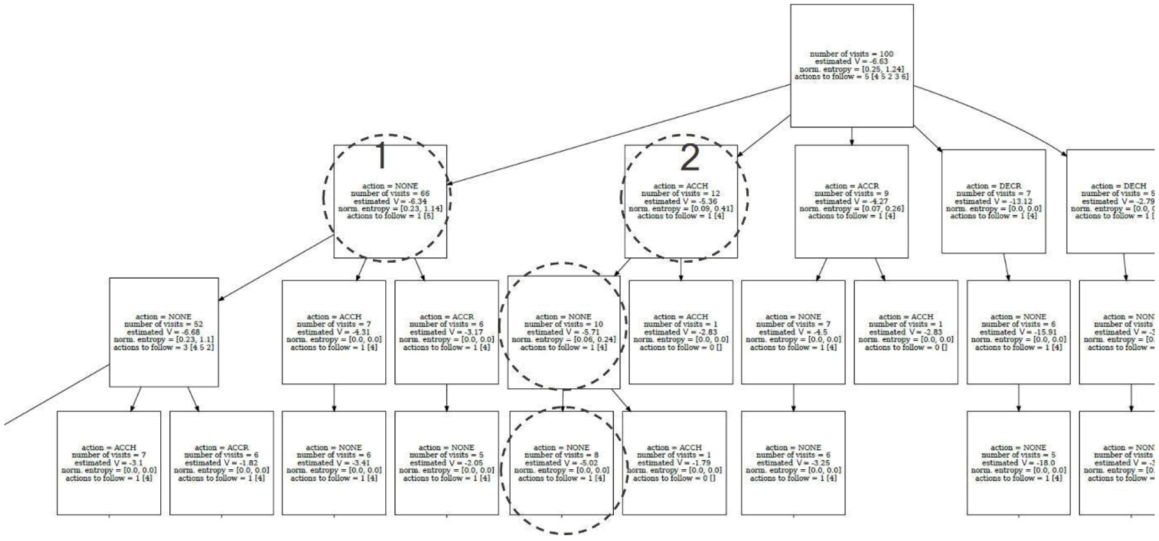


Fig. 7. The nodes circled are those considered by the Contrastive XAI algorithm in the MCTS data structure, developed at time 9 seconds of the algorithm when the action chosen for execution at the root state is "None". The explanation output for this case is shown in Fig. 6.

over time, that is, it maximizes the expected discounted accumulation of a reward function r. The immediate reward obtained at time t is given by $r(s_t)$ and γ represents the discount factor.

$$R_t = \sum_{j=0}^{\infty} \gamma^j r(s_{t+j})$$

For our study, we assume that the function *r()* is a linear weighted combination of several functions r_i *()* – the components of the immediate reward. Let's assume there are k such components in the immediate reward function. These components can help identify relevant factors affecting the behavior the AI planner is controlling (for example, safety features such as related to the road topography, user comfort or trust features in an automated driving domain). Consequently, the planner can reward or penalize these factors according to values of features of the road or the drive affecting the comfort of the driver:

$$r(s_t) = \sum_{i=1}^{k} w_i \, r_i(s_t)$$

Combining these two equations, we write the discounted accumulated reward as a linear combination of its components:

$$R_t = \sum_{j=0}^{\infty} \gamma^j r(s_{t+j}) = \sum_{j=0}^{\infty} \gamma^j \sum_{i=1}^{k} w_i \; r_i(s_{t+j}) = \sum_{i=1}^{k} \sum_{j=0}^{\infty} \gamma^j w_i \; r_i(s_{t+j}) = \sum_{i=1}^{k} R_{t,i},$$

where $R_{t,i} = \sum_{j=0}^{\infty} \gamma^j w_i r_i(s_{t+j})$.

The expectation of these future accumulations of discounted rewards are approximated in the standard manner as done in MCTS [6] by performing selection, accumulation, and accounting for the number of accumulations (the number of visits in that node in our implementation of the MCTS algorithm, for example). Each node holds two values, the number of times it has been visited and a value that corresponds to the total reward of all playouts that passed through this state (so that is an approximation of the node's game-theoretic value). It can be noted in Fig. 8 that we adjusted the AlphaZero planner to store at each state in the tree its accumulated value per component (and adjusted the NN module to return the estimated value per component as well).

$$V_{i,t} = E\left[\sum_{j=0}^{\infty} \gamma^j w_i r_i(s_{t+j})\right]$$

For example, in a simulated automated driving domain, we can have the following reward components:

- r_1 can refer to the curvature of the road segment current point relative to current radius,
- r_2 can refer to the headway between the host vehicle and the vehicle immediately in front,
- r_3 can refer to the current speed of the host vehicle
- r_4 can refer to the current jerk of the host vehicle (change in acceleration over t-(t-1))

From a semantic perspective, we can associate each of these components to an abstracted semantic category, for example r1 and r2 are related to the safety of the ride, r3 refers to the progress made by the host vehicle along its ride and r4 might refer to the comfort of the drive. Our implementation (evaluated in the study reported later in the paper) for r1-r3 is standard as can be seen in (e.g., [30] implementation for r2 and r3).

The third XAI algorithm (see Fig. 8) provides an explanation based on information available to the planner only at its first level (that of the root's children nodes) unlike the Plan-Next and Contrastive XAI algorithms which process trajectories. The calculation of the value components, though, has been done throughout the tree (at each node). The combination of additional information from the potential future with the values of the components' assessment is an interesting direction of future work (i.e., combining what actions will probably be chosen or what semantic behavior can be interpreted by a sequence of expected actions together with the expected value can provide information that could be better expressed in semantic terms and might be beneficial for the user to understand the behavior of the automated system).

This XAI algorithm first creates a probability vector by shifting the value components to hold only positive values, normalized to be in [0,1]. The most significant components in this vector are returned after computing the Jensen-Shannon divergence measure between these value components and the four relevant uniform distributions (see Appendix A, where this function is provided in detail and generalized to any length of probability vector). The resulting list of the most dominant components is parsed to return a semantic and human understandable explanation (see Table 2). An example of an automated notification, explaining its value-components is shown in Fig. 9 for a simulated curve following driving maneuver, processed from the information collected in the tree data structure as shown in Fig. 10.

4.4. When shall we present an explanation?

In a temporal process, such as autonomous driving, timing can affect the understanding and trust levels of a user experiencing the ride. Providing predictive information about a driving maneuver before the maneuver takes place might be perceived as too early and

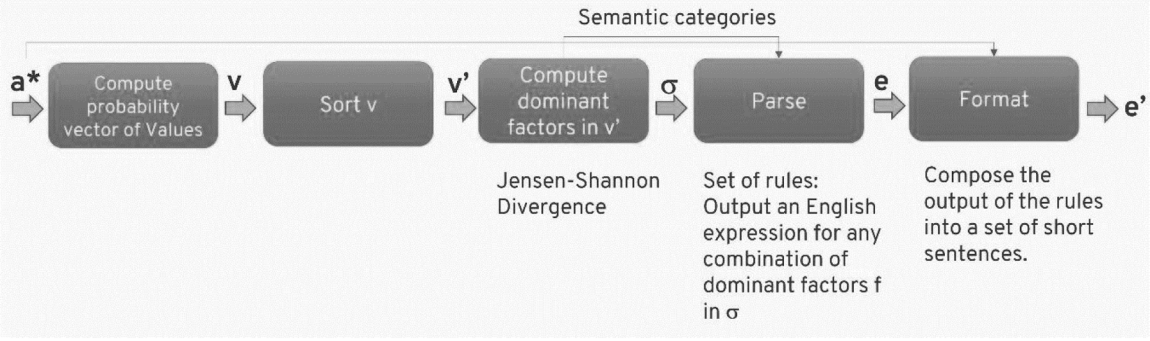


Fig. 8. XAI Algorithm #3: How to compute an explanation e' of type VAL (Value Components), for an action a* chosen for execution by AlphaZero Stochastic Planner.

Table 2
A Parser Example for Expressing the Most Dominant Value Components in English.

Value components reasoning [if statement refers to the dominance of those factors]	Explanation String
if curvature and not headway and not vel and not jerk:	return "Curve ahead.\n Driving safely."
if not curvature and headway and not vel and not jerk:	return "Vehicle ahead.\n Driving safely."
if not curvature and not headway and not vel and jerk:	return "Adjusting acceleration to maintain comfort."
if not curvature and not headway and vel and not jerk:	return "Speed set to its optimum."
if curvature and headway and not vel and not jerk:	return "Curve and vehicle ahead. \n Driving safely."
if curvature and not headway and not vel and jerk:	return "Curve ahead.\n Adjusting acceleration to drive safely."
if curvature and not headway and vel and not jerk:	return "Speed set to its optimum.\n Curve ahead.\n Driving safely."
if not curvature and headway and not vel and jerk:	return "Vehicle ahead. \n Adjusting acceleration to maintain comfort."
if not curvature and headway and vel and not jerk:	return "Vehicle ahead. \n Adjusting acceleration to maintain comfort."
if curvature and headway and not vel and jerk:	return "Curve, headway and jerk are affecting driving.\n Taking care of a safe and comfortable driver."
if curvature and headway and vel and not jerk:	return "Curve and vehicle ahead.\n Speed set to drive safely."
if curvature and headway and vel and jerk:	return "Curve and vehicle ahead. \n Adjusting acceleration to maintain comfort and drive safely."
If no curvature, and not headway, and not vel and not jerk	return null
If curvature and not headway and vel and jerk	return "Curve ahead. \n Adjusting acceleration to maintain comfort and drive safely."
If not curvature and headway and vel and jerk	return "Vehicle ahead. \n Adjusting acceleration to maintain comfort and drive safely."
If not curvature and not headway and vel and jerk	return "Adjusting acceleration to maintain comfort and drive safely."

Fig. 9. An Example of a Value Explanation for a Curve Following Scenario.

as a result, the explanation will not help that user understand the maneuver. Providing such predictive information during the maneuver or after it might be perceived as too late. Another concern related to timing refers to the frequency of information provision. Providing information too frequently might overload a user while providing it less frequently might miss opportunities that a user could have benefited from (e.g., they might have decided to keep engaged in an automated driving experience because of getting that additional information). The question of finding the optimal time to provide an explanation is a research question on its own that remains for future work (i.e., for a given content of explanation, what would be the optimal timing to provide it to the user or alternatively can we compute both optimal content and timing at the same time?). In this paper, we focused on studying the value of the explanations in terms of their content (also described here as the types of the explanations). In our solution, we have implemented three options for scheduling the time to provide the automated explanations as described below.

Constant-Duration: a duration period is set through a configuration file. Without additional conditions, an explanation will be provided to the user at constant gaps of duration time.

Configurable Rules: a configuration file allows designers to determine when an explanation is required. Rules can be designed to determine the contextual conditions for an explanation to be provided. These rules can be further personalized according to the driver profile (settings directly managed by the driver or by learning some model representation about their preferences). For example:

1. If the vehicle makes a left or right lane change, it will always display an explanation,
2. If the vehicle brakes or accelerates harshly, it will always display an explanation,
3. If the vehicle is driving through a certain area in a map, it will always display an explanation,
4. If the vehicle's driving profile is highly different from some regular behavior, then it will display an explanation.

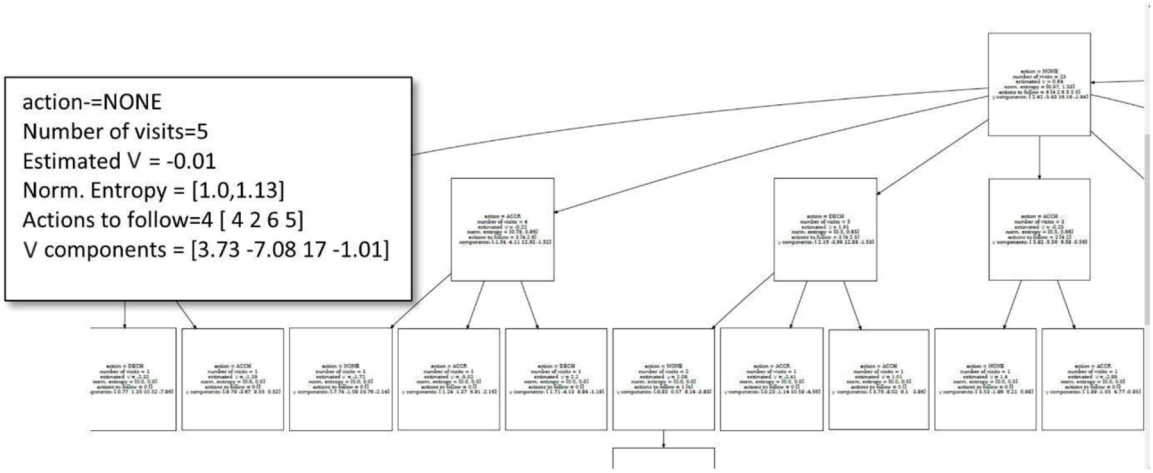


Fig. 10. The values shown in the zoom-in box are those stored at the node being chosen for execution at time 20 seconds of the MCTS algorithm (action "NONE" is chosen for execution). The VAL XAI algorithm will process these value components to output an explanation on the most impactful components as shown in Fig. 9.

Dominant actions: An automated explanation scheduler can be developed according to the reasoning process of the planning system. This scheduling solution can be determined according to the uncertainty that the planner is facing when deciding to execute an action (reflected in the current tree). For example, we might want to instruct the system not to provide explanations when it is too uncertain about the choices it is making, but to provide explanations about its behavior when it is more certain about them. The XAI algorithm will provide an explanation only when there is a single dominant action, or two dominant actions (i.e., the minimal Jensen-Shannon divergence was found for the uniform distributions over either one or two elements). In the first case of a single dominant action, an explanation projecting the most likely future will be provided, while in the second case of two dominant actions, the XAI algorithm will explain the contrast between these two dominant actions. This measure is computed at the state currently evaluated by the planner. Other implementations relevant to this explanation scheduler include providing explanations when more than two actions are found dominant. This remains for further study, requiring the definition of similarity between actions to provide concise explanations.

4.5. How shall we present explanations?

Once the content of an explanation has been computed, an important question would be how to present that information to the user. In our work, we assume that the vehicle is fully autonomous, therefore speech solutions might interfere with the users' activities. Speech outputs will interrupt other users' activities such as phone calls or listening to music. We opted for a textual option to present the explanations (in an autonomous driving setup, the user is not required to control the vehicle as in manual driving or as in partially automated driving). For a recent review on text design for in-car displays see [62].

We have run a usability study to compare between two possible representations of the information provided to users. In one representation, the participants viewed static figures of a graphical display with information describing the maneuver. In the second representation, the participants viewed the same graphical representation with additional text (the explanation expressed directly as an English sentence). In both cases the same HMI display was shown. The textual explanations were handcrafted according to the information that could have been computed as shown in the XAI algorithms reported here. This study was run on static images of displays and no simulated animated movie was created for these. Reporting on this usability study is out of scope. However, for the sake of completeness of this article, we report on the main conclusions: textual notifications are necessary in addition to the graphical display to help users build trust in the system. Users also expressed their desire to be able to configure this request for information. For example, with time, users might change their needs for the amount and the frequency of information provided.

5. Explainable AI algorithms - empirical testing

This section reports on the empirical evaluations conducted to assess the value of XAI notifications, computed with the algorithms presented in this paper. Usually, automated vehicles provide situation awareness information to their drivers through a dynamic HMI display. Our study explores whether XAI notifications can add value to these displays in the sense of increasing trust and understanding of their users.

5.1. User Study: The value of XAI real-time notifications

This paper is focused on the planner's behaviors for fully autonomous vehicles, assuming the scenarios experienced are all safe. In

that context, trusting such technology is essential for this technology to succeed and penetrate the market (see [38] and references cited within). Thus, it is important to evaluate whether the new type of automated notifications proposed here would help drivers increase their subjective understanding of the autonomous vehicle and could then help them in increasing their trust in this new technology leading to higher usage. That means that it is not a question of operating the vehicle where misuse or disuse can result from over-trusting or under-trusting the automation. Therefore, measuring objective understanding by observing how a user operates a system is not relevant to this study. We, thus, report on self-reported measures including subjective understanding and trust. This explains why our hypothesis are phrased with an increase in these measures instead of establishing a calibrated level of these measures. This study also enables us to learn what factors might have additional effects on the value of these automated notifications such as gender, age, and prior Advanced Driver Assistance System (ADAS) experience of the drivers.

