Towards Balancing Preference and Performance through Adaptive Personalized Explainability

Andrew Silva*
andrew.silva@tri.global
Toyota Research Institute
Cambridge, Massachusetts, USA

Pradyumna Tambwekar
ptambwekar3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Mariah Schrum*
mariahschrum@berkeley.edu
University of California, Berkeley
Berkeley, California, USA

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

arXiv:2504.13856v1 [cs.HC] 21 Mar 2025

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 (xAI) has made great strides to enable such communication, these advances often assume that one xAI approach is ideally suited to each problem (e.g., decision trees to explain how to triage patients in an emergency or feature-importance maps to explain radiology reports). This fails to recognize that users have diverse experiences or preferences for interaction modalities. In this work, we present two user-studies set in a simulated autonomous vehicle (AV) domain. We investigate (1) population-level preferences for xAI and (2) personalization strategies for providing robot explanations. We find significant differences between xAI modes (language explanations, feature-importance maps, and decision trees) in both preference ($p < 0.01$) and performance ($p < 0.05$). We also observe that a participant's preferences do not always align with their performance, motivating our development of an adaptive personalization strategy to balance the two. We show that this strategy yields significant performance gains ($p < 0.05$), and we conclude with a discussion of our findings and implications for xAI in human-robot interactions.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Explainability, Personalization, User Studies

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 [8, 76]. Researchers have identified *explainability* as a necessary component of high-quality human-robot interactions in many domains [26, 85]. While several approaches for explainability are under active investigation (e.g., natural language explanations [24], decision-tree extraction [95], counterfactual presentation [52], saliency-based explanations [83, 102], etc.), existing studies on human-use of explanations is almost entirely confined to treating explanation as a "one-size-fits-all" problem [68, 72, 81]. However, explanations have different functional roles with respect to deployment context [4, 34], suggesting that personalization and contextualization of explanations is an important and understudied avenue to bring explainability to the real world. If individual preferences and expertise affect the success of an explanation [104], a natural next step is to identify which xAI modalities should be shown to an individual user for any given decision.

Within the field of xAI, simply measuring the accuracy or fidelity of an explanation (with respect to the underlying agent or algorithm) is not enough to know that an explanation was useful. If explanations do not carefully consider a user's expertise or expectations, the simple act of showing an explanation can cause the user to blindly trust an agent's advice, leading to adverse effects on performance and trust [81, 97]. This counter-intuitive result presents a key problem: explanations encourage inappropriate compliance. If users see explanations and defer to robots without critically examining the robot suggestion, then researchers must develop a deeper understanding of the relationship between explanations and compliance while also improving an xAI agent's ability to expose faulty decision-making to human users [28]. By improving our understanding of such relationships and by better calibrating to end-users, we can produce xAI systems that are not only easier and more enjoyable to use, but also improve outcomes and efficiency of human users [31, 87].

Ultimately, xAI research seeks to help humans understand when to rely on vs. override their AI assistants, using explanations to determine if decisions are sound and trustworthy [42]. Such a dynamic exists when humans collaborate with fallible AI assistants– a scenario that we recreate in this work. Our work aims to understand the diverse preferences of untrained humans with potentially-faulty assistants that use xAI to support human decision making. We present a set of studies in which participants interact with a virtual AV to navigate through an unknown city with the assistance of a digital agent, replicating a common problem of navigating in a new place. Crucially, this assistive agent is not always correct, and incorrect advice is signalled with the inclusion of red-herring features (e.g., if the agent refers to "weather" in its explanation, the suggestion is wrong). Our work therefore simulates the use

*Work completed while at the Georgia Institute of Technology. This work reflects solely the opinions and conclusions of the authors and not of TRI or any Toyota entity.

of xAI for explanatory debugging [19, 22] with concept-based explanations [21], also called the "glitch detector task" [41, 43]. We investigate how xAI may improve people's mental models for AI [2, 10], and how *personalized* xAI will affect people's ability to accurately identify when their assistant is correct or incorrect (i.e., if the agent adapts to the user, will the user make fewer mistakes?). Our contributions include:

