The Design and Preliminary Results of a User Study Measuring Diverse Explainability Preferences

Pradyumna Tambwekar
Georgia Institute of Technology
Atlanta, Georgia, USA
pradyumna.tambwekar@gatech.edu

Andrew Silva
Georgia Institute of Technology
Atlanta, Georgia, USA
andrew.silva@gatech.edu

Matthew Gombolay
Georgia Institute of Technology
Atlanta, Georgia, USA
matthew.gombolay@cc.gatech.edu

ABSTRACT

As robots and digital assistants are deployed in the real world, these agents must be able to communicate their decision-making criteria to build trust, improve human-robot teaming, and enable collaboration. While the field of explainable artificial intelligence has made great strides in building a set of mechanisms to enable such communication, these advancements often assume that one approach is ideally suited to one or more problems (e.g., decision trees are best for explaining how to triage patients in an emergency room), failing to recognize that individual users may have different past experiences or preferences for interaction modalities. In this work, we present the design and results of a user study in a virtual self-driving car domain, in which the car presents navigational assistance to the human and uses varying explanation modalities to justify its suggestions. We find significant differences between explanation baselines for subjective ranking preferences ($p < 0.01$) and objective performance with respect to incorrect compliance ($p < 0.05$). However, we find that some participants have strong preferences that go against our population-level findings, which makes suggesting the majority-preference an inappropriate solution. Our analysis shows that personalization is crucial to maximize the subjective and objective benefits of explanations with diverse users.

CCS CONCEPTS

• **Information systems** → **Personalization**; • **Human-centered computing** → **User studies**.

KEYWORDS

explainability, user studies, personalization

ACM Reference Format:
Pradyumna Tambwekar, Andrew Silva, and Matthew Gombolay. 2023. The Design and Preliminary Results of a User Study Measuring Diverse Explainability Preferences. In *LEAP-HRI*. ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

As robots and digital assistants are deployed to the real world, these agents must be able to communicate their decision-making criteria to build trust, improve human-robot teaming, and enable collaboration [2, 15]. Researchers have identified *explainability* as a necessary component of high-quality human-robot interactions in many domains [5, 18]. While numerous approaches to explainability are under active investigation (e.g., natural language explanations [4], decision-tree extraction [21], counterfactual presentation [12], or saliency-based explanations [17, 23]), we hypothesize that some explainability mechanisms may be better suited to certain problems and individuals than others in contrast to the "one-size-fits-all" approach of current research. Individual dispositional preferences and expertise can have an impact on the success of an explanation both in terms of subjective satisfaction and objective utility [24].

Explainable AI is not a domain wherein the "accuracy" can simply be measured by the explanations interpretability of the underlying algorithm. If explanations do not carefully consider an individual's human expertise or expectations, the simple act of showing an explanation can cause humans to blindly trust an agent's advice, leading to adverse effects on performance and trust[16, 22]. While explainability may seem to be a useful tool to deploy alongside an imperfect decision-making aid, this counter-intuitive result presents a key problem: explanations encourage inappropriate compliance. If some users see explanations and defer to robots without critically examining the explanation or robot suggestion, a natural follow-up question is: Can we understand the relationship between a user's preferences for usability of an explanation with their compliance? By attaining an understanding of the factors which influence inapropriate complaince, we can counter the potential elicitation of unwarranted trust in a system [6, 8, 19].

We aim to understand the diverse preferences of untrained human users with potentially-faulty assistants that use explanations. In our study, participants interact with a virtual self-driving car to navigate through a foreign environment, where an agent provides navigational assistance to help guide the user, reflecting a potentially common future use case. Crucially, this advice is not always correct, and incorrect advice is signalled by the inclusion of red-herring features (e.g., we designed the study such that using "weather" in an explanation signals an error). Users must decide whether to accept or reject the advice by examining the agent's explanations, which hold clues as to whether or not the agent is offering incorrect advice. Our results demonstrate that, while there are significantly different preferences for explanation modality at a population level ($p < 0.01$), individual users were not always reflective of this trend. Through our novel study, we show that personalization in the context of explainability with human users is crucial by showing that preferences at an individual or group level vary from global population differences.