To achieve these goals, we raise the following hypotheses:

- Hypothesis #1: Age, gender, and ADAS experience affect the levels of subjective understanding and trust of users, when they are provided with XAI explanations in addition to a dynamic HMI display.
- Hypothesis #2: Automatic XAI explanations (i.e., Plan-next, Contrastive and Val XAI algorithms presented in this paper) presented together with a dynamic HMI display help the users reach levels of trust higher than those attained when the dynamic HMI display was provided alone.
- Hypothesis #3: Automatic XAI explanations (i.e., Plan-next, Contrastive and Val XAI algorithms presented in this paper) presented together with a dynamic HMI display help the users reach levels of subjective understanding higher than those attained when the dynamic HMI display was provided alone.

This study also tested the case of partially automated driving where a transfer of control between the vehicle and the user might be needed to control the driving operation. The study evaluated whether the additional automated notifications could affect drivers to avoid taking manual control when such shift of control might happen due to the driver's need (e.g., a driver might get anxious or worried if they do not understand what the vehicle is about to do or why it is reacting in a certain way). In these cases, automatic explanations can clarify the behavior of the vehicle, thus reducing the misunderstanding level of the user. For this question, our study is an initial attempt to evaluate XAI notifications in such scenarios requiring humans and automated systems to collaborate. A fourth hypothesis was raised:

- Hypothesis #4: Automatic XAI explanations (i.e., Plan-next or Contrastive XAI algorithms presented in this paper) reduce the number of manual takeovers in scenarios when users might have preferred to take over when provided a dynamic HMI display only.

5.1.1. Methodology - Apparatus

We have developed the AlphaZero and MCTS planning system on a simulator platform and implemented the three XAI algorithms described in Section 4 with the different timing schedules. A fixed schedule option has been assigned to each explanation type; therefore, these timing schedules did not create any experimental condition. We chose three driving scenes on a highway map to perform this evaluation: ramp merge into a highway scene, straight driving, and curve following. We trained the planner to drive autonomously in each one of these scenes. Then, we ran this planner for each one of the scenes and output animated simulated movies from every execution of the planner. The length of each movie was around one minute. We chose three movies (one for each scene) for running the online study. All driving scenarios were safe and the autonomous vehicle followed a reasonable driving profile learned by the planning system.

Fig. 11. A Snapshot taken from the ramp merge driving animated movie, showing a contrastive explanation next to the dynamic HMI display.

Fig. 11, Fig. 12 and Fig. 13 show 3 snapshots from each one of the movies respectively, displaying (A) driving on a ramp merge into a highway (B) straight driving in a traffic congested highway including a lane change maneuver and (C) following a curve on a highway. In the merge scenario, the timing of the explanations was determined by the dominant actions timing algorithm. The content of the explanations was, thus, determined to be either Plan-Next or Contrastive accordingly. In the straight driving and curve following scenarios, the explanations were of type Value-components, and they were provided with the timing schedule of constant duration of two seconds. See Section 4 for the details of each algorithm.

5.1.2. Experimental design and procedure

This 2 × 3 between-subjects study was performed on-line during the month of September 2023. All participants experienced all conditions (3 driving scenarios movies). Participants were assigned to two groups based on the information that was provided to them in addition to the dynamic HMI display: XAI notifications or none.

5.2. Participants

The study included the recruiting of 800 participants (400 females and 400 males) between 30 years old and 89 years old (M=57.1 years old, SD=12.8). A total of 1054 possible respondents entered the survey, 185 were screened out and an additional 69 were removed for data quality concerns, to get the 800 total responses desired. These participants were also balanced across two age groups. The first group included 400 participants of age 30–59 years old (referred to as the *young group: M*=46.5 years old, SD=8.32), and the second group included the other 400 participants of age 60–89 years old (referred to as the *old group: M*=67.8 years old, SD=5.57). The distinction between these two groups is due to the results presented in a prior study [26], showing that older populations (above 60 years old) react differently than the younger populations when information is provided during simulated driving scenarios. Our aim was to deepen our understanding of this group, so we recruited an equal number of users in both age groups.

A vendor, AYTM (see https://aytm.com), was utilized to field the questionnaire via their proprietary online consumer panel, Paid Viewpoint. Participants are either recruited or joined of their own accord by filling out a demographic profile. They are informed up front that they will be invited to participate in various surveys that they may qualify for, based on their profile, and compensated for each survey they click into. The amount of compensation varies based on length of survey, quality of their responses, and how many questions they answered. By clicking into a survey based on its description each participant is agreeing to share their opinions and has the option to stop and leave the survey at any time.

This was a blind study, where participants were not informed of who the questionnaire was for. The inclusion criteria for each participant included:

1. Must be 30 years old or above.
2. Is not employed in a sensitive industry that may bias their responses: Advertising or Public Relations, Promotions, Automotive, or Marketing/Market Research.
3. Must be a frequent driver, driving at least once a week.

The results reported here were collected from participants who answered all the questions and attested that they could see the videos properly. A test question was included of a video of an otter followed by a question asking what animal was shown in the video. This served to confirm participants were able to watch the assessment videos and that they are paying attention and not a bot. Each participant's set of responses were reviewed by at least two analysts for any anomalies, such as straight lining, speeding, verbatim responses, or response that did not make logical sense in comparison to their other responses, that may indicate they were not

Fig. 12. A snapshot taken from the straight driving animated movie showing a VAL explanation next to the dynamic HMI display.

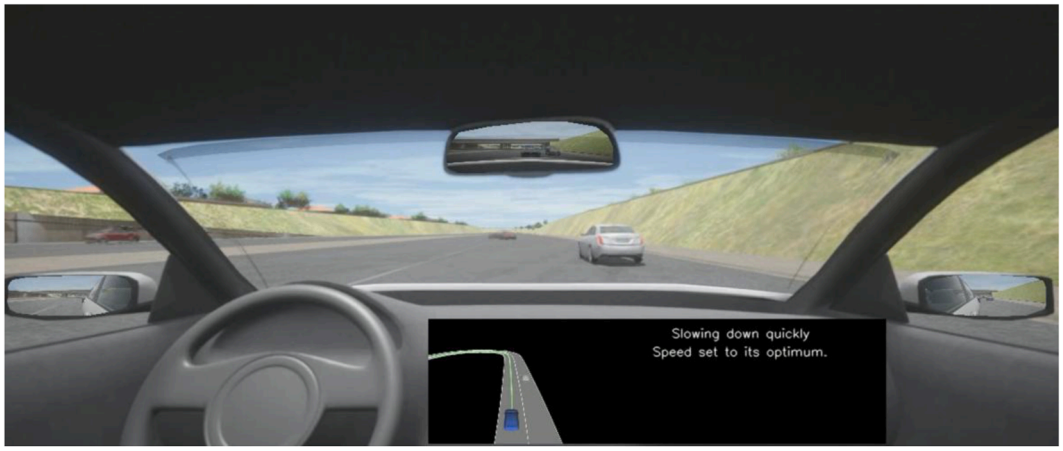


Fig. 13. A snapshot taken from the curve following animated movie showing a VAL explanation next to the dynamic HMI display.

answering honestly or thoughtfully. Participants who had at least two or more marks against them were removed from the sample set. By the end of the study, the users were asked whether the videos presented were clear and lacked any buffering or downloading issues. Data was collected only from those with no issues.

5.3. Study Procedure

Each participant started the study by filling up demographics data, including frequency of driving and experience with ADAS features (for example Adaptive Cruise Control). All participants were asked about their prior trust towards automation in the context of advanced driving features. For the specific questions and possible answers see Appendix C.

Then, all participants went through a practice assessment where one animated movie of an autonomous driving maneuver was shown, followed by the questions they would be required to answer in the actual study. Appendix C comprises the introductory text and the set of questions presented to each participant for each one of the three animated movies created for each one of the driving scenarios as discussed.

The movies, showing the autonomous driving behavior of a host vehicle, were presented in a random order. Each participant was shown the three movies according to the group they were assigned to, either "HMI only" or "HMI + text". HMI-only refers to those participants that watched the movie showing the drive through the vehicle windshield, and a dynamic HMI display was shown on the vehicle's dashboard. HMI+text refers to those participants that watched the same movie as those in the HMI-only group, but also could see textual automated notifications shown to the right of the dynamic HMI display on the vehicle's dashboard where text refers to the automated explanation computed in real-time. Half of the population was assigned the first group and the other half was assigned to the second group.

After watching all three movies and answering all related questions (i.e., trust and understanding levels), each one of the 800 participants was asked about the helpfulness of the overall information provided to trust the vehicle. Then, finalizing the study, we showed all the participants an additional movie, different from the previous three. This movie was presented to all participants with the dynamic HMI display only, with no notifications. The driving scene showed the host vehicle trying to merge on a highway. We asked all participants whether they would have considered taking manual control of the vehicle during the automated driving maneuver shown in that movie. For those that did not answer negatively, we presented possible automated notifications computed with the Plan-Next and the Contrastive XAI algorithms. We asked the participants whether any one of these types of information would have changed their minds so that receiving that information might prevent them from taking manual control of the vehicle if that would have been an option (See Question 17 in Appendix C).

The design of this study allowed for data collection about users' subjective understanding and trust in a simulated autonomous driving vehicle. This data could be compared when users were provided automated notifications in addition to the dynamic HMI display and when the HMI was the only information provided on the display.

5.4. Metrics

Trust is widely studied in the social sciences as well as between humans and automation. Schilke et al. [59] extended the usually studied dichotomy between two types of trust, where on one extreme trust can be conceived as generalized trust (i.e., general trust in some group independent of the identity of the trustee) and on the other, trust can be characterized as particularized trust (i.e., trust in a specific situation). In the particularized trust case, the trustor and the trustee have mutual interests in succeeding in their interaction and there is no reason to believe that the trustee will behave opportunistically. Their extension opens this framework to include gradual levels of trust in each dimension including the trustor, the trustee, and the object of trust. For our study, we are interested in the notion of humans' trust in automation [40]. Trust in automation also refers to a measure like the generalized trust as the prior trust

attitude of a person in automation. When studying trust in automation, other trust measures include the direct trust in a specific interaction (see also [45]) and the dynamic trust formation through time. Taking from these approaches to trust, we have assessed trust at three different points in time in the simulated automated driving maneuvers presented to participants: prior trust at the start of the experiment (before they watched any movie), direct trust (after each scenario was displayed) and post trust (after all three scenarios were watched).

According to the research done in the area of trust in automation [42], trust can be measured through understanding and trust sub-scales (the Competence and Intention of Developers subscales are not relevant to this study). Our subjective understanding question follows the Understanding subscale ("I was able to understand why things happened"). The direct trust question corresponds to the Trust in Automation sub scale ("I trust the system"). The question whether the user will disengage from a partially automated driving system in case the automated behavior is not clear to them, is related to the second question under this sub scale "I can rely on the system". Our prior trust question is related to the Familiarity subscale ("I already know similar systems" and "I have already used similar system"):

1. A-priori trust attitude to automated driving features: prior to showing any simulated movie. This is related to the innate propensity to trust as noted in [40].
2. Intermediate trust assessment: after watching each one of the three movies, we presented a direct question asking the participant to grade their trust level (asking directly about the user's trust level is discussed in [45]) in a 5 scale (strongly distrust, somewhat distrust, neither trust nor distrust, somewhat trust, strongly trust).
3. Posterior trust assessment: After all movies were watched, participants were asked about the degree to which the information provided, was helpful to trust the vehicle (possible answers were mandatory to help me trust the vehicle, help me trust the vehicle but not necessary, indifferent -may or may not help me trust the vehicle, does not help me trust the vehicle)

Understanding is also one of the trust components evaluated in human automation interactions (see [40,83]). As explained at the beginning of Section 5.1, we measured subjective understanding since in a fully autonomous driving setup the user is not operating the vehicle.

The data collected in this study served us to measure the independent variable, the information provided as HMI-only or HMI with XAI with these five measures: subjective understanding, prior trust attitude, direct trust, post trust and avoidance of manual takeovers.

The reliability of the subjective understanding and direct trust measures was verified by Cronbach's alpha. The control variables included age group (young (30–59 years old) and old (above 60 years old)), gender (female or male) and ADAS experience (exists or not). For the statistical analysis reported below (ANOVA F and z-test), we chose an alpha of 0.1 corresponding to a confidence level of 90%, even though we are considering a relatively large study (800 participants), due to the subjective nature of the questions. We have observed in [26] that such questions regarding trust and subjective understanding tend to have higher variance and multiple dependent variables, not all evident to us.

We use interaction plots to identify leading impacting variables on the outcome of our examined dependent variables. To verify the significance of these observations we conduct Analysis of Variances (ANOVA) to determine whether these differences are indeed statistically distinguishable. In the ANOVA, the F-statistics is employed (ANOVA F) with confidence of 90% (alpha=0.1). We performed the analysis on the Trust and Understanding questions over all three scenarios and the Take Over question, comparing between the population that received XAI (HMI + text) and the population that received only the HMI. We also employ ANOVA F on variables recognized as impactful in the interaction plots, specifically ADAS experience (see post-hoc analysis section).

Our goal was to estimate the value of adding XAI notifications to the dynamic HMI display in terms of understanding and trust measures. Therefore, we also performed an in-depth analysis examining the proportions of the participants giving specific answers. The graphs included show the percentage of participants that have answered any one of the possible answers for each question. That is, we present the percentage of the participants that chose certain scores indicating strong preference, weak preference or indifference for those measures. To determine statistical significance between the compared populations, we conducted hypothesis testing over proportions using the z-test. To depict this, we added confidence intervals corresponding to alpha equals 0.1 (90% confidence). Thus, the compared results are statistically significant if and only if the confidence intervals do not overlap.

Post-hoc analysis - Following upon the insight observed for the question posed on prior trust attitude of the participants, we conclude that the effect of XAI notifications needs to be studied by splitting the whole population pool into two distinguishable groups: those with ADAS experience and those that lack ADAS experience.

Additional questions on the participants' feelings of comfort and safety during the simulated drives were also presented. We found out that such measures are problematic to assess in an online simulated setup, as performed here. First, from our experience with simulated driving studies, there is a need to create scenarios interesting enough so the users can experience the automated driving in a simulated scene. Consequently, in some cases the driving might seem somewhat more aggressive than in real life. Second, since the experience is online and simulated, we might receive inaccurate feedback related to notions such as comfort and safety. Therefore, our focus in this paper is on the questions related to the understanding and trust levels of users as explained above.

5.4.1. Empirical results

The effects of adding XAI notifications to a dynamic HMI display in three simulated autonomous driving scenarios are summarized below along the four hypotheses raised.

H1: Age, gender, and ADAS experience affect the levels of subjective understanding and trust of users, when provided with XAI explanations in addition to the dynamic HMI display.

We start by emphasizing that the participants in both groups, XAI (HMI + text) and HMI only, viewed three different simulated driving movies. Thus, we first show that the questions regarding Trust and Subjective Understanding over these three movies are consistent. For this purpose, we conducted Cronbach's Alpha analysis.

The question on direct Trust (Question 10 in Appendix C) was presented to two groups of 400 participants respectively for three simulated driving movies. The value for Cronbach's Alpha for this question was 0.895 when the participants were provided XAI notifications in all three driving scenes they watched. The Cronbach Alpha was 0.864 when the participants were provided HMI only information. In both cases, the questions on Trust reached a good level of reliability.

The question on Subjective Understanding (Q8 in Appendix C) was presented to two groups of 400 participants respectively for three simulated driving movies. The value for Cronbach's Alpha for this question was 0.779 when the participants were provided XAI notifications in all three driving scenes they watched. The Cronbach Alpha was 0.770 when the participants were provided HMI only information. In both cases, the questions on Subjective Understanding reached an acceptable level of reliability.