(1) We design two studies in which participants interact with xAI modalities randomly or using personalization.
(2) We empirically study participants' preferences for certain types of explanations, as well as their performance with such explanations, finding that language explanations are significantly preferred ($p < 0.05$) and lead to fewer mistakes ($p < 0.05$) relative to other modalities.
(3) We develop a novel adaptive personalization approach to dynamically balance a participant's preference- and performance-based needs depending on their progress in a task.
(4) We find that adapting to a participant's preferences while also maximizing their performance leads to fewer mistakes relative to naive, randomly-chosen explanations or a preference-maximization approach ($p < 0.05$), and leads to significantly greater perceptions of preference-accommodation relative to an agent that does not personalize ($p < 0.05$).

Unlike prior work on personalizing xAI, which only considers user-preferences [17, 55, 58, 63], our work is the first to directly account for the user's task-performance when personalizing xAI. Our work takes a crucial first step towards understanding *how* future work should consider personalization in xAI as well as *why* such personalization matters.

2 RELATED WORK

In this work, we investigate the effects of personalizing xAI mechanisms to users, focusing on three primary modalities for explanation: language generation and counterfactuals [51, 52, 88], feature-importance maps [35, 83], and decision-tree explanations [85, 95]. We investigate the domain of AVs to study personalized xAI, a domain of increasing interest to the HRI community [1, 32, 62, 71].

Personalization – With the proliferation of digital assistants and machine learning in consumer products, the problem of personalization has become more pressing. While conventional machine learning applies a single model to all data, the real-world contains many problems where the same sample may have different labels depending on the user (e.g., people wanting to customize a social robot's greeting, behavior, or appearance [33, 57, 82]). This problem setup requires personalization of the shared model, such that individual users can receive personally-tailored responses from the learned model [16, 61], ideally without needing to retrain the entire model from scratch. There have been several approaches for personalization with such shared models, including personal model heads for each user [3, 16, 25, 53, 61, 64, 80, 86] or meta-learned models that can rapidly adapt to users [15, 23, 30, 36, 37, 50, 66].

Personal Embeddings – In this work, we build on personalization via personal embeddings [45, 77, 89, 90, 96, 98, 103], in which a unique embedding is assigned to each user and appended to the input data or hidden representations of the network, thereby allowing the model to adapt its decision-making by conditioning on this unique per-person embedding.

Adaptive Personalization – While personalizing to different human users in this work, our system must balance between two objectives– user-preference and task-performance. We refer to this balancing act as "adaptivity". Prior work [5, 6, 84] also develops "adaptive" approaches to personalization, though prior definitions only consider task-performance in a single domain. In contrast, our approach is generally applicable to any explanatory debugging scenario with concept-based explanations, and dynamically balances between task-performance and human preferences.

Explainability – While personalization helps to bring machine learning to wider audiences and a greater diversity of problems, the inherent unpredictability of models remains an obstacle to wider deployment of learned solutions. There are legal [106] and practical criteria for machine learning models to be used in many contexts [26, 27]. xAI is a subfield of machine learning research seeking to help justify a model's decision-making using a variety of approaches. The field has contributed many techniques, such as developing neural network models that can be readily interpreted [107], pursuing natural language generation for explanations [14, 29, 72, 109], presenting feature-importance maps for input samples [13, 35, 48, 83, 101, 102, 109], presenting relevant training data [11, 12, 54, 94], and other techniques [42, 44, 67].