2 STUDY DESIGN

In this section, we detail the design of our study, set in the domain of navigating through an unknown city with the assistance of a virtual XAI agent. A visual overview of the study is in Figure 1.

LEAP-HRI '23, 2023,
2023.

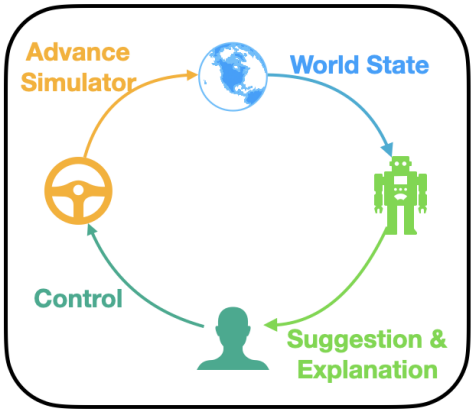


Figure 1: A visual walkthrough of our study cycle, (entry and exit surveys omitted for simplicity). As the agent provides suggestions and explanations to participants, participants continue to direct the virtual car based on navigation suggestions and explanations. Participants complete eleven navigation tasks in each study session before completing post-study surveys and providing qualitative feedback.

2.1 Navigation Tasks

Our study consists of eleven navigation tasks, in which the participant must navigate through the city to a goal location by directing a self-driving agent through intersections in a city. We build our study in the AirSim simulator [20], shown in Figure 3. For each task, the user is moved to a new starting location, and the objective is moved to new goal location, mitigating any learning effects or ability to memorize the same routes. Furthermore, the environment contains obstacles that force the car to turn around, obstructing navigation and constraining possible paths. This encourages reliance on the agent's navigational suggestions and discourages participants from attempting to self-navigate, as the agent knows the positions of all obstacles, but participants do not.

The environment is a set of city blocks joined by four- and three-way intersections. At each intersection, the car comes to a full stop and asks the participant to select the next direction (straight, right, or left). The participant is shown a directional suggestion from the navigation agent, as well as an explanation for the given suggestion. Finally, the participant also has access to a mini-map that shows their position and heading, as well as the goal position. This helps participants to feel better oriented in the map, and gives them the ability to navigate independently. However, as they do not know the positions of obstacles scattered throughout the city, participants are still heavily reliant on the agent's assistance to reach the objective in an optimal number of turns.

For all but the first task, 30% of explanations and suggestions are incorrect. Suggestions in the first task are all correct to allow participants time to become accustomed to the domain and the agent. Each task in the main body of the study is completed with only one explainability mechanism, and we rotate which mechanism is used after each task. Tasks have a fixed number of incorrect explanations, which are shown at random times. Participants are warned that their assistant will occasionally make mistakes during the navigation tasks, and that mistakes are signaled by the inclusion of a "red-herring" feature, including things such as the "time of day, radio station, rush hour traffic," and several other features (e.g., "We should turn left because it's near noon."). All explanations are scripted via Wizard-of-Oz, allowing us to control for differing levels of sophistication in state-of-the-art explainability research.

2.2 Explanation Modalities

In our study, compare explanations using three different mechanisms, that have been widely utilized in prior work on explainable AI [7, 11, 15]:

(1) Natural Langugage: Explanations are offered in natural language, describing the decision and the justification.
(2) Decision Tree: A decision tree describing the car's logic, with relevant nodes highlighted.
(3) Salience Map: A feature-importance heatmap, highlighting objects in the scene that may be relevant for a decision.

2.3 Research Questions and Metrics

Our primary research questions in this study are:

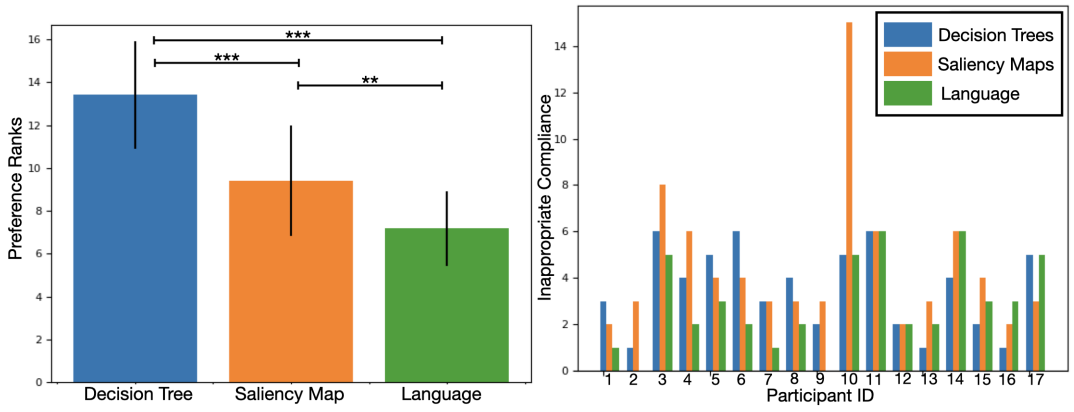
- **RQ1 – Preferences**: Will one explanation modality be significantly more preferred than the others?
- **RQ2 – Performance**: Will one explanation modality lead to significantly better performance than the others?
- **RQ3 – Alignment**: Will participants prefer to use the modality that gives them the best performance?

Finally, for all of our research questions, we are interested in both population and individual-level data. Our study will signal a need for personalization if we show that many individuals in our population do not adhere to the trends found across the broader population (i.e., we cannot simply apply the significant majority preference if it means that many users will be left behind).

We propose to measure **XAI modality rankings (RQ1)**, participant rankings of explainability mechanisms (where rank 1 is best and rank 3 is worst), and **inappropriate compliance (RQ2)**, the number of times that participants accept faulty advice.

2.4 Study Timeline

Upon arrival to the onsite location, participants complete consent forms and are briefed on their task. Participants are introduced to each of the explanation mechanisms employed in the study, the interface for directing the car, and a mini-map that will assist them for each task. After completing NARS[14] and "Big-Five" personality[13] surveys to control for the effects of personality or comfort with robots on our results and being briefed on the study tasks, participants begin the eleven navigation tasks in our study. The first two tasks are both practice, allowing the users to become acquainted with the simulator, controls, and explanations. We arrived at two tasks after pilot studies revealed that very little experience was required to learn the navigation task. In this practice phase, explanations are randomly sampled from any of the three mechanisms used in our study (Section 2.2). After completing the practice phase, participants begin the main body of the study, which consists of 9 tasks. We then conclude the study by administering two final surveys to our participants. First, we have participants



(a) Preference Rank Results

(b) Inappropriate Compliance Results for All Participants

Figure 2: Visualized results for preference rankings (a) and objective performance metrics (b). Lower is better for both. Preference ranks are the summation over five questions, in which modalities are ranked one to three (i.e., scoring five is the best possible rank, scoring fifteen is the worst possible rank).

rank their preferred explanation modality according to which was their favorite, which they felt was their most/least preferred, which was the fastest, the slowest, the easiest to use, and the hardest to use. Finally, participants complete a Likert scale measuring their desire to personalize explanations (Section 6).

2.5 Sample Population

We gathered data from 17 participants (14 Males, 3 Females) between 18-35. Our participants reported a small to medium degree of experience with computer science.

3 RESULTS

We performed a repeated-measures multivariate analysis to compute the effect of each condition on explanation modality ranking and on inappropriate compliance. The explanation modality was modelled as a fixed-effect covariate, and the participant id was a random effects covariate. We utilized the AIC metric to determine which covariates to include in our linear model. We then applied an analysis of variance (ANOVA) to identify significance across baselines, and further employed a Tukey-HSD post-hoc test to measure pairwise significance. For our linear regression model, we tested for the normality of residuals and homoscedasticity assumptions and found that the assumptions all passed for the explanation ranking model; however, the residuals of the inappropriate compliance model were not found to be normally distributed. In prior work, it has been shown that an F-test is robust to non-normality [1, 3, 9, 10]. Therefore, we choose to proceed with a linear regression analysis.