Establishing the consistency shows that we can indeed consider all three scenarios when examining the effect on Trust and Subjective Understanding. Before exploring the effect of secondary variables, such as age, gender, ADAS experience and prior trust, we examine whether the two populations of those receiving XAI (HMI+text) and those receiving HMI only are distinguishable with respect to the Trust, Subjective Understanding and Taking control questions.

To examine the significance of these results we performed the ANOVA F analysis. For the direct Trust question (Question 10 in Appendix C) we had 1200 elements for each group (400 participants with 3 scenarios, n=1200). We had two groups (m=2). The mean values were 2.69 for the HMI+text group and 2.78 for the group of HMI only. Even though the difference in mean seems very small, the two populations are indeed statistically significantly distinguishable with respect to the Trust questions ($F=3.35617$, $p < 0.1$). For the full details on computing the ANOVA F analysis see Appendix D.

On the other hand, when examining the Subjective Understanding question as a whole, with the same degrees of freedom and means of 3.3475 for HMI+text and 3.29 for HMI only, we got no statistical significance ($F=1.6684$, $p>0.1$).

This suggests two important aspects that influenced our choice of analysis. First, that the mean analysis (as performed above) is too crude for the measure of Subjective Understanding and the alternative of examining the proportions of the different responses to the Subjective Understanding question might shed more light on this measure. Second, that examining the populations of XAI and HMI only as a whole might also be too crude, and additional (secondary) variables dividing the population might be significant in the understanding of the effect of XAI on the Subjective Understanding measure. This is directly related to the H1 hypothesis, and the interaction plots analysis performed next.

For the question whether a person would have taken control or not, we had 800 responses. The participants had three optional answers: 1 – Yes. Will take manual control. 2 – Maybe. 3 – No. Will not take manual control. We have two groups (those with XAI and those with only the HMI display, so m=2). The mean values were 1.375 for the group with XAI (experienced XAI in the previous questions), and 1.4525 for those that did not receive any XAI notifications.

Also, in this case we observe that there is a significant difference between the two groups with respect to the Taking Control question ($F=3.1414$, $p<0.1$). However, note that this difference is due to the previous experience of the participants within the experiment, since the movie presented to them for this question was identical and did not include any XAI. A reason for the difference could be that those that previously received XAI expected to continue and receive XAI and thus, were more likely to indicate that they will take control or might take control. This explains also the lower average value of their responses compared to those participants that have not received XAI (the HMI only group).

As mentioned above, we now want to examine the effect of age, gender, ADAS experience and even prior trust attitude using interaction plots. Our primary independent variable was, by definition, the XAI versus the HMI only. The secondary variable was either one of these independent variables. Within each group we averaged the Trust results over all users and all three scenarios (shown to be consistent with respect to the Trust and Subjective Understanding measures). A similar method was performed for the Subjective Understanding question. For the taking over question we averaged over all users, as each user received a single question (after

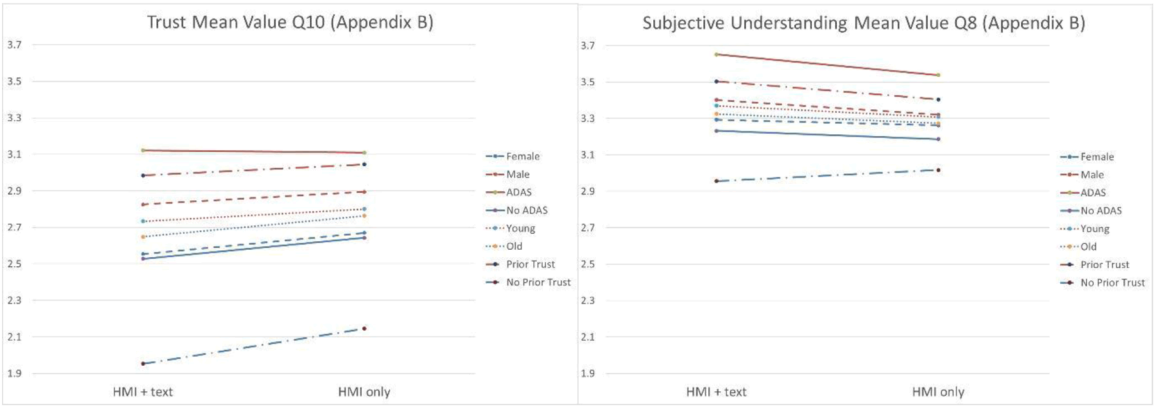


Fig. 14. Interaction plots showing the effect of the control variables on the Trust and Understanding measures.

observing the same scenario with no XAI). The interaction plots for Trust and Subjective Understanding are shown in Fig. 14.

In the Trust case we can clearly see that ADAS experience is the variable with the most impact, followed by prior trust and age. Gender seems to have the least impact on the Trust measure. For Subjective Understanding, prior trust is the most impactful, followed by ADAS experience, gender and age. Recall that users marked an option in the scale of 1 to 5, higher being that they either trust more or understand more. It can also be seen that the average results for Trust are lower than those for Subjective Understanding.

Similarly, we also created an interaction plot for the Taking Control question (see Fig. 15). Since the range of responses to this question is 1–3 the scale of the y-axis is different compared to the interaction plots of Trust and Understanding.

In this case we see that all variables seem to have considerable effect on the behavior, compared to the Trust and Understanding interaction plots. Similarly, we observe that ADAS experience is the most impactful variable. As mentioned, all participants viewed an additional movie with no XAI and we asked them whether they would consider taking control. Since this was done after they all watched the 3 movies (with and without XAI), we observed that the results are influenced by the participants' previous experience with these movies. Therefore, we did not include "taking manual control" in this hypothesis and left this aspect for future research. Still, we think it is interesting to view the effect of the secondary variables as supporting evidence to these factors' effect on Trust and Subjective Understanding.

We observe that ADAS experience is a leading variable of impact on all three dependent variables: Trust, Understanding and Taking Control. The second dominant variable is the prior trust (which in some cases is even stronger than the ADAS experience). ADAS experience is also an objective variable compared to prior trust which is self-reported and harder to quantify. Moreover, from a practical sense, a system of XAI based on ADAS experience can be easily implemented (e.g., previous usage of ADAS systems in the vehicle is a direct indicator). Thus, we focus on ADAS experience as a leading variable influencing the value of XAI, while keeping in mind that prior trust is also very significant and will be used in our post-hoc analysis.

As further support of the importance of ADAS experience and its effect on the outcome of the dependent variables, we conducted an ANOVA F analysis considering only this factor. With respect to Trust, the results indicate statistical significance ($F=138.882$, $p<0.1$); meaning there is a distinguishable difference between the two groups, ADAS and No ADAS experience with respect to the Trust measure. Similarly, for the Subjective Understanding ($F=91.227$, $p<0.1$); a distinguishable difference between the two groups with respect to the Subjective Understanding measure.

To emphasize the validity of these results and their significance, we also performed the ANOVA F analysis over the Taking Control measure, considering the ADAS and no ADAS populations. In this case we found no statistical significance ($F=0.32$, $p>0.1$).

Our explanation for this result is that the Taking Control question is mostly influenced by the prior experience of the participants during the study, either receiving XAI or not in the previous movies, as seen by the difference in mean over these two populations. All other effects seem minor compared to this main effect. As a result, in the sequel we examine the answers to this question for only those participants that received XAI, so as to remove this effect.

Since the two leading affecting variables are ADAS experience and prior trust we examined the correlation between the two.

A-priori-Trust attitude: *"I trust vehicles that enable automated driving features (such as Adaptive Cruise Control and future technologies). I will engage in automated driving once my vehicle has such a feature."*

Fig. 16 shows the distribution of the participants according to their prior trust attitude to ADAS. Interestingly, as shown further in Table 3, we found that prior experience with ADAS features (one of the questions presented to the participants at the beginning of the

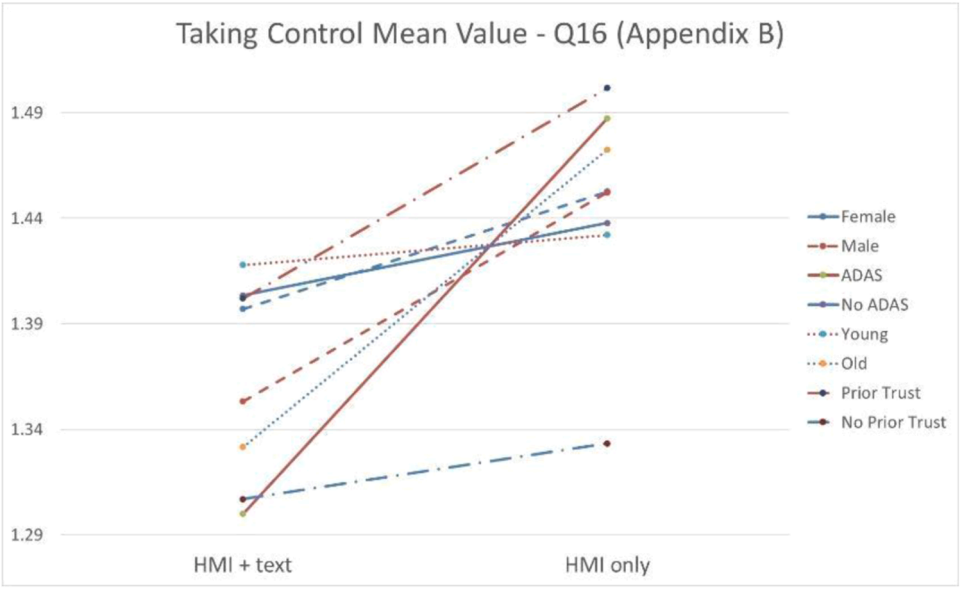


Fig. 15. Interaction plot showing the effect of the control variables on whether a user would take manual control.

study) is highly correlated with a positive prior attitude to trust. Correspondingly, lack of such ADAS experience is correlated with a negative prior attitude to trust automated driving features. As will be detailed in the next sections, this positive prior is also an important factor affecting when the XAI notifications help users increase their levels of trust and understanding of automated driving maneuvers.

Table 3 (top) shows the percentages of the population that attested to have experienced some ADAS feature (at least Adaptive Cruise Control) and answered positively to the prior trust attitude question (left top entry) or answered negatively to that question (left bottom entry). In the right column, the percentages of the population that answered positively or negatively to the prior trust question are shown for those participants with no previous ADAS experience. The values compared between the columns were found to be statistically significantly different. When running the Fisher's exact test, p value was almost 0, comparing the a-priori trust attitude to the ADAS experience ($p<0.000001$, this p value was obtained as part of the analysis results through the Qualtrics Stats IQ Software used for the analysis of the study results). This test shows high correlation between whether a participant agrees or disagrees to the trust statement and whether they have experience with ADAS features. The Chi-square statistical test also reached a similar p value as in the Fisher's exact text, showing that 84.1% is indeed statistically significantly larger than 43.8% and 15.9% is statistically significantly smaller than 56.2%.

Focusing on the Effect Size (Cramer's V), we see that the value of 0.335 for the Chi Squared Test and 0.379 for the Fisher's Exact Test resulting in a Medium Effect Size, supporting the low p value interpretation that Prior Trust and ADAS Experience are related. The Chi-squared and Fisher's exact tests were both run to assure that sample size is not a factor when determining if there is a relationship between the two variables, ADAS experience and prior trust attitude (Fisher usually is performed for smaller sample sizes while Chi-Squared is for larger sample sizes). We also utilized the SPSS Software for the Cronbach's Alpha and to double check the Chi Square analysis. Table 3(bottom) contains the standard deviation values for the variables in Table 3 (top). We see that Prior Trust has a standard deviation of 1.10 while ADAS Experience has a standard deviation of 0.45, indicating that Prior Trust responses are more dispersed with greater variance.

To better understand how ADAS Experience, Gender, and Age explain the Prior Trust, a regression model was employed. With a McFadden's R-squared of 9.1% we see that our tested variables explain part of the variance, but not all of it. The main driver of variance in Prior Trust was ADAS Experience, with a Relative Importance of 78%, followed by Gender at 15%, and Age at 6%. While the variables tested explain a small portion of the variance in Prior Trust the primary drivers were not captured in this study and may be difficult to quantify and assess in future studies due to the subjective nature of Trust. This further supports the importance of experience in emerging technologies playing a role in how willing and accepting people will be as advancements are made and enter the general market.

Both statistical tests, together with the regression analysis show that having ADAS experience is strongly correlated with strong prior positive trust attitude, while not having ADAS experience is strongly correlated with a strong prior positive for distrust attitude.

We now turn to examine the effect of each secondary variable. We do so by examining the proportions of the responses to get a more in depth understanding since we saw that examining the average responses smooths the distinction between the groups, specifically for the subjective understanding control variable. Our interest is thus in evaluating whether the distribution as a function of each control variable is different within each of the 2 groups (with or without XAI).

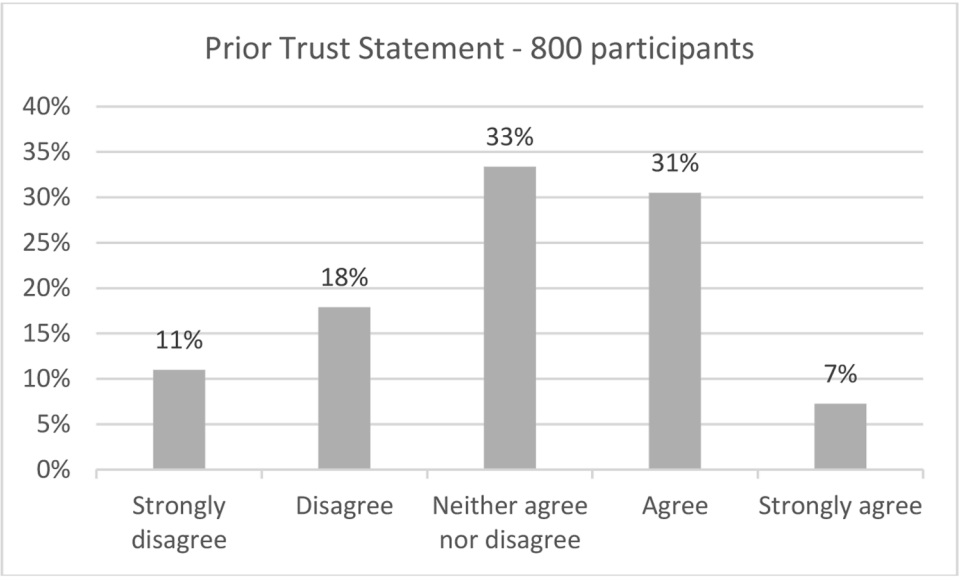


Fig. 16. A-Priori Trust Attitude Distribution in the Whole Population.

Table 3
(Top table) Prior Trust Attitude According to the ADAS experience of the Participants; (Bottom table) Standard Deviations.

	ADAS Experience Updated		
Trust Statement T2B B2B	ADAS Experience	No ADAS Experience	Total
Strongly agree/Agree (Top 2 Box)	84.1% B	43.8%	
Strongly disagree/Disagree (Bottom 2 Box)	15.9% B	56.2%	
Total	100%	100%	
	A	B	

Variable	Standard Deviation
ADAS Experience	0.452007674
Prior Trust (Question 1)	1.100510535
Both	1.618339716

Fig. 17 and Fig. 18 show that the distribution of trust and subjective understanding of participants with prior ADAS experience (no matter the information provided) is different than those with no ADAS experience. This can be seen by comparing the distribution of the light gray bars compared to the dark gray bars in either the HMI+text case (on the left-hand-side) or the HMI only case (the right-hand-side). In addition, we observe that for both trust and subjective understanding, ADAS experience helps participants trust and subjectively understand the system more than without ADAS Experience (for example, the values shown in Fig. 18, subjective understanding when XAI was provided is statistically significantly higher for those with ADAS experience, 66% and 48% respectively. Also, we reached significantly lower misunderstanding when XAI was provided for those with ADAS experience, 13% and 24% respectively).