While research has begun to investigate the effects of explanations on user's ability to understand and forecast network behavior [2, 39, 42, 47, 68, 73, 81, 95, 97, 99, 105] or the social implications of working with robots that explain their behavior [9, 20, 56, 100], there is considerably less work on understanding how such dynamics would unfold if a human were afforded the ability to influence the xAI agent more directly, such as through controlling what types of explanations are provided. Prior work has studied the effects of explanations on compliance with inaccurate suggestions [81, 97]; however, it is possible that such explanations were simply ill-suited to the study participants and that personalization would help mitigate this problem. Prior work has also shown that user-characteristics and dispositional factors can significantly impact a user's interaction with an explainable agent [69, 70, 92, 93], and that aligning explanations with a user's expertise can lead to higher perceived utility [79]. In this work, we address these oversights in prior work and enable active personalization based on user feedback, studying compliance with personalized xAI.

3 STUDY SETUP

In our work, we present two separate studies using the same environment and driving agent. The first is a **population study** with a within-subjects design targeted at identifying population-wide trends, and serving as a data capture for our personalization model. The second is a **personalization study** with a mixed design to test the effects of different personalization strategies on how well they align with participant's preferences and how effectively they maximize task performance. We also present the design of an adaptive personalization agent that seeks to jointly satisfy both objectives. In this section, we provide background for the shared environment, driving agent, metrics, and research questions in the two studies,

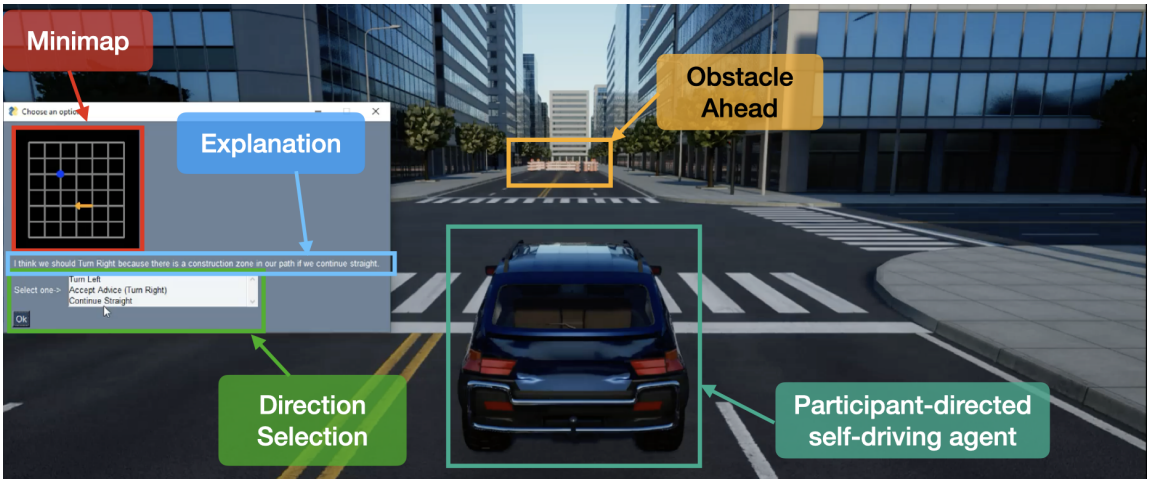


Figure 1: Here we show an example interaction with a language explanation and a correct suggestion. Taking a wrong turn (e.g., going straight) will lead directly into a roadblock, forcing participants return to this intersection and repeat the interaction.

before detailing study procedures, metrics, and results in Sections 4 & 5. All studies in this work took 60-75 minutes, and participants were compensated $20 for their time. All studies were approved by an Institutional Review Board (IRB).

3.1 Environment

The domain we employed for our experiments was a simulated driving domain. The participant interacts with an AV, reflecting a robot-deployment that is becoming increasingly common in the real world. Furthermore, autonomous driving is an accessible and easily-understandable domain for a non-expert end-user, and is a highly pertinent area of study for human-robot communication and xAI [59, 60]. In our study, the human is responsible for all navigational direction, but the robot handles all actual control of the vehicle. This task was setup through the AirSim driving simulator [91] and built in the Unreal Engine. Our domain features a simulated city with a seven-by-seven grid layout, effectively putting the participants into a small maze that they will navigate for the duration of each task. Participants direct an AV through the maze to the goal location in the city, working with assistance from a self-driving agent and a small mini-map.