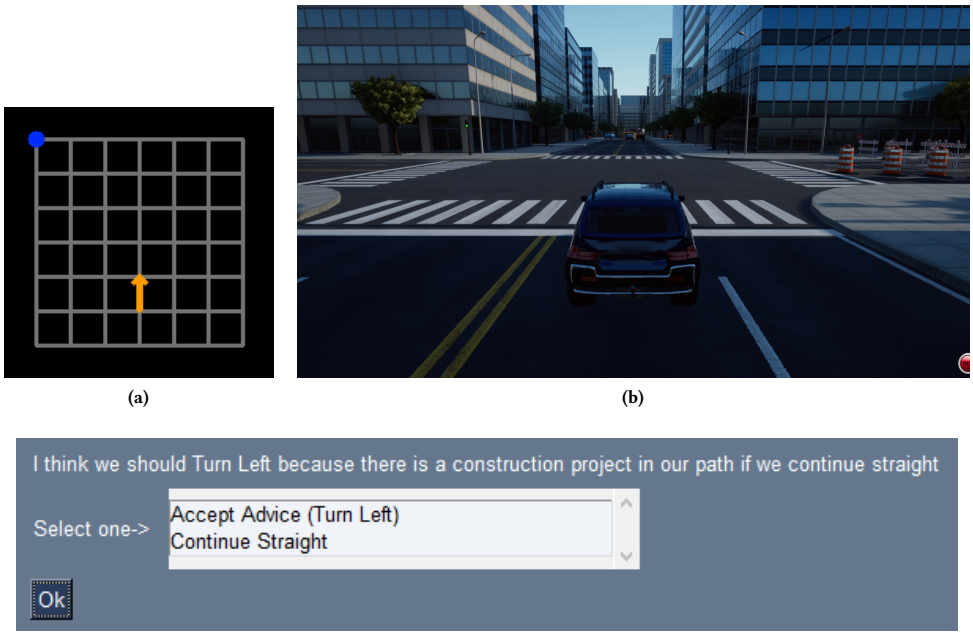
Firstly, our ANOVA for explanation modality rankings yielded a significant difference across baselines ($F(2, 45) = 27.994, p < 0.001$). A Tukey post-hoc test showed a pairwise difference between each pair of explanation modalities included in this study. We observed that language ranked significantly higher than saliency maps ($p < 0.01$) and decision trees ($p < 0.001$), and saliency maps were ranked significantly higher than decision trees ($p < 0.001$). A comparison of means of all preference rankings is shown in Figure 2a.

An ANOVA on inappropriate compliance yielded significant difference across conditions ($F(2, 32) = 4.3593, p < 0.05$). A Tukey post-hoc revealed that language explanations significantly reduced inappropriate compliance relative to saliency maps ($p < 0.01$). Inappropriate compliance counts for individual participants in the study are given in Figure 2b, where we see that over a quarter of all participants (13-17) exhibit their highest performance with a modality other than language.

4 DISCUSSION

RQ1 – Our results identify significant preferences across our population of participants, i.e. language » saliency » decision tree. However, we find that these significant differences are not always reflected at an individual level. For example, while our population level differences suggest that saliency map explanations are significantly more preferred than decision tree explanations, we find that 20% of our participants report the exact opposite. This finding signals a need for personalization to individual users, as naively assigning the modality with the highest average rankings could lead to increases in innapropriate compliance. Upon completion of our data collection process, we plan to conduct correlation analysis, and other similar analyses, on various sub-groupings of users within the data to identify relevant XAI insights to for these groupings.

RQ2 – Our results for participant performance with each explanation modality are more nuanced than our preference results. We observe a statistically significant difference between language and saliency modalities, showing that language is a superior explanation modality for participants to determine when an explanation may be faulty. However, we observe that over a quarter of our participants show equal or better performance with a modality other than language (as shown by participants 10-17 in Figure 2. In other words, while the majority of our population exhibits superior performance with language, a non-trivial minority shows improved performance with decision trees or saliency maps. Our performance results reinforce the need for personalization because there is no globally superior explanation modality for all participants.



(a) (b) (c)

Figure 3: At each intersection, each participant is shown a mini-map of the city (a), in order to assist them in their decision making. The mini-map provides the location and heading of the car, as well as the location of the goal. Participants select a direction from a pop-up to direct the car. In this example, the pop-up includes a language explanation.