Similarly, the distributions between male and female over the responses are also statistically significantly different for the trust measure (see Fig. 19). We observe that male participants trust the automated vehicle statistically significantly more than the female participants no matter what information was provided to them. Additionally, significantly more female participants distrust the automated maneuvers, no matter the information provided to them. The differences are not as sharp as we saw previously with respect to the ADAS experience control variable. For the subjective understanding, we hardly notice any difference between the distributions (see Fig. 20).

The results observed in Fig. 19 are consistent with similar results obtained across different areas, showing the influence that gender might have on self-reported measures. Gentile et al. [23] showed that gender may affect self-esteem levels in different specific domains, for example men scored significantly higher than women in physical appearance, athletic, personal self, and self-satisfaction

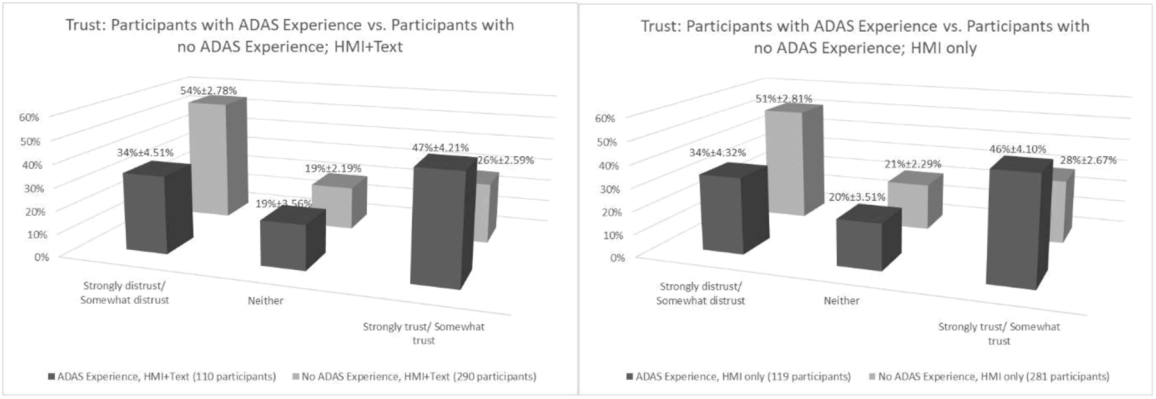


Fig. 17. Direct Trust Levels Comparing Participants with ADAS Experience to Participants without ADAS Experience when Automated Explanations Were Presented (Left graph) and when Automated Explanations were not Presented (Right graph).

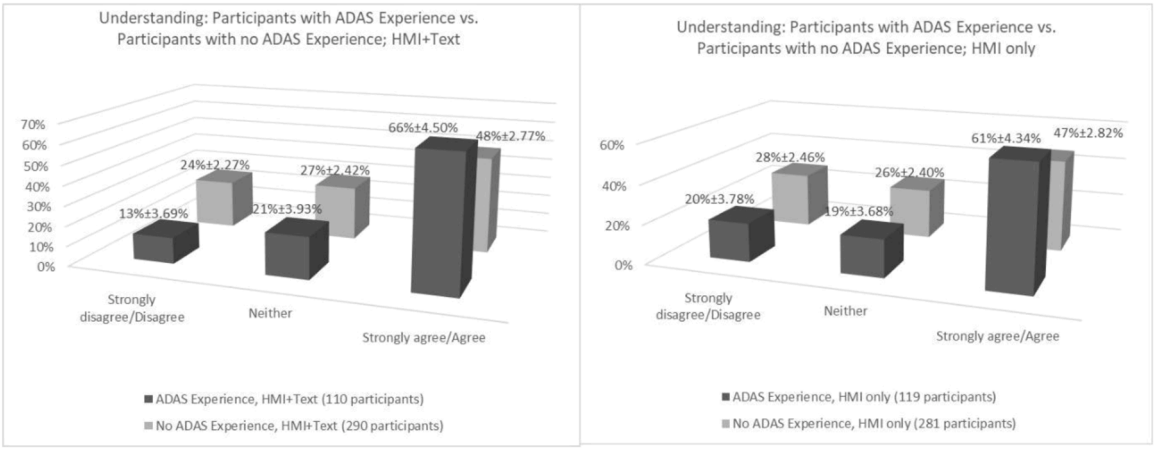


Fig. 18. Subjective Understanding Levels According to ADAS Experience Comparing Participants Provided with Automated Explanations and with HMI Only.

Fig. 19. The Gender Effect on Direct Trust Levels when Automated Explanations Were Presented and When These Were Not Presented.

self-esteem. Self-perceptions were also found to differ in tasks considered masculine (see Beyer et al. [70] where female self-evaluation of performance was inaccurately low. Females significantly underestimated their performance in tasks considered more masculine than feminine). Gender effect on driving tasks has also been studied in Song et al. [76] showing how the gender stereotype effect analyzed with traits and driving interest influences poor driving performance.

Fig. 21 and Fig. 22 show even a lesser difference between the distributions of the responses with respect to age, specifically in the HMI+text population. When HMI only information was provided, the difference is very significant between the two populations (young and old) whereas, when XAI notifications are provided this difference becomes almost indistinguishable.

H2: Automatic XAI explanations presented together with a dynamic HMI display help the users reach levels of trust higher than those attained when the dynamic HMI display was provided alone.

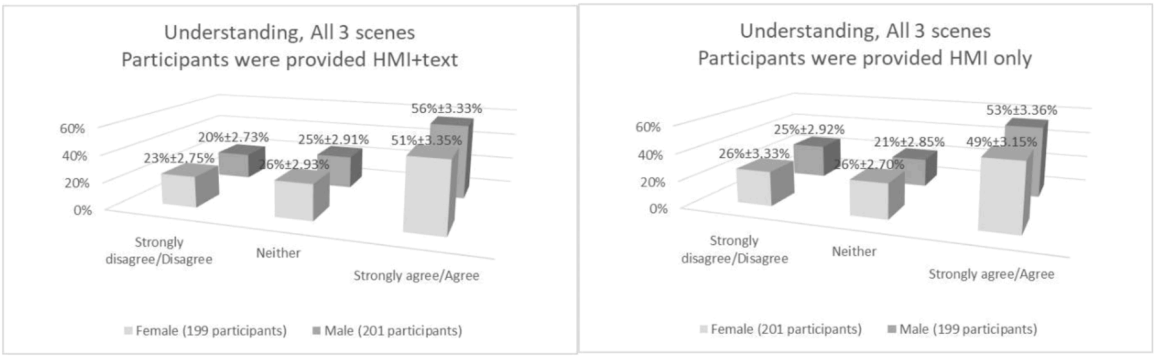


Fig. 20. The Effect of Gender on the Subjective Understanding Levels Over the Whole Population.

Intermediate Trust Level (Q10): *"Please indicate which of the statements below best fits your level of trust in the vehicle during the automated drive you just watched"* (Answers between 1–5).

Over the whole population, we do not see an effect favoring XAI notifications over HMI only. Fig. 23 shows the proportions of participants according to the answers submitted to the direct trust question. We gather all answers over all scenarios for each group (HMI only or HMI+text), assuming these answers are independent. The reported trust for both conditions was not high in general, and both values attained for the direct trust variable did not differ significantly (32% and 34% respectively).

We investigated, then, as part of a post-hoc analysis, groups of participants combining those control variables which were shown to affect the trust levels (see analysis in the context of Hypothesis 1). We refer to participants with a positive prior trust attitude to ADAS if they answered with the highest scores of 4 or 5 to the prior trust question (Question 1 in Appendix C). It is worth noting that focusing on groups of participants with both characteristics of having ADAS experience and positive prior trust attitude results in a relatively small pool of 143 participants, which makes the desire for reaching statistical significance more challenging. Therefore, we present results for the direct trust question, answered by participants with non-negative prior trust attitudes. That is, the group evaluated includes those participants that provided a score of 3 in the score scale for this question in addition to those that input the highest scores in the scale of 4 or 5. This larger pool of 202 participants allows for the inclusion of those that felt indifferent to the trust attitude question. These participants might not be willing to commit to trust ADAS features but they are not negative about them either.

We observe in Fig. 24 that XAI notifications significantly increase the direct trust levels of participants compared to the HMI only case when these participants are over 60 years old, with ADAS experience and with a non-negative prior attitude to trust ADAS features. For this group, Hypothesis 2 is supported. Interestingly, we note in Fig. 24, that if we only take the ADAS experience away of this group, Hypothesis 2 is not supported anymore, both values attained for the direct trust variable differ significantly (34% and 39% respectively). As can be seen in the Fig., all values for the younger group overlap considering the confidence values.

To evaluate how helpful the XAI notifications were, we refer to the data collected for Question 12 in Appendix C:

Posterior Trust Assessment (Q12): "Thinking back to the various automated driving maneuvers you just watched, which of the following statements best describes how you feel about the information, in addition to the traffic display that was provided in the videos?" The question asks whether the computed notifications, provided in addition to the HMI display, help you trust *the automated vehicle?*

This question was given to all the 800 participants, 400 in the "HMI only" group and 400 in the "HMI+text" group. It is problematic to compare between these two populations with respect to this question, because each population looked at different solutions. Since in this study we are interested in the value of automated explanations, we focus here on the 400 participants that did see the explanations in addition to the HMI display.

More than half of these participants (53% of 400 participants) find the XAI notifications helpful to trust the automated vehicle (see Fig. 25). Having a positive trust attitude to ADAS features, increases this number to 75%. The age of the participants affects this number as well: 78% (80%) of the younger group find the automated explanations helpful to trust automated driving (with No ADAS or with ADAS experience respectively); whereas 70% (72%) of the older group find the information helpful for trusting the vehicle (with No ADAS or with ADAS experience respectively).

Hypothesis 2 is supported by the results obtained for the group of participants over 60 years old, with ADAS experience and non-negative prior trust attitude to ADAS features.

H3: Automatic XAI explanations presented together with a dynamic HMI display help the users reach levels of subjective understanding higher than those attained when the dynamic HMI display was provided alone.

Subjective Understanding Level (Q8): "Thinking about the video you just watched, to what extent do you agree or disagree with the following statement: "I have a better understanding of why the vehicle performed the maneuver shown in the scenario." (Answers between 1–5).

Over the whole population of 800 users (Fig. 26), most of the participants understood well the driving maneuvers, expressed by the high scores entered of 4 or 5 (51% of the participants in the group assigned "HMI only" information and 53% of the participants in the group assigned "HMI+text" information). However, no statistically significant differences were found between these two options. Nevertheless, note that significantly less users that received XAI notifications understood less than those that received the HMI only

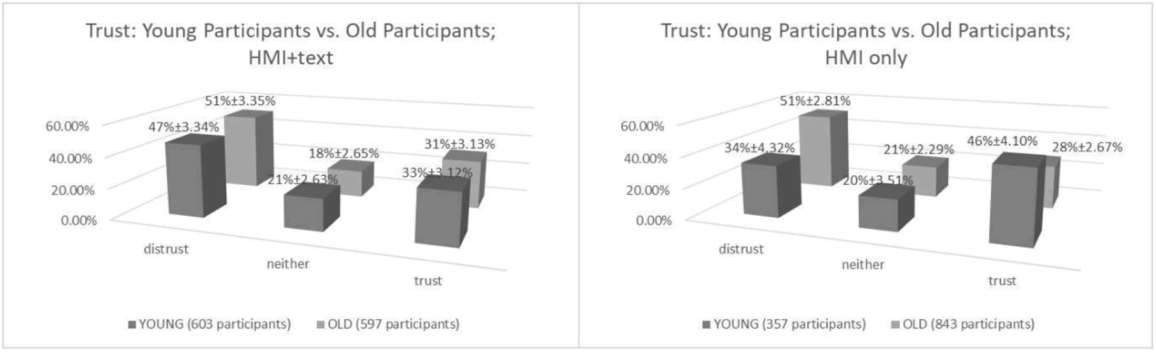


Fig. 21. The Effect of Age on the Direct Trust Levels Over the Whole Population.

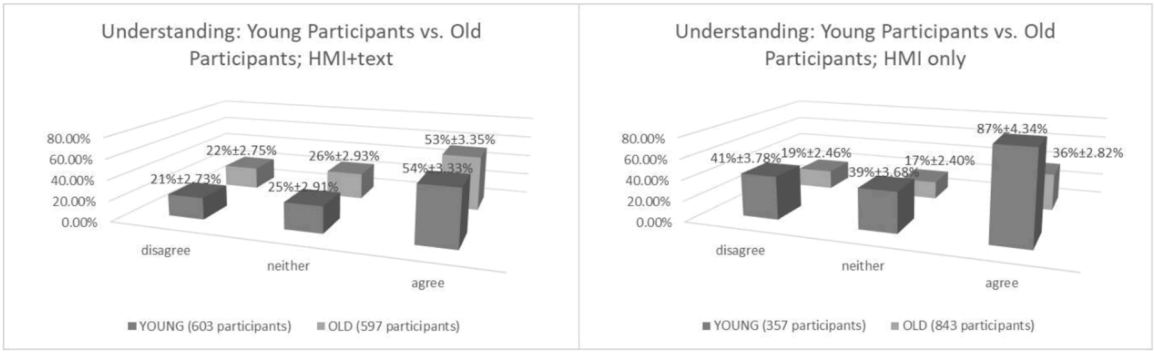


Fig. 22. The Effect of Age on the Subjective Understanding Levels Over the Whole Population.

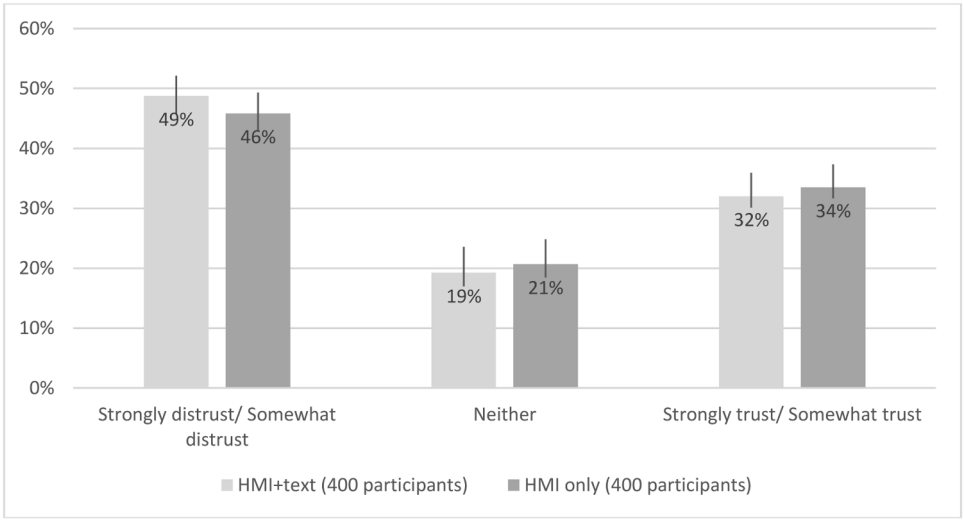


Fig. 23. Direct Trust Levels Over the Whole Population (800 participants).

display (21% and 25% respectively).

For the subjective understanding measure, we also investigated different groups of participants as part of a post-hoc analysis, combining those control variables which were shown to affect the self- reported understanding levels (see analysis in the context of Hypothesis 1). A negative attitude to automation (scores of 1 or 2 in the prior trust attitude), decreases the level of understanding of the participants, no matter the information provided (see Fig. 27). Interestingly, significantly less participants that received the computed explanations understood well the driving maneuvers, compared to those that were provided only an HMI display (32% compared to 43%; also 30% and 22% were found statistically significantly different). There might be an effect of providing too much information with XAI to those that have a negative prior trust attitude to automated driving. This "overflow" of information might cause more confusion than benefit. Alternatively, people with prior negative attitude might require different types of information to first calibrate their trust attitude more positively towards an automated system. Such users might be more familiar with the information presented in HMI displays so they are more prone to understand this information more naturally. This would need to be validated in a further study.