Each intersection in the domain presents an opportunity to select a direction to progress through the environment, giving the participant all available navigation options (e.g. “turn left,” “turn right,” or “continue straight”) alongside a directional suggestion from the self-driving agent and an explanation justifying the suggestion. Participants consider the suggestions and explanations to help them decide how to navigate through the city. After making a decision, participants are also asked to provide binary positive/negative feedback on whether or not they would like to see more explanations with the modality that they received. We provide an example of one such interaction in Figure 1.

The city contains several roadblocks, thereby creating a single optimal path to the goal, with any deviation resulting in either a U-turn (as the car drives down a road with a roadblock and must turn around) or a significantly slower path. For each task, the participant has a new starting and goal location, and roadblocks are moved around the map. This relocation prevents participants from memorizing routes through the city, and encourages reliance on navigational assistance from the self-driving agent.

3.2 AI Driving Agent and Explanations

At each intersection in the domain, a digital agent suggests a direction and also presents an explanation for its suggestion to the participant. Explanations are either a sentence in natural language, a feature-importance map, or a decision-tree (Figure 2). All explanations were manually generated before the study, rather than autonomously generated via an existing machine learning method [52, 83, 95], to control for existing explainability research and to more closely examine the modalities themselves. To identify the optimal direction, the agent uses a breadth-first search planner over the grid to find the shortest path to the goal.

Approximately 30% of the time, the agent will suggest the *opposite* of the optimal direction, alongside a flawed explanation attempting to rationalize the incorrect suggestion. Participants are trained to identify these *incorrect* explanations before beginning the study. The threshold for performance was chosen following prior work on agent reliability in user studies [78, 108, 110]. Further details on how incorrect suggestions are provided and signalled are given in the supplementary material.

3.3 Metrics

Shared Metrics – In both studies, we employ the following metrics:

- *Inappropriate Compliance* - The proportion of incorrect advice accepted by participants. The better a participant understands a particular xAI method, the lower this metric should be.
- *Mistakes* - The number of mistakes made by the participant, as an additional gauge of the participant’s ability to interpret an explanation modality.