RQ3 – Finally, considering the alignment between participant's preferences and efficiency, we find that there are often discrepancies between a participant's preferred explanation modality and the one that maximizes their task performance, and there are several instances in which a participants preferred modality is the worst for their performance. Again, there are no statistically significant results across the entire population– rather, each individual participant presents a unique combination of prior experiences and preferences that inform their performance and ranking for each explanation modality during the study. Without the ability to personalize to individual's preferences or to intelligently balance between performance and preference, we will not be able to optimize performance or satisfaction in human-robot teams.

Summary – In summary, our results underscore the need for personalization in the context of explainability with human users of autonomous systems. Naively applying the explanation that fits the average user will lead to increases in inappropriate compliance in a non-negligible section of the population of end-users. **Similarly, simply accepting a user's request to use their preferred explanation modality could result in sub-optimal performance, as preferences and performance do not always align.** To maximize a human's efficiency and satisfaction with an XAI system, we must develop agents that rapidly personalize to users, balancing between maximizing efficiency and accommodating preferences.

5 IMPLICATIONS AND FUTURE WORK

Personalization within explainability stands as a key challenge to be solved in order to improve adoption and utilization of XAI systems by stakeholders in the real-world. Unlike prior work, our study explicitly models incorrect compliance. Incorrect compliance in safety-critical domains such as self-driving cars and healthcare, can be extremely problematic, and our proposed study and analysis will provide some useful insights to understand these problems.

Our results suggest that personalization is crucial to the usability of widespread explainable AI, as individuals present personal preferences that are not simply matched to the population average. This finding presents two interesting areas for further exploration: (1) which personalization mechanism to use (e.g., using language instead of saliency maps) and (2) personalizing explanations within one mechanism (e.g., refining the phrasing of a language explanation to conform to a user's expectations).

6 CONCLUSION

As machine learning and robotics are deployed to the real world, it is imperative that the research community maintains an accurate understanding of how such technologies are received and used by human users. We have presented the design and early results of a user study that targets explanation and personalization in machine learning with humans. Our study design enables us to gather data on the unique preferences of individual users presented with the same set of explanation mechanisms, and our results build a strong case for the need for personalized machine learning. We conclude with directions for future work, including research into personalizing what types of explanations are given to different users and research into how we can further refine individual explanations to meet a user's personal knowledge base, expertise, and preferences.

REFERENCES

[1] María José Blanca Mena, Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré, Rebecca Bendayan, et al. 2017. Non-normal data: Is ANOVA still a valid option? *Psicothema* (2017).

[2] Kathleen Boies, John Fiset, and Harjinder Gill. 2015. Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. *The Leadership Quarterly* 26, 6 (2015). https://doi.org/10.1016/j.leaqua.2015.07.007

[3] William G Cochran. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics* 3, 1 (1947), 22–38.

[4] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429* (2019).

[5] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[6] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[7] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[8] Andrea Ferrario and Michele Loi. 2022. How explainability contributes to trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1457–1466.

[9] Gene V Glass, Percy D Peckham, and James R Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research* 42, 3 (1972), 237–288.

[10] HRB Hack. 1958. An empirical investigation into the distribution of the F-ratio in samples from two non-normal populations. *Biometrika* 45, 1/2 (1958), 260–265.

[11] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. 2021. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence* 301 (2021), 103571.

[12] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

[13] Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology* 52, 1 (1987), 81.

[14] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies* 7, 3 (2006), 437–454.

[15] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, and Matthew Gombolay. 2021. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. In *Proceedings of the Conference on Neural Information Processing Systems*.

[16] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*.

[18] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2021. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251* (2021).

[19] Nadine Schlicker and Markus Langer. 2021. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Proceedings of Mensch und Computer 2021*. 325–329.

[20] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*. Springer.

[21] Andrew Silva, Matthew Gombolay, Taylor Killian, Ivan Jimenez, and Sung-Hyun Son. 2020. Optimization Methods for Interpretable Differentiable Decision Trees Applied to Reinforcement Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Silvia Chiappa and Roberto Calandra (Eds.), Vol. 108. PMLR.

[22] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2022. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction* (2022), 1–15.

[23] Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2020. Finding experts in transformer models. *arXiv preprint arXiv:2005.07647* (2020).