On the other hand, observing those with positive trust attitude, we found that there are significantly more participants that did not understand the driving maneuvers when HMI only was presented (18% and 12% are statistically different with 90% confidence). See Fig. 27.

Analyzing the data for the groups of participants with ADAS experience (see Fig. 18 presented in the context of H1), we observe that providing XAI notifications to this population is enough to reach significantly larger levels of subjective understanding than with the dynamic HMI display alone (66% and 61% are indeed statistically significantly different with confidence of 90%). Furthermore, a significantly smaller group of participants that received XAI notifications understood less than those that saw the display with the HMI only option (20% and 13% were found to be statistically significantly different with confidence of 90%). When the participants do not have ADAS experience, the levels of understanding are indistinguishable. It is noteworthy mentioning the analysis of the data of larger sets of participants, including those in both age groups, with non-negative prior trust attitude (i.e., including those participants that entered a score of 3, 4 or 5 to the prior trust attitude question). For this group, even when they lack ADAS experience, XAI notifications added to the dynamic HMI display helped these users to reach higher levels of subjective understanding compared to those that

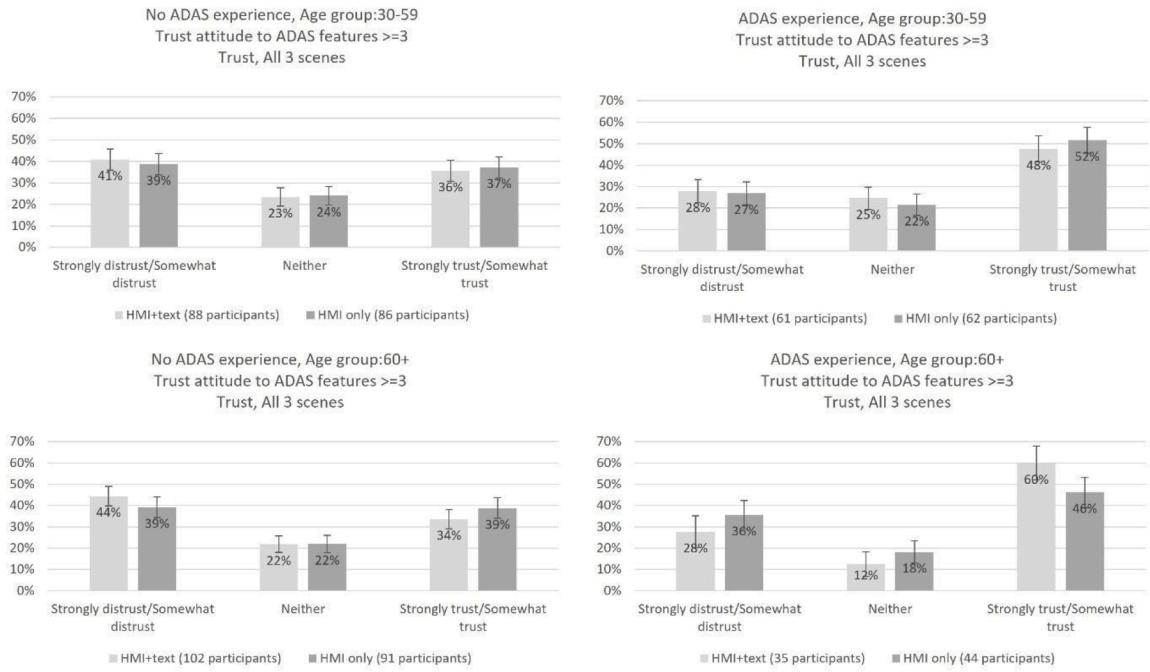


Fig. 24. Direct Trust Levels Obtained in Populations Split According to Age, Prior Attitude to ADAS and ADAS Experience.

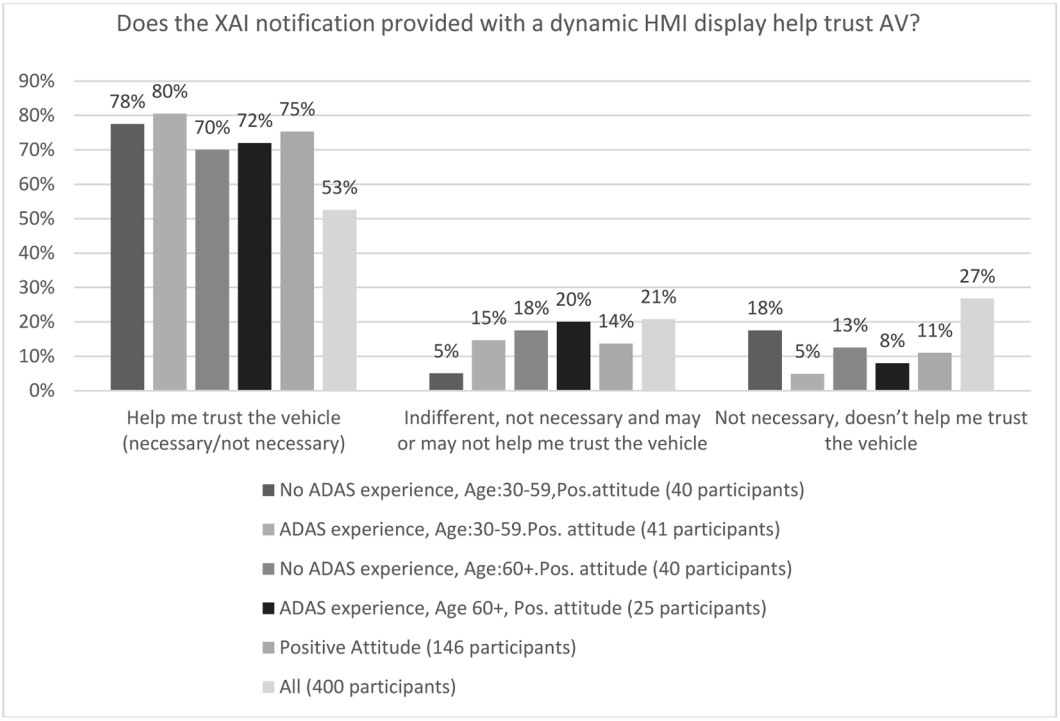


Fig. 25. Posterior Trust Levels of Participants Provided with XAI Notifications.

received the HMI only option (see Fig. 28, left top and left bottom graphs): In the older group, 56% and 39% were found to be statistically significantly different as well as 18% and 26%. XAI notifications helped the older group with ADAS experience and non-negative prior trust attitude to understand significantly better than those that received the HMI alone (all values are statistically significantly different in the right bottom graph).

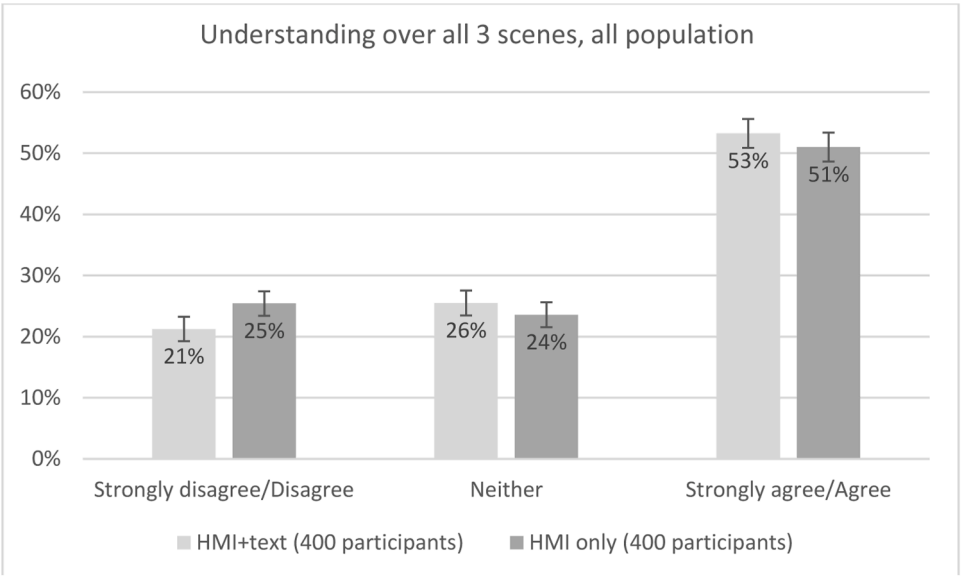


Fig. 26. Understanding Levels of the Whole Population Comparing Participants that received Automated Explanations to Those that Did Not Receive Automated Explanations.

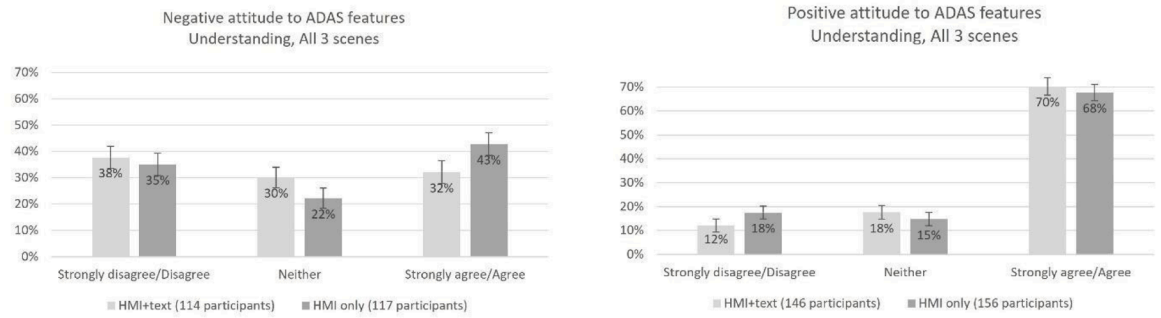


Fig. 27. Understanding Levels of the Participants According to Prior Attitude to ADAS. Comparing Automated Explanations Provided to- HMI-only.

In summary:

- XAI notifications together with HMI reduce the subjective misunderstanding level of participants over the whole population compared to those that received only the HMI display.
- Participants with ADAS experience that were provided XAI information reach statistically significantly higher levels of subjective understanding (and lower levels of subjective misunderstanding) when comparing to those that received HMI only information.
- Lacking ADAS experience but having a non-negative prior trust attitude to ADAS also results in XAI providing value in higher subjective understanding and lower subjective misunderstanding of automated maneuvers compared to HMI only information. Statistical significance was found for the older group.
- We noticed that participants with negative prior attitude to ADAS features can understand less than those provided with HMI only information. And participants with positive prior trust attitude understand less when provided HMI only rather than XAI notifications.

H4: Automatic XAI explanations (i.e., Plan-next or Contrastive XAI algorithms presented in this paper) reduce the number of manual takeovers in scenarios when users might have preferred to take over when provided a dynamic HMI display only.

We have presented the whole population, of 800 participants, a fourth animated movie, showing a simulated ramp merge automated driving maneuver. All participants were provided solely a dynamic HMI display (with no additional textual explanations). A large majority of this population, 80% (744 participants), answered that they would be interested in taking manual control of the automated driving vehicle in this scenario.

These participants were then shown two pictures with two possible notifications, explaining the maneuver. These explanations were computed following the XAI algorithms as described in Section 4 (Plan-Next and Contrastive). These participants were then asked

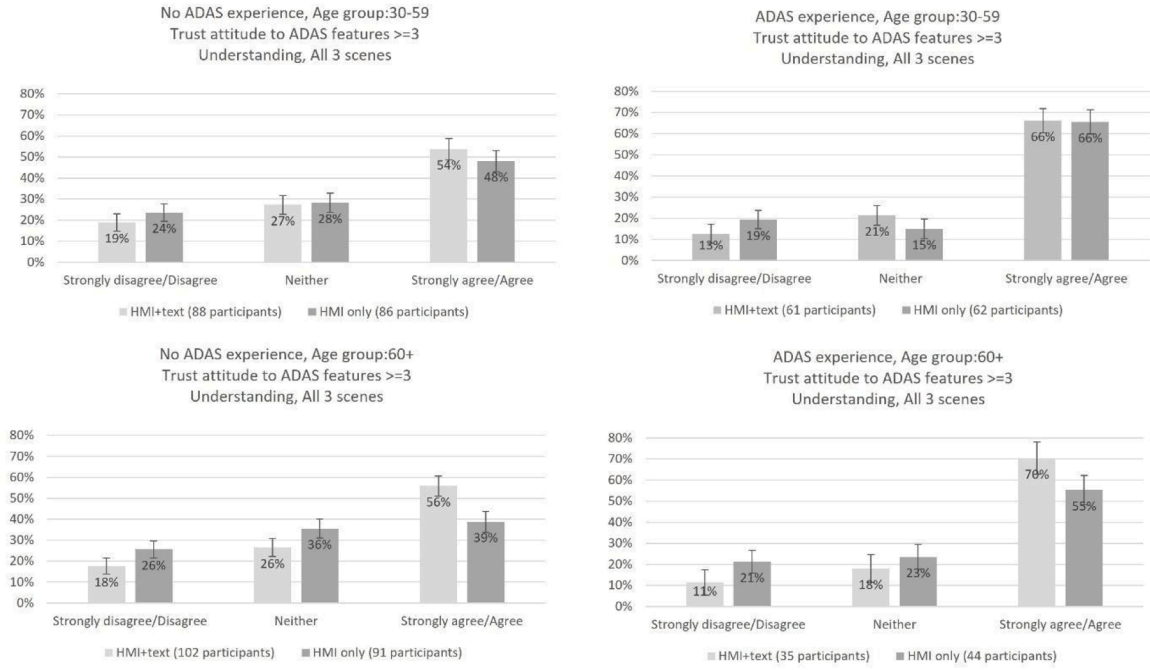


Fig. 28. Understanding Levels Obtained in Populations Split According to Age, Non-negative Prior Trust Attitude and ADAS Experience.

whether any one of these informative notifications might have prevented them from taking manual control of the vehicle. The results presented in Fig. 29 (lowest bar) show that from the 744 participants in the whole population, willing to take manual control, 48.3% of them would have avoided the manual take over if provided any one of these explanations. That is, the XAI notifications helped almost half of the participants to remain engaged in automated driving, no matter their prior trust attitude to ADAS features. This group of participants include people with different prior trust attitudes (positive, negative and indifferent). We note (consistently with the insights presented with respect to H1) that this prior attitude affects the response of these users to take manual control in case that automated notifications are presented to them explaining the vehicle behavior.

The XAI notifications mostly affected the group with the positive prior trust attitude (shown in the second bar from the bottom in Fig. 29: 68.7% of the 278 participants would have avoided the manual take over if any one of the explanations computed was provided. No statistically significant difference was found among the information options, Plan-Next and Contrastive (see on users' preferences

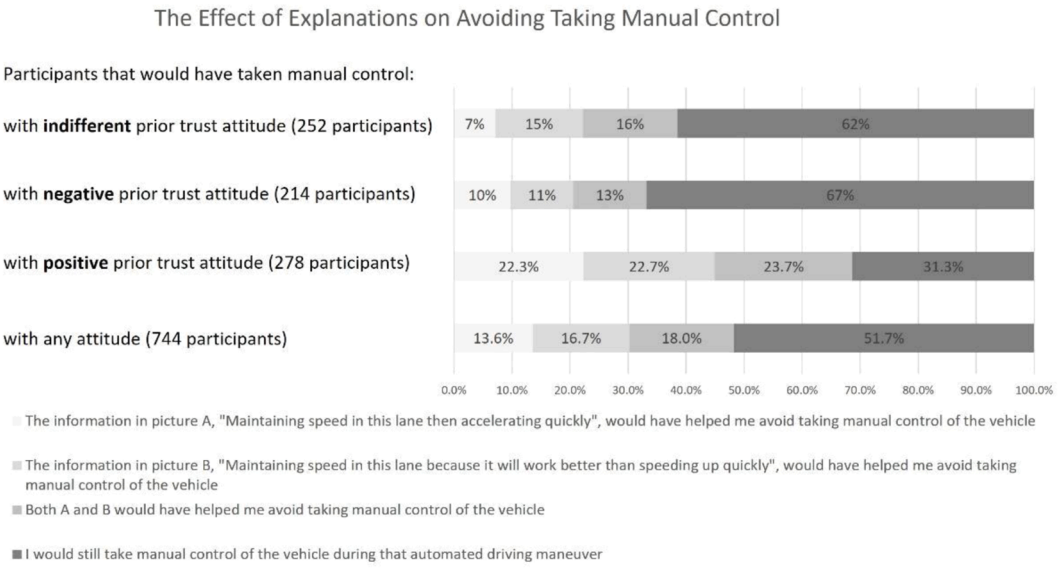


Fig. 29. The Effect of XAI Automated Explanations on Users' Decision to Avoid Manual Takeovers.

for types of explanations in the next section). As can also be observed from Fig. 29, 38% and 34% of those with indifferent or negative prior trust attitudes (respectively) would have also avoided taking manual control when XAI notifications were provided to them. Thus, **Hypothesis 4 is supported by this data**, at the minimum by 34% of those participants with negative prior trust attitude to ADAS and at the best it is supported by 68.7% of those participants with a positive prior trust attitude to ADAS (68.7% and 31.3% are statistically significantly different with 90% confidence).