[24] Pradyumna Tambwekar and Matthew Gombolay. 2023. Towards Reconciling Usability and Usefulness of Explainable AI Methodologies. *arXiv preprint arXiv:2301.05347* (2023).

EXAMPLES OF EXPLANATION MODALITIES

Here, we show examples of the two visual explanation modalities in our study. In Figure 4, we show a saliency map explanation. Here, the agent highlights various components of the scene, signifying that they are relevant to the decision, and the agent also highlights directions that it may move. In Figure 5, we show the decision tree explanation. The tree is consistent in all explanations, and shows various criteria that may be reflected in the environment. Red nodes indicate that the decision criteria are not met (i.e., the node evaluates to "False"), and the final decision is highlighted in green.

PERSONALIZATION LIKERT SCALE

Here, we include the items in our personalized XAI Likert scale. All items are rated on a seven-point scale from "Strongly Disagree" to "Strongly Agree."

(1) I would like to be able to interactively personalize the types of explanations I receive
(2) I would like to be able to provide feedback regarding the suitability of the explanations to me.
(3) I would like to work with an explainable agent where I could provide feedback regarding the suitability of the explanation.
(4) It would be detrimental if the explainable agent did not consider the circumstances or my personal preferences while generating explanations
(5) I only care about how accurately the explanation describes the behavior of the agent
(6) I don't care if it takes me a long time to understand/parse the explanation
(7) I care more about accuracy of an explanation rather than its ease-of-understanding to me
(8) I need an explanation which explains every aspect of the underlying AI algorithm
(9) The type (e.g. feature importance, language, etc.) of explanation I receive does not matter to me
(10) I can work with any type of explanation
(11) I do not need an explanation personalized to me
(12) I am satisfied as long as I receive an explanation