5.5. Study limitations

The end goal of the proposed computational solution should be verified in-vehicle ride scenarios when drivers interact with an automated driving system in real-time, in real traffic scenarios. Since it is currently not possible to deploy this setup for practical and safety reasons, we opted to assess the automated explanations through simulation. Clearly, there is a gap between simulated scenes and real-life scenarios.

Another limitation resulting from performing this study on simulated automated driving scenarios refers to the self-reporting of measures by the users. In that sense it can be claimed that users' responses about trust suffer from some weaknesses. For example, self-report of trust measures has been noted to be not perfectly aligned with actual trust behavior [40]. However, these authors also mentioned that self-reporting is the most common method of assessing trust in automation. The alternative to measuring trust indirectly through physiological measurements would have its own limitations: a limited set of participants will comprise such a study that would require physical participation, and the need to place hardware and sensors on the participants which might have potential effect on the users' comfort during the study. This might also affect the results. Interviewing the participants can provide additional information regarding their understanding of the scenario presented that could confirm or not the users' self-report of understanding levels. In an autonomous driving scenario as assumed, the user is not requested to intervene in the driving maneuver. In other domains, a more objective understanding could have been assessed by testing how the user performs if they understood correctly the system behavior. This was not the case in the domain we chose. Also, there is a trade-off between the size of the participants' pool that provides more data and thus the option to reach statistical significance versus a physical interview that can provide more direct information. We believe that we have counterbalanced this weakness with a very large size of participants' pool (of 800 participants) that have been recruited for this study, providing us with sufficient self-reported data. This number is very high compared to the size of pools usually reported in the literature.

Another limitation is that due to the time and cost limitations to run a simulated study with participants, we did not invite the same participants to repeat this study many times. That is, we could show the participants only three different automated driving maneuvers. Hence, we could not assess their dynamic trust formation in the system [40].

It was also shown in [40] that many studies lack the assessment of trust prior and after the study. Here, we have indeed evaluated the trust levels at three points during the time of the study: prior attitude to trust in ADAS features, direct trust question after each simulated driving scenario has finished playing and after all scenarios have been shown covering prior, direct and post questions about trust.

6. Summary and conclusions

Establishing the appropriate trust level of users in automated systems is a critical and non-trivial challenge. This paper studies explainable AI algorithms applied to decision-making under uncertainty planners to provide users with real-time automated notifications. The value of these explanations in establishing trust in automated driving maneuvers has been assessed as part of a large online user study with 800 participants. Our XAI solution included computing three different types of automated explanations (Plan-Next, Contrastive and Value), three scheduling mechanisms to time these explanations (constant, rules and dominance) and an assessment of the value of textual information compared to only graphical information. The computational capabilities of this solution were demonstrated on a model-based reinforcement learning planner, the AlphaZero planning algorithm (including a MCTS module). This emphasizes the wide applicability of our XAI solution to real-life stochastic sequential decision-making problems that can benefit from higher transparency in the interactions with their users. The proposed explainable AI algorithms are general, in the sense that they can be applied to any tree data structure, resulting from an AI planning process (i.e., a data structure resulting from the information directly stated in a given model or a data structure implied by the information collected during the training or execution stages of a model-free planner as explained in Section 3). The algorithms are not strictly about the driving scenes presented but about the reasoning of the planning system.

The main insights gained from the analysis conducted on the data collected in the large-scale online study are summarized below:

- Hypothesis #1: Age, gender, and ADAS experience affect the levels of subjective understanding and trust of users, when provided with XAI explanations in addition to the dynamic HMI display: Age, gender and ADAS experience were found to affect the subjective understanding and trust measures. ADAS experience and prior trust attitude were shown to be the main factors among the control variables examined. We also observed that ADAS experience is highly correlated with prior trust attitude, allowing to focus mostly on the more practical variable of ADAS experience.
- Hypothesis #2: Automatic XAI explanations (i.e., Plan-next, Contrastive and Val XAI algorithms presented in this paper) presented together with a dynamic HMI display help the users reach levels of trust higher than those attained when the dynamic HMI display was provided alone: The data that supported this hypothesis, with statistical significance, indicated that the participants over 60 years old with ADAS experience and a non-negative prior trust attitude to ADAS features reached higher levels of trust when

provided additional XAI notifications compared to the HMI only display. Interestingly, only deducting the ADAS experience variable to this particular group results in a significant reduction in trust level when XAI was provided, comparing this to the HMI only case. For the young age group, the trust levels were found to be indistinguishable when comparing XAI with dynamic HMI to dynamic HMI alone.

- Hypothesis #3: Automatic XAI explanations (i.e., Plan-next, Contrastive and Val XAI algorithms presented in this paper) presented together with a dynamic HMI display help the users reach levels of subjective understanding higher than those attained when the dynamic HMI display was provided alone: Over the whole population, we found that those provided with additional XAI information misunderstood significantly less than those provided only a dynamic HMI display (supporting this hypothesis only in one direction). Those participants with ADAS experience understood significantly better when provided XAI compared to HMI only (they also misunderstood less) supporting fully this hypothesis. Moreover, those participants that lacked ADAS experience but had a non-negative prior trust attitude to ADAS also understood the automated driving maneuvers better than those provided only HMI (and they also misunderstood less). Statistical significance was found for the older group. Interestingly, the data from those participants with a negative prior trust attitude to ADAS does not support this hypothesis: providing additional XAI notifications resulted in significantly lower subjective understanding than that reached by those provided with a dynamic HMI display only.
- Hypothesis #4: Automatic XAI explanations (i.e., Plan-next or Contrastive XAI algorithms presented in this paper) reduce the number of manual takeovers in scenarios when users might have preferred to take over when provided a dynamic HMI display only: The data analyzed supports this hypothesis. All participants have chosen any one of the explanations provided, to avoid a manual take over. The highest proportion of participants that reduced this number the most included those participants with positive prior trust attitude. In this case, 68.7% of the users that would have wished to take manual control when shown only an HMI, changed their answers to have avoided this manual takeover if XAI notifications would have been provided. This result, of course, needs to be validated in a real driving setup where drivers can actually perform a manual take over.

These results contribute to the progress of the study of automated explanations towards improving human-machines' interactions and beneficial use of automation. This paper supports the efforts of developing XAI algorithms for information that was previously shown to be beneficial when it was provided manually or hand-crafted. These XAI algorithms were developed on a stochastic planner that can handle real life, complex problems and were evaluated in a very large-scale study compared to other studies reported in the literature of a much smaller scale. One specific group of participants, including those over 60 years old with ADAS experience and non-negative prior trust attitude was found to fully support Hypotheses 2 and 3 (see Fig. 30). Larger and more varied groups of participants benefited from the XAI notifications in terms of increasing their subjective understanding of an automated maneuver. We did not find similar significant results for the trust measure for other groups. Future work is still needed for developing further types of explanations that can accommodate contextual human preferences (i.e., where XAI information can help increase both subjective understanding and trust), optimizing the timing when to provide an explanation, and real-life validation tests with sufficiently large pools of participants.

CRediT authorship contribution statement

Claudia V. Goldman: Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration,

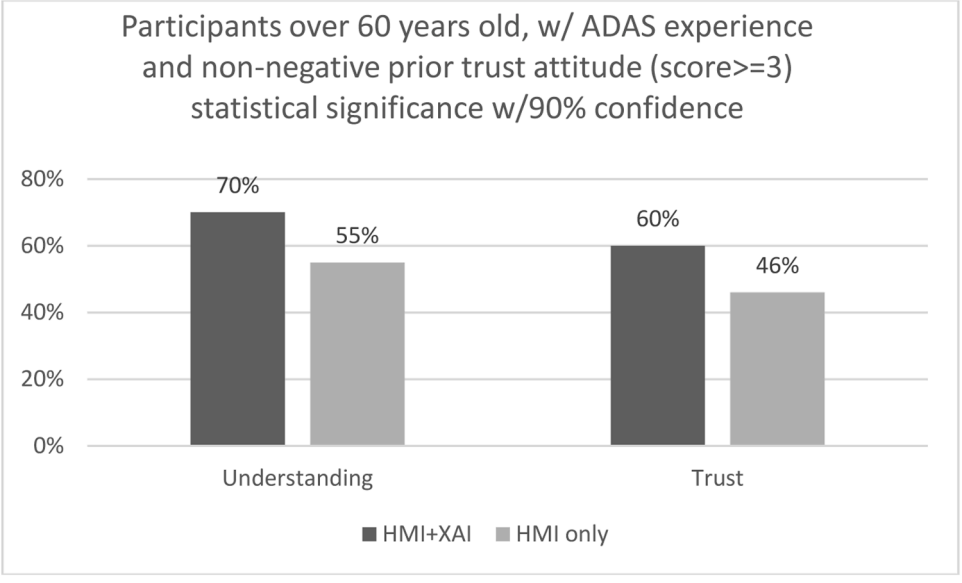


Fig. 30. The set of participants that benefited the most from XAI notifications in both measures (trust and subjective understanding).

Methodology, Investigation, Formal analysis, Conceptualization. **Ronit Bustin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Wenyuan Qi:** Validation, Software. **Zhengyu Xing:** Validation, Software. **Rachel McPhearson-White:** Investigation, Formal analysis, Data curation. **Sally Rogers:** Supervision, Resources, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Claudia V. Goldman has patent Generation and presentation of explanations related to behavior of an automated control system pending to GM Global Technology Operations LLC. Claudia V. Goldman and Ronit Bustin have patent System and methods for computing automatic notifications about a given AI policy of behavior pending to GM Global Technology Operations LLC. Claudia V. Goldman and Ronit Bustin have patent System and method for soft understandings of automated decisions pending to GM Global Technology Operations LLC. Claudia V. Goldman and Ronit Bustin have patent System and method for analyzing the structure of a probability tree pending to GM Global Technology Operations LLC. Claudia V. Goldman and Ronit Bustin have patent Human Data Driven Explainable AI System and Methods pending to GM Global Technology Operations LLC. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank their colleagues at GM, Omer Tsimhoni, Yael Shmueli, Andrea Forgacs, Yuval Zak, Gershon Celniker, Santhosh Thirunavukkarasu, Monica Malden-Stevens, Jeff Cooke, Rob McCulloch, Grady Beerck, Nate Rogers, Joe Vichich, Lisa Talarico, Ginamarie Mick, Guy Zohar, and Ofer Saraf for contributing at different stages of this work.

Claudia V. Goldman was partially supported by the David Goldman Data-driven Innovation Research Center at the Hebrew University Business School at the Hebrew University.

Appendix A. Dominant Elements

Given a vector of elements (either natural values or a probability vector), determining which of the elements are the dominant ones does not have a single answer. Naturally, it depends on how we define dominance. A very basic approach is to determine some threshold value. Every element above this threshold will be considered as a dominant element, whereas all those below this threshold will be considered not dominant. This requires us to determine the threshold, which is a problematic task. Instead, we used a different approach which does not require us to set a threshold and tries to mimic what humans would consider dominant.

The approach suggested here reduces the input vector by finding the subset of elements within the vector that are most likely to be uniformly distributed assuming all other elements do not play a role [16].

The approach uses the Jensen-Shannon Divergence (JSD), which given two distribution measures P and Q is given as follows: $JSD(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$, where $M = \frac{1}{2}(P+Q)$ and $D()$ is the Kullback-Leibler (KL) divergence measure, $D(P \parallel Q) = \sum_{x \in X} P(x)\ log\ \frac{P(x)}{Q(x)}$. The JSD measure is symmetric and non-negative. It equals zero iff $P = Q$ as measures.

The input to this approach is a probability vector (thus, any input must be non-negative and normalized). We denote the input probability vector as P. The steps of the algorithm are as follows:

a). Sort P from high to low $\rightarrow \widetilde{P}$, remember the permutation of P
b). Initialize *results* as a zero vector of length similar of P
c). For $j \in \{1,\ 2,\ \ldots len(P)\}$:
  I. Construct $Q = \left(\frac{1}{j} \ldots \frac{1}{j}, 0 \ldots 0\right)$ with j non-zero elements. Q is of the same length as P.
  II. $results[j] = JSD(\widetilde{P} \parallel Q)$
d). Optimal_index =argmin{results}
e). Return Q calculated with j equal to the Optimal_index and after reversing permutation (permutation done to P).

Thus, the approach is to compare the input probability vector P to several uniform vectors, each over a different subset of elements. However, we do not need to consider all subsets of P since if an element is dominant, any element greater or equal in value will also be dominant. Thus, we need to consider only $len(P)$ such uniform vectors and calculate the JSD between them and the input vector (sorted). The idea is that being close to uniform is similar to how humans conceive the first order structure of the vector. For example, we will determine that the vector has a single dominant element if the uniform vector closest to it (minimum JSD value) is a single shot, meaning $(1,\ 0 \ldots 0)$. We will determine that there are two dominant elements if the closest uniform vector is $\left(\frac{1}{2},\ \frac{1}{2},\ 0 \ldots 0\right)$, etc.

As an example, assume that the set of actions, available to the planning system consists of five possible actions: RightLaneChange, LeftLaneChange, Accelerate, Decelerate and None (None refers to the action that keeps the vehicle driving at the same speed on the

same lane). Our objective is to compute which actions are dominant in a particular node, for example at the root node of an MCTS tree (referred to as s). We compose the distribution vector corresponding to the probabilities of choosing actions at s, by referring to the total number of visits, this node s has been visited by the planning algorithm (*NumVisits(s)*) and by referring to the number of visits this node s has been visited for any particular action in the set of the five discrete actions (*NumVisits(s,a_i)*). The first parameter stores the number of times the root node s was evaluated, and the second parameter refers to the number of times that a particular action a_i was evaluated at this root node (leading to the same child node). Given the five actions assumed, we can compose the following probabilities' vector:

$$p = \left[\frac{NumVisits(s,a_1)}{NumVisits(s)}, \frac{NumVisits(s,a_2)}{NumVisits(s)}, \frac{NumVisits(s,a_3)}{NumVisits(s)}, \frac{NumVisits(s,a_4)}{NumVisits(s)}, \frac{NumVisits(s,a_5)}{NumVisits(s)}\right]$$

Then, we compute the Jensen-Shannon divergence measure between this current vector (sorted in decreasing order of their values in [0,1]) and each one of the following uniform distributions. The options are:

- 5-dominant actions [1/5, 1/5, 1/5, 1/5, 1/5]
- 4-dominant actions [1/4,1/4,1/4,1/4, 0]
- 3-dominant actions [1/3, 1/3, 1/3, 0, 0]
- 2-dominant actions [1/2,1/2,0,0,0]
- 1 dominant action [1,0,0,0,0]

Finally, this process returns the distribution that attained the minimal Jensen-Shannon divergence.