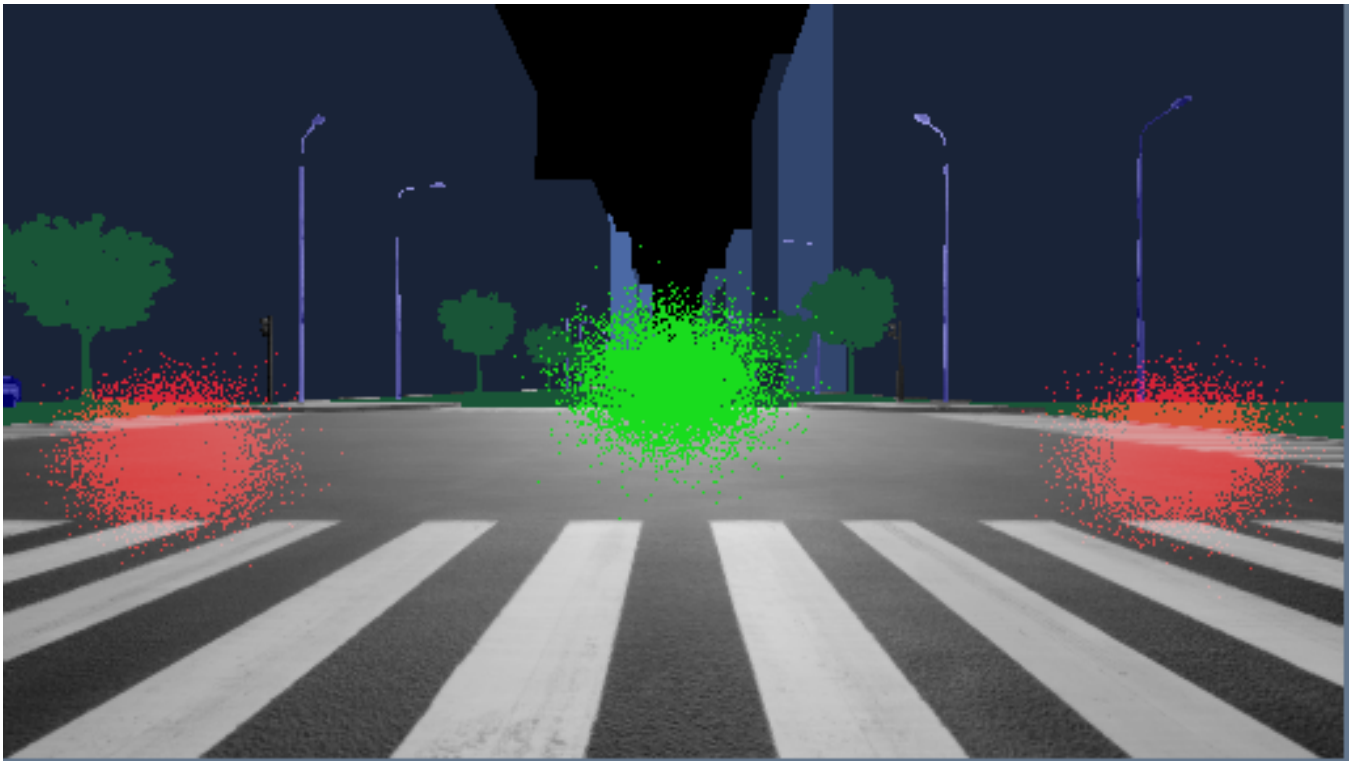


Figure 4: An example of the saliency map explanation from our study.

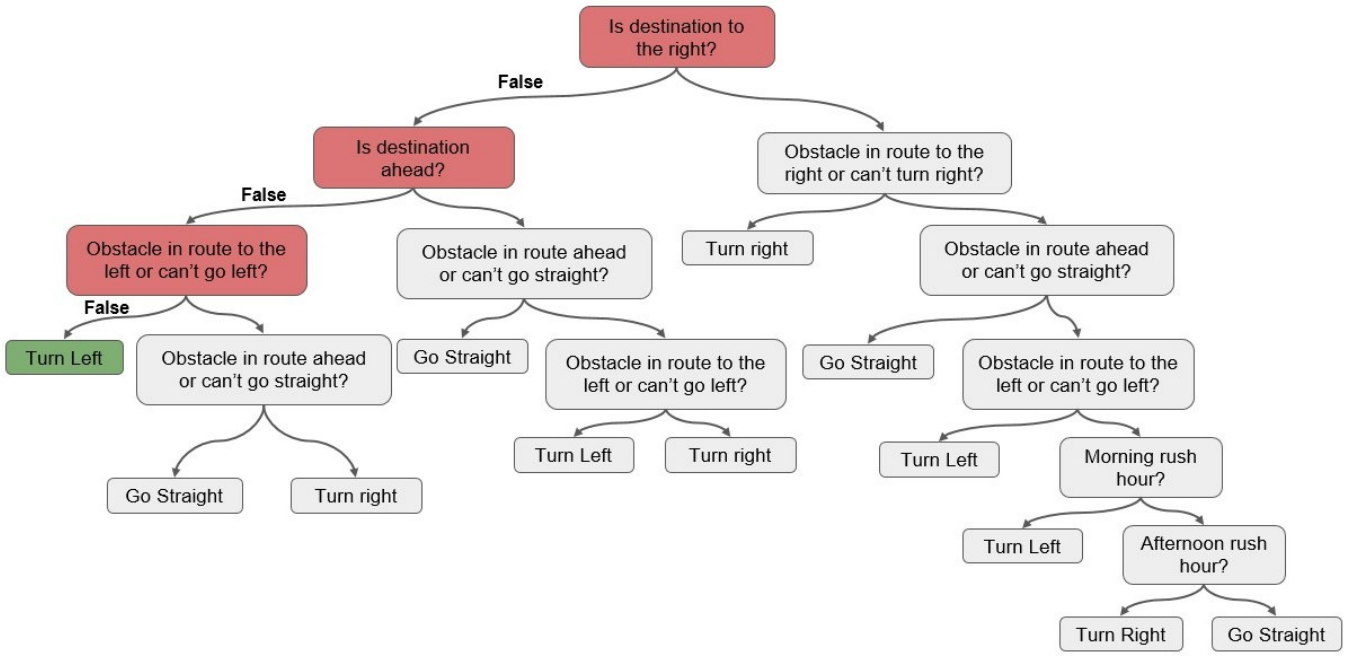


Figure 5: An example of the decision tree explanation in our study.