A second example, for which we compute this Jensen-Shannon divergence, includes finding out the dominant factors of the reward function when a certain action is chosen for execution by the planner. In the example implemented, we had 4 reward factors. We express these factors as a probability vector by shifting and normalizing this vector. Then, we find the JSD measure between this probability vector to any one of the following uniform distributions:

- 4 dominant factors: [1/4, 1/4, 1/4,1/4]
- 3 dominant factors: [1/3,1/3,1/3,0]
- 2 dominant factors: [1/2,1/2,0,0]
- 1 dominant factor: [1,0,0,0,]

Finally, this algorithm returns the distribution that reached the minimal JSD, from which we can understand which factors were the most impactful for the planner's choice of action.

Appendix B

As mentioned, in our setting we used a discrete action space composed of seven driving maneuvers. The state space, on the other hand, is a fuller description of the environment. It contains information about the ego vehicle, and relative information about up to 20 surrounding vehicles. These vehicles are chosen by their distance from the ego (the 20 closest vehicles, with a limit on the radius considered). The information about the ego includes the following: its longitudinal speed and lateral speed (both normalized by a preset maximum speed), a lane value normalized by the total number of lanes in the road and a lane change status which is a trinary value taking one of {-1,0,1} to indicate whether there is a lane change to the right, no lane change or a lane change to the left. Finally, we also added the distance to the ego's destination normalized by the diagonal of the map.

For each of the surrounding vehicles the state contains a binary indicator whether the vehicle is in radius or not (to handle cases in which there are less than 20 vehicles of interest), a longitudinal and latitudinal relative position, a longitudinal and latitudinal relative speed, a relative heading and a trinary lane change status as provided for the ego vehicle. The definition of the state is motivated by the one described in [30].

AlphaZero is an algorithm used in the framework of reinforcement learning [79]. It has both an off-line stage in which a neural network (NN) is trained, and an on-line stage. In both cases the operation is similar so we will first focus on the on-line stage and then on the training stage. For every state encountered during the drive, for which the planner needs to determine the next action to perform, the algorithm defines this state as the root of a tree and begins expanding the tree following the Monte Carlo Tree Search (MCTS) algorithm, meaning it follows four distinctive steps:

1. Selection: starting from the root node, the selection process goes down the tree until it reaches a leaf node. The specific selection used is the Upper Confidence Bound applied to Trees (UCT) which balances exploitation of actions that have shown to be favorable, and exploration of actions that have not been sufficiently explored.
2. Expansion: when reaching a leaf node, it can be a terminal node, in which case it cannot be expanded. If the node is not a terminal node the tree is expanded to the set of actions that can be considered from this node.
3. Simulation/Neural Network (NN): the properties of the expanded node can be evaluated using a Monte-Carlo simulation, where multiple rollouts reaching a terminal state are performed to assess the value of the expanded node, and a prior distribution over its actions. In the AlphaZero family of algorithms the off-line learning of the value and prior per state (node) replaces this. The off-line

learning is summarized in a NN that for a given state gives an estimation for the value of the node (the accumulated discounted reward from the corresponding state) and an estimation for the prior distribution over the possible actions from this node.

4. Back-propagation: after expanding and extracting the value and prior distribution of the node, this new information is back-propagated all the way up the tree to the root node. The update is to the accumulated value at each node and the number of visits at each node.

Once the time budget ends, for the tree created that far, the algorithm chooses what action to execute based on the action taken at the root state that attained the maximal number of visits (information stored and processed in each node according to the above 4 steps).

For the implementation of the MCTS we used the open-source available ray library (https://docs.ray.io/en/latest/data/user-guide.html) version 1.9. Under rllib/contrib/AlphaZero the library included an implementation of AlphaZero which we slightly adjusted to our specific needs. We performed a decision step every 1 second, meaning every 1 second a MCTS has been constructed. We allowed a budget of 60 expansions of the MCTS during training and 100 expansions during evaluation (real-time). For the discounted accumulated reward, we used the default value for γ equals to 0.9. We gave equal balance to the exploration and exploitation elements in the UCT selection.

Training was done iteratively. The starting point was a randomly initialized NN. We first use the current NN to simulate driving through a specific scenario, and collect tuples of (state, value, value components, prior) from the root of the constructed MCTS. These samples are then used for retraining of the NN. We performed between 30 – 100 such iterations, depending on the complexity of the scenario. The training of the NN was done over train batches of 200 and with a rollout fragment length of 28. The NN contained three heads, the first two producing the value and value components and the third head producing the priors. The NN itself used two layers of convolutional layers and 2 layers of joint fully connected layers and another two fully connected layers for each of the required heads. The loss function used for the stochastic gradient descent (SGD) optimization step followed the one suggested in [30], meaning it contained three terms, the policy loss measured using cross-entropy, the value loss measured using mean-square-error (MSE) and a regularization term on the parameter's size. The main difference in our case was the adjustment of the value loss component to include not only the value but also the value components. Thus, the value was half influenced by the loss of the value, and half of it by the loss of the value components.

Appendix C

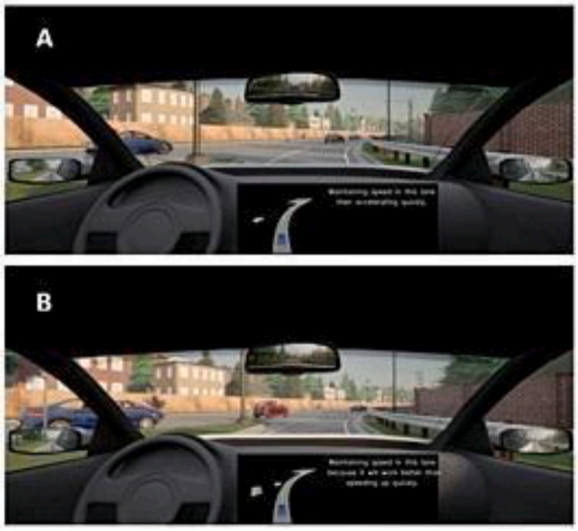
Questions:

1. To what extent do you agree or disagree with the following statement?
   "I trust vehicles that enable automated driving features (such as ACC or Adaptive Cruise Control and future technologies). I will engage in automated driving once my vehicle has such a feature."
   Answer options
   1. Strongly disagree
   2. Disagree
   3. Neither agree nor disagree
   4. Agree
   5. Strongly agree
2. Please read:
   Instruction text
   We will now move onto the main evaluation. The purpose of this assessment is to understand the experience drivers have with automated vehicle technology.
   Throughout this section of the survey, you will be shown videos of a vehicle performing different driving maneuvers automatically, all of which are safe to perform.
   We ask you to imagine yourself as the driver of the vehicle. During the time the maneuver is being shown, you are not able to manually control the vehicle; your brakes, speed, and steering control are automated.
   Our aim is to learn about how you perceive these maneuvers to better understand what causes concern, confusion, and trust in automated driving vehicles.
3. Please read:
   Instruction text
   While watching each video, you will notice a window in the bottom center of the screen where the infotainment panel would be in a vehicle. This window displays what the vehicle is observing, the lane it is driving in, and the vehicle's actions.
   When answering the questions that follow each video, please do not spend a lot of time thinking about your ratings. Your first reactions are usually best.
   Please consider the content of the text that is presented to you on the display, and not if this is an effective way of presenting the information.
   Please note: the videos do not have sound.
   As a reminder, all driving maneuvers are safe.
4. [Practice Assessment, text skipped here, questions are similar to the following questions]

5. [Main Assessment, questions presented below were repeated for each one of the three scenarios. Here we write these questions only once.]
   Please read:
   Instruction text
   The main evaluation will begin on the next page.
   As a reminder, when answering the questions that follow each video, please do not spend a lot of time thinking about your ratings. Your first reactions are usually best.
   Please note: the videos do not have sound.
   All driving maneuvers are safe.
6. Please read the below information before proceeding to the video:
   Instruction text
   In the upcoming video, [we are requesting the vehicle automatically merges onto a highway from the on ramp] [we are requesting the vehicle continues to drive straight] [we are requesting the vehicle automatically follows the curve in the road.]
7. Please watch the entire video carefully.
   Note: It is recommended to open the video in **full screen**, by clicking on the middle square symbol in the bottom right hand corner, to ensure the best viewing experience.
   [Video]
8. Thinking about the video you just watched, to what extent do you agree or disagree with the following statement:
   "I have a better understanding of why the vehicle performed the maneuver shown in the scenario."
   Answer options
   1. Strongly disagree
   2. Disagree
   3. Neither agree nor disagree
   4. Agree
   5. Strongly agree
9. Please indicate your level of comfort with the vehicle during the automated drive you just watched.
   Answer options
   1. Extremely uncomfortable
   2. Somewhat uncomfortable
   3. Neither comfortable nor uncomfortable
   4. Somewhat comfortable
   5. Extremely comfortable
10. Please indicate which of the statements below best fits your level of trust in the vehicle during the automated drive you just watched.
    Answer options
    1. Strongly distrust
    2. Somewhat distrust
    3. Neither trust nor distrust
    4. Somewhat trust
    5. Strongly trust

11. How much did the information displayed in the video add to the feeling of safe driving?
    Answer options
    1. Not at all
    2. Not very much
    3. Somewhat
    4. Very much
    5. Extremely
12. [Text presented to the HMI+Text group] Thinking back to the various automated driving maneuvers you just watched, which of the following statements best describes how you feel about the information, in addition to the traffic display that was provided in the videos? [Text presented to the HMI only group: Thinking back to the various automated driving maneuvers you just watched, which of the following statements best describes how you feel about the traffic display that was provided in the videos?]
    Answer options
    1. Mandatory to help me trust the vehicle
    2. Nice to have, helps me trust the vehicle but not necessary
    3. Indifferent, not necessary and may or may not help me trust the vehicle
    4. Not necessary, doesn't help me trust the vehicle
13. Is there anything different the vehicle systems could do or show you that would help you feel more comfortable or trusting of a vehicle performing automated driving maneuvers? Please explain.

14. Please read:
    Instruction text
    We have one final video for you to watch.
    Please read the below information before proceeding to the video:
    In the upcoming video, we are requesting the vehicle automatically merges onto a highway from the on ramp.
15. Please watch the entire video carefully.
    Note: It is recommended to open the video in full screen, by clicking on the middle square symbol in the bottom right hand corner, to ensure the best viewing experience.
    [Video]
16. Would you consider taking manual control of the vehicle during the automated driving maneuver you just watched?
    Answer options
    1. Yes
    2. Maybe
    3. No
17. [Show if answer to Question 16 was yes/maybe] If the information shown in the images below would have been provided to you during the driving maneuver you just watched, which statement best reflects how you would have reacted?



    Answer options
    1. The information in picture A, *"Maintaining speed in this lane then accelerating quickly"*, would have helped me avoid taking manual control of the vehicle
    2. The information in picture B, *"Maintaining speed in this lane because it will work better than speeding up quickly"*, would have helped me avoid taking manual control of the vehicle
    3. Both A and B would have helped me avoid taking manual control of the vehicle
    4. I would still take manual control of the vehicle during that automated driving maneuver
18. Thinking back to the various automated driving maneuvers you just watched, what information would best help you **avoid** taking manual control of the vehicle?
19. Throughout this survey, were you able to clearly watch each video, without any buffering or downloading issues?
    Answer options
    1. Yes
    2. No
    3. Not sure

Appendix D

To evaluate the XAI algorithms proposed in this work, we performed a 2×3 between-subjects on-line study. All participants experienced all conditions (3 driving scenarios movies). Participants were assigned to two groups based on the information that was provided to them in addition to the dynamic HMI display: XAI notifications or none. The study included the recruiting of 800 participants (400 females and 400 males) between 30 years old and 89 years old (average 57.1 years old, standard deviation 12.8). These participants were also balanced across two age groups. The first group included 400 participants of age 30–59 years old (referred to as the *young group* with an average age of 46.5 years old with a standard deviation of 8.32), and the second group included the other 400 participants of age 60–89 years old (referred to as the *old group* with an average age of 67.8 years old with a standard deviation of 5.57).

H1: Age, gender, and ADAS experience affect the levels of subjective understanding and trust of users, when provided with XAI

explanations in addition to the dynamic HMI display.

We showed, through Cronbach's Alpha analysis, that the questions regarding Trust and Subjective Understanding over the three movies presented to all participants were consistent. Here, we include the full details of the ANOVA F analysis performed on these results to show their significance.

For the direct Trust questions, we had 1200 elements for each group (each one of the 400 participants watched all 3 scenarios, n=1200). We had two groups (m=2). The mean values were 2.69 for the HMI+text group and 2.78 for the group of HMI only. The two populations tested were found to be statistically significantly distinguishable with respect to the Trust questions ($F=3.35617$, $p < 0.1$): the Sum-of-Squares Between (SSB) divided by degrees of freedom (df) of 1 was 4.95, while Sum-of-Squares Within (SSW) divided by df of 2398 (i.e., m(n-1)) was 1.475. Thus, F-statistics $= 3.35617 >$ critical F value for alpha $= 0.1$ with the corresponding df $\cong 2.7$. Therefore, we can conclude that even though the difference in mean seems very small the two populations are indeed distinguishable with respect to the Trust questions.

On the other hand, when examining the Subjective Understanding question as a whole, with the same df and means of 3.3475 for HMI+text and 3.29 for HMI only, we got no statistical significance ($F=1.6684$, $p>0.1$). In this case, the Sum-of-Squares Between, SSB divided by df $= 1.9267$, SSW divided by df $= 1.154$ and thus, F-statistics $= 1.6694 <$ critical F value for alpha $= 0.1$ with the corresponding df $\cong 2.7$. In the case of the Subjective Understanding question, we cannot conclude that the two populations are distinguishable.

For the question whether a person would have taken control or not, we had 800 responses and three optional answers: 1 – Yes. Will take manual control. 2 – Maybe. 3 – No. Will not take manual control. We have two groups (those with XAI and those with only the HMI display, so m=2). The mean values were 1.375 for the group with XAI (experienced XAI in the previous questions), and 1.4525 for those that did not receive any XAI notifications. In this case, the SSB divided by df of 1 was 1.2, while the SSW divided by df of 798 was 0.382, thus: F-statistics $= 3.1414 >$ critical F value for alpha $= 0.1$ with the corresponding df $\cong 2.7$. Also, in this case we observe that there is a significant difference between the two groups with respect to the Taking Control question ($F=3.1414$, $p<0.1$).

We observe that ADAS experience is a leading variable of impact on all three dependent variables: Trust, Understanding and Taking Control (Fig. 14 and Fig. 15). As further support of the importance of ADAS experience and its effect on the outcome of the dependent variables, we conducted an ANOVA F analysis considering only this factor. With respect to Trust, the results indicate statistical significance ($F=138.882$, $p<0.1$); the SSB/df_SSB (df_SSB_ $= 1$) was 138.07 while SSW/df_SSW (df_SSW $= 3424$) was 0.99, thus: F-statistics $= 138.88215218526435$; meaning there is a distinguishable difference between the two groups, ADAS and No ADAS experience with respect to the Trust measure.

Similarly, for the Subjective Understanding ($F=91.227$, $p<0.1$); we received SSB/df_SSB $= 71.87$, and SSW/df_SSW=0.7878 (same df values), thus: F-statistics $= 91.22738959162265$, a distinguishable difference between the two groups with respect to the Subjective Understanding measure.

To emphasize the validity of these results and their significance, we also performed the ANOVA F analysis over the Taking Control measure, considering the ADAS and no ADAS populations. In this case we found no statistical significance ($F=0.32$, $p>0.1$) received SSB/df_SSB $= 0.0856$ (df_SSB=1) and SSW/df_SSW=0.2684 (df_SSW=1140), thus, F-statistics =0.32; meaning that in this case we find that the two groups are not distinguishable.

Data availability

The authors do not have permission to share data.

References

[1] A. Agogino, R. Lee, & Giannakopoulou, D. (2019). Challenges of explaining control. Proceedings of 2nd ICAPS Workshop on Explainable Planning (XAIP-2019). Berkeley, CA. USA.
[2] M. Akamatsu, P. Green, K. Bengler, Automotive Technology and Human factors Research: past, present and future, Int. J. Veh. Technol. (2013), https://doi.org/10.1155/2013/526180.
[3] M.S. Alam, & Y. Xie (2022). Appley: approximate Shapley value for model explainability in linear time. Proceedings of the IEEE International Conference on Big Data, (pp. 95–100). Osaka, Japan.
[4] A. Anderson, J. Dodge, A. Sadarangani, Z. Juozapaitis, E. Newman, J. Irvine, M. Burnett (2019). Explaining reinforcement learning to mere mortals: an empirical study. Proceedings of the 28th International Joint Conference on Artificial Intelligence.
[5] M. Beggiato, F. Hartwich, K. Schleinitz, J.F. Krems, & I. Othersen (2015). What would drivers like to know during automated driving? Information needs at different levels of automation. Proceedings of the 7th Conference on Driver Assistance. Munich, Germany. doi:10.13140/RG.2.1.2462.6007.
[6] C. Browne, E. Powley, D. Whitehouse, S. Lucas, P.I. Cowling, P. Rohlfshagen, S. Colton, A survey of Monte Carlo tree search methods, IEEe Trans. Comput. Intell. AI. Games. 4 (1) (2012).
[7] N. Bussmann, P. Giudici, D. Marinelli, J. Papenbr, Explainable machine learning in credit risk management, Comput. Econ. 57 (2021) 203–216.
[8] R. Bustin, & C.V. Goldman (2024). Structure and reduction of MCTS for explainable AI. Proceedings of the 27th European Conference on Artificial Intelligence. Santiago de Compostela, Spain.
[9] T. Chakraborti, S. Sreedharan, & S. Kambhampati (2020). The emerging landscape of explainable automated planning & decision making. Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), (pp. 4803–4811). Yokohama, Japan.
[10] C.C. Chang, R.A. Grier, J. Maynard, J. Shutko, M. Blommer, R. Swaminathan, R. Curry, Using situational awareness display to improve rider trust and comfort with an av taxi, Human Factors Ergon. Soc. Ann. Meet. 63 (1) (2019) 2083–2087.
[11] F. Chen (2020). Special issue on HMI and autonomous driving. Automotive Innovation. doi:10.1007/s42154-020-00094-1.

[12] H. Chen, I.C. Covert, S.M. Lundberg, et al., Algorithms to estimate Shapley value feature attributions, Nat. Mach. Intell. 5 (2023) 590–601, https://doi.org/10.1038/s42256-023-00657-x.
[13] J.Y. Chen, S.G. Lakhmani, K. Stowers, A.R. Selkowi, Situation awareness-based agent transparency and human-autonomy teaming effectiveness, Theor. Issues. Ergon. Sci. 19 (3) (2018) 259–282.
[14] J. Chen, E. Li, M. Tomizuka, Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning, IEEE Trans. Intell. Transport. Syst. 99 (2021) 1–11, https://doi.org/10.1109/TITS.2020.3046646.
[15] J. Craig, & M. Nojoumian (2021). Should self-driving cars mimic Human driving behaviors? In H. Krömker (Ed.), Lecture Notes in Computer Science:HCI in Mobility, Transport, and Automotive Systems (Vol. 12791). Springer, Cham. doi:10.1007/978-3-030-78358-7_14.
[16] I. Csiszar, The Method of Types, IEEe Trans. Inf. Theory. 44 (1998) 2505–2523.
[17] F. Doshi-Velez, & et al. (2017). Towards A rigorous science of interpretable machine learning. Retrieved from https://arxiv.org/abs/1702.08608.
[18] A.D. Dragan, K.C. Lee, & S.S. Srinivasa (2013). Legibility and predictability of robot motion. Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), (pp. 301–308).
[19] F. Elizalde, E. Sucar, J. Noguez, A. Reyes, Generating explanations based on Markov decision processes, in: Proceedings of the Mexican International Conference on Artificial Intelligence, Springer, 2009, pp. 51–62.
[20] M. Finkelstein, L. Liu, Y. Kolumbus, D.C. Parkes, J.S. Rosenschein, S. Keren, et al., Explainable reinforcement learning via model transforms, Adv. Neural Inf. Process. Syst. 35 (2022) 34039–34051.
[21] M. Fox, D. Long, & D. Magazzeni (2017). Explainable planning. XAI workshop at IJCAI. Melbourne, Australia. Retrieved from https://arxiv.org/abs/1709.10256.
[22] J. Fritz (2024, February). Wireless technology fuels industry growth and innovation. Retrieved from https://www2.deloitte.com/us/en/pages/technology-media-and-telecommunications/articles/wireless-technology-fuels-innovation-in-key-industries.html.
[23] B.C. Gentile, S. Grabe, B.J. Dolan-Pascoe, J.M. Twenge, B.E. Wells, A. Maitino, Gender differences in domain-specific self-esteem: a meta-analysis, Rev. Gen. Psychol. 13 (2009) 34–45, https://doi.org/10.1037/a0013689.
[24] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M.A. Specter, & L. Kagal (2018). Explaining explanations: an approach to evaluating interpretability of machine learning. Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics, (pp. 80–89). Turin, Italy. doi:10.1109/DSAA.2018.00018.
[25] C.V. Goldman, & M. Baltaxe (2021). Why are you predicting this class? Proceedings of the 32nd IEEE Intelligent Vehicle Symposium. Virtual.
[26] C.V. Goldman, & R. Bustin (2022). Trusting explainable autonomous driving: simulated studies. Proceedings of the IEEE Intelligent Vehicles Symposium. Aachen, Germany.
[27] C.V. Goldman, M. Baltaxe, D. Chakraborty, C.A. Escobar, J. Arinez, Interpreting learning models in manufacturing processes: towards explainable AI methods to improve trust in classifier predictions, J. Ind. Inf. Integr. 33 (2023) 100439 doi.
[28] A. Gross (2022, May). Consumer skepticism toward autonomous driving features justified. Retrieved from https://newsroom.aaa.com/2022/05/consumer-skepticism-toward-active-driving-features-justified/.
[29] B. Hayes, & J.A. Shah (2017). Improving robot controller transparency through autonomous policy explanation. Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), (pp. 303–312). Viena, Austria. doi:10.1007/978-3-030-7772.
[30] C. Hoel, K. Driggs-Campbell, K. Wolff, L. Laine, M.J. Kochenderfer, Combining planning and deep reinforcement learning in tactical decision making for autonomous driving, IEEE Trans. Intell. Veh. 5 (2) (2020).
[31] B.E. Holthausen, R.E. Stuck, & B.N. Walker (2022). Trust in automated vehicles. Studies in computational intelligence user experience design in the era of automated driving, 980, 29–49. doi:10.1007/978-3-030-7772.
[32] T. Huber, K. Weitz, E. Andŕe, O. Amir, Local and global explanations of agent behavior: integrating strategy summaries with saliency maps, Artif. Intell. (2021) 301, https://doi.org/10.1016/j.artint.2021.103571.
[33] Info Tech Research Group. (2024). Tech Trends. Retrieved from https://www.infotech.com/research/ss/tech-trends-2024.
[34] A. Jacovi, A. Marasovĭc, T. Miller, & Y. Goldberg (2021). Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, (pp. 624–635). Virtual. doi:10.1145/3442188.3445923.
[35] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, & F. Doshi-Velez (2019). Explainable reinforcement learning via reward decomposition. Proceedings of the International Joint Conference on Artificial Intelligence, A Workshop on Explainable Artificial Intelligence.
[36] D. Kahneman (2013). Thinking, fast and slow. Farrar, Straus and Giroux.
[37] O.Z. Khan, P. Poupart, & J.P. Black (2009). Minimal sufficient explanations for factored Markov decision processes. Proceedings of the 19th International Conference on Automated Planning and Scheduling, (pp. 194–200). Thessaloniki, Greece. doi:10.1609/icaps.v19i1.13365.
[38] S. Khastgir, S. Birrell, G. Dhady, P. Jennings, Calibrating trust through knowledge: introducing the concept of informed safety for automation in vehicles, Transport. Res. Part C Emerg. Technol. 96 (2018) 290–303, https://doi.org/10.1016/j.trc.2018.07.001.
[39] J. Kim, & J. Canny (2017). Interpretable learning for self-driving cars by visualizing causal attention. Proceedings of the IEEE international conference on computer vision, (p. 2961). doi:10.1109/ICCV.2017.320.
[40] S.C. Kohn, E.J. de Visse, E. Wiese, Y.C. Lee, T.H. Shaw, Measurement of trust in automation: A narrative review and reference guide, Front. Psychol. 12 (604977) (2021).
[41] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, C. Nass, Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust and performance, Int. J. Interactive Design Manuf. 9 (4) (2015) 269–275.
[42] M. Körber (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), Proceedings of the 20th Congress of the International Ergonomics.
[43] A. Kraft, C. Maag, M. Baumann, Comparing dynamic and static illustration of an HMI for cooperative driving, Accid. Anal. Prevent. (2020) 144, https://doi.org/10.1016/j.aap.2020.105682.
[44] A. Kulkarni, Y. Zha, T. Chakraborti, S.G. Vadlamudi, Y. Zhang, & S. Kambhampati (2019). Explicable planning as minimizing distance from expected behavior. Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Montreal, Canada.
[45] J.D. Lee, N. Moray, Trust, self-confidence, and operators' adaptation to automation, Int. J. Human Comput. Stud. 40 (1) (1994) 153–184.
[46] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, Hum. Factors 46 (1) (2004) 50–80.
[47] B.Y. Lim, A.K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2009, pp. 2119–2128.
[48] Z. Lin, K. Lam, & A. Fern (2021). Contrastive explanations for reinforcement learning via embedded self predictions. Retrieved from https://arxiv.org/abs/2010.05180.
[49] A. Lucic, H. Haned, & M. de Rijke (2020). Why does my model fail? Contrastive local explanations for retail forecasting. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, (pp. 90–98). Barcelona, Spain.
[50] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, S.I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67.
[51] P. Madumal, T. Miller, L. Sonenberg, & F. Vetere (2020). Explainable reinforcement learning through a causal lens. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, NY, USA. doi:10.1609/aaai.v34i03.5631.
[52] McKinsey. (2023). Retrieved from McKinsey Center for Future Mobility: https://www.mckinsey.com/features/mckinsey-center-for-future-mobility/overview.
[53] S. Milani, N. Topin, M. Veloso, F. Fang, Explainable Reinforcement Learning: a survey and comparative review, ACM. Comput. Surv. 56 (7) (2024), https://doi.org/10.1145/3616864.
[54] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[55] T. Miller, Contrastive explanation: a structural-model approach, Knowl. Eng. Rev. (2021) 36, https://doi.org/10.1017/s0269888921000102.

[56] B. Mittelstadt, C. Russell, & S. Wachter (2019). Explaining explanations in AI. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, (pp. 279–288). Atlanta, GA, USA. doi:10.1145/3287560.3287574.
[57] B. Moye (2023, March). Fear of self-driving cars on the rise. Retrieved from AAA: https://newsroom.aaa.com/2023/03/aaa-fear-of-self-driving-cars-on-the-rise/ .
[58] S.B. Nashed, S. Mahmud, C.V. Goldman, & S. Zilberstein (2023). Causal explanations for sequential decision making under uncertainty. Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS), (pp. 2307–2309). London, UK.
[59] O. O. Schilke, M. Reimann, K.S Cook, Trust in social relations Annu Rev. Sociol. 47 (2021) 239–259.
[60] R. Parasuraman, V. Riley, Humans and automation: use, misuse, disuse, abuse, Hum. Factors 39 (2) (1997) 230–253.
[61] J. Petch, S. Di, W. Nelson, Opening the black box: the promise and limitations of explainable machine learning in cardiology, Can. J. Cardiol. 38 (2) (2022) 204–213.
[62] N. Phongphaew, A. Jiamsanguanwong, Text-based information design for in-vehicle displays: a systematic review, Transport. Res. Part F Traffic Psychol. Behav. 103 (2024) 442–459, https://doi.org/10.1016/j.trf.2024.04.025.
[63] M. Ribeiro, et al. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 1135–1144). San Francisco, CA, USA.
[64] F. Rietz, S. Magg, F. Heintz, et al., Hierarchical goals contextualize local reward decomposition explanations, Neural Comput. Appl. 35 (2023) 16693–16704, https://doi.org/10.1007/s00521-022-07280-8.
[65] R. Roscher, B. Bohn, M.F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, IEEe Access. 8 (2020) 42200–42216.
[66] E. Rovira, K. McGarry, R. Parauraman, Effects of imperfect automation on decision making in a simulated command and control task, Hum. Factors 49 (1) (2007) 76–87.
[67] B. Rozemberczki, L. Watson, P. Bayer, H.T. Yang, O. Kiss, S. Nilsson, & R. Sarkar (2022). The Shapley value in machine learning. Retrieved from https://arxiv.org/abs/2202.05594.
[68] J. Russell, & E. Santos (2019). Explaining reward functions in Markov decision processes. Proceedings of the 32nd International Florida AI Research Society Conference (FLAIRS).
[69] S.J. Russell, & P. Norvig (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
[70] S. Beyer, E.M. Bowden, Gender differences in self-perceptions: convergent evidence from three measures of accuracy and bias, Pers, Soc. Psychol. Bull. 23 (1997) 157–172, https://doi.org/10.1177/0146167297232005.
[71] SAE. (2021, May). SAE levels of driving automation. Retrieved from https://www.sae.org/blog/sae-j3016-update.
[72] A. Saranya, R. Subhashini, A systematic review of explainable artificial Intelligence models and applications: recent developments and future trends, Decis. Anal. J. (2023) 7, https://doi.org/10.1016/j.dajour.2023.
[73] B.D. Seppelt, J.D. Lee, Keeping the driver in the loop: dynamics feedback to support appropriate use of imperfect vehicle control automation, Int. J. Human Comput. Stud. 125 (2019) 66–80.
[74] L.S. Shapley (1953). A value for n-person games. In H. W. Kuhn, & A. W. Tucker (Eds.), Contributions to the Theory of Games (pp. 307–317). Princeton University Press.
[75] D. Silver, J. Schrittwieser, K. Simonyan, et al., Mastering the game of Go without human knowledge, Nature 550 (2017) 354–359, https://doi.org/10.1038/nature24270.
[76] J. Song, X. Song, J. Bai, et al., The influence of gender driving stereotype threat on the driving performance of female drivers and its mechanism, Curr. Psychol. 43 (2024) 17213–17224, https://doi.org/10.1007/s12144-024-05695-1.
[77] S. Sreedharan, T. Chakraborti, S. Kambhampati, Foundations of explanations as model reconciliation, Artif. Intell. 301 (103558) (2021), https://doi.org/10.1016/j.artint.2021.103558.
[78] K. Stubbs, P.J. Hinds, D. Wettergreen, Autonomy and common ground in human-robot interaction: a field study, IEEe Intell. Syst. 22 (2) (2007) 42–50.
[79] R.S. Sutton, & A.G. Barto (2018). Reinforcement Learning: An Introduction (Second ed.). MIT Press.
[80] J. van der Waa, J. van Diggelen, K. van den Bosch, & M. Neerincx (2018). Contrastive explanations for reinforcement learning in terms of expected consequences. Proceedings of the first workhop on XAI at IJCAI. Stockholm, Sweden.
[81] N. Wang, D.V. Pynadath, S.G. Hill, The impact of POMDP-generated explanations on trust and performance in human-robot teams, in: Proceedings of the International Conference on Autonomous Agents & Multiagent Systems, International Foundation for Autonomous Agents and Multiagent Systems, Singapore, 2016, pp. 997–1005.
[82] W. Wang, I. Benbasat, Recommendation agents for electronic commerce: effects of explanation facilities on trusting beliefs, J. Manage. Info. Syst. 23 (4) (2007) 217–246.
[83] H.M. Wojton, D. Porter, S.T. Lane, C. Bieber, P. Madhavan, Initial validation of the trust of automated systems test (TOAST), J. Soc. Psychol. 160 (2020) 735–750, https://doi.org/10.1080/00224545.2020.1749020.
[84] H. Yau, C. Russell, & S. Hadfield (2020). What did you think would happen? Explaining agent behaviour through intended outcomes. Retrieved from https://arxiv.org/abs/2011.05064.
[85] Y. Zhang, Q.V. Liao, & R.K. Bellamy (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, (pp. 295–305). doi:10.1145/335.