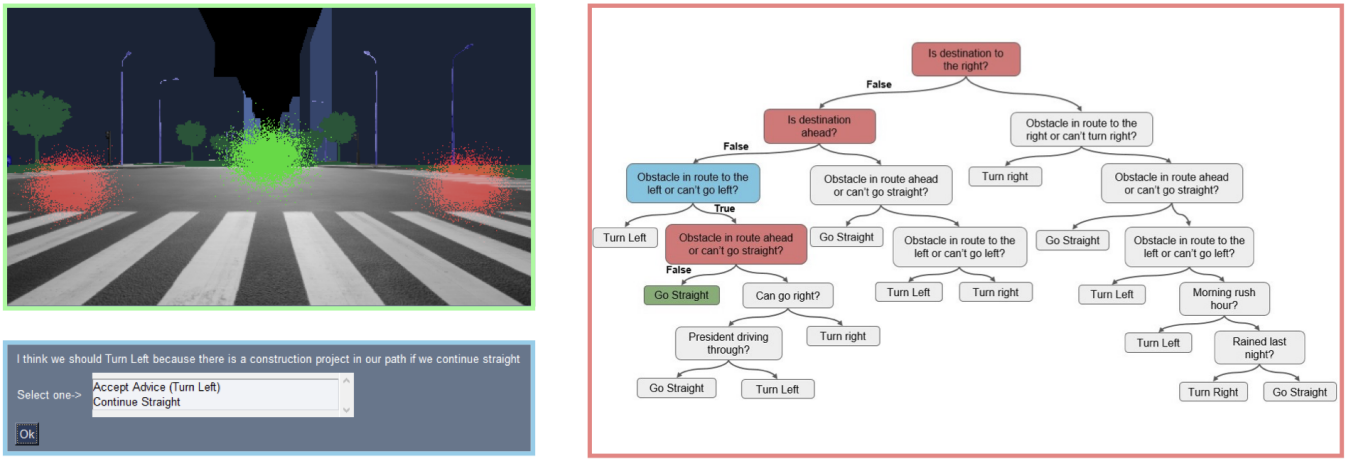


Figure 2: We compare three xAI modalities in this work: feature-importance maps, (top left) in which highlighted regions indicate possible directions and relevant elements of the image, such as green indicating the suggested direction, language explanations (bottom left) that are a sentence justifying one direction over another, and decision trees (right) in which the highlighted path leads to the suggested direction. Red blocks mean "false" and blue blocks mean "true".

- *Binary Feedback Ratings* - Participants' answers to a "yes/no" question about whether the participant would like to work with a specific xAI modality again, which is asked to the participant after every intersection. We record the total number of positive and negative responses for each xAI modality across the study.

Population Study Metrics – In the population study, we also measure:

- *Preference Rankings* - Rankings from a 5-item ranking survey between the explanation modalities. These values will be high if participants felt that the condition did a better job of accommodating their personal preferences.
- *Consecutive mistakes* - Back-to-back mistakes as a result of rejecting correct suggestions or accepting incorrect suggestions.
- *Consideration Time* - The amount of time a participant considers an explanation prior to making a decision.

In the population study, participants work with a fixed xAI modality for an entire task. Therefore, through measuring consecutive mistakes, we are able to infer a participants' reaction to making a mistake, i.e. does the specific xAI modality enable them to better reflect on what they missed in the previous iteration, or will they make repeat mistakes? Similarly, consideration time tells us if one modality is slower or faster than others.

Personalization Study Metrics – Finally, the personalization study also measures:

- *Steps Above Optimal* - The number of steps to complete a task, relative to the optimal solution.
- *Preference Annotations* - Free form text responses to how well the different agents accommodated participant preferences. Free form text allowed participants to describe the various successes and failures of different personalization strategies without being confined to a predefined ranking survey.

We do not measure consideration time or consecutive mistakes in this study, as such metrics target the xAI modalities themselves rather than the personalization strategies we seek to compare and because the xAI modality can change between interactions (i.e., modalities are not fixed, as in the population study). We also change the approach to measuring preference, giving us more insight into why participants preferred one option over another by requiring participants to describe their preferences [82].

3.4 Research Questions

Our work aims to understand both (1) population-wide trends on preference and performance for diverse xAI modalities, and (2) the effects of different personalization strategies on human-robot teaming with xAI. As our study uses a novel domain, we first sought to verify whether there was a specific modality that led to the highest average preference or performance. Furthermore, recent work has highlighted a nuanced relationship between preference and performance, often in relation to external factors, such as expertise [76, 104]. To gain insight into these relationships, our population study is designed with the following research questions in mind:

- **RQ1.1 – Preferences**: Will one xAI modality be significantly more preferred than others?
- **RQ1.2 – Performance**: Will one xAI modality lead to significantly better performance than others?
- **RQ1.3 – Alignment**: Will participants prefer to use the modality that maximizes their performance?

The personalization study seeks to examine the degree to which balanced personalization affects participants' performance on a task and on their perceptions of the agent's accommodation of their preferences (i.e., does balanced personalization make people feel like the agent is listening to them while also helping them perform better?). The primary research questions are then:

- **RQ 2.1 – Preferences** Will balanced personalization be significantly more preferred than other personalization strategies?

- **RQ 2.2 – Performance** Will balanced personalization lead to significantly fewer mistakes than other personalization strategies?
- **RQ 2.3 – Comparison to known-best** Will balanced personalization match or exceed task-performance and preference metrics when compared to the a-priori known best xAI modality for the study task (i.e., language explanations).

4 POPULATION STUDY

The population study enables us to study overall trends for preference and task-performance with our chosen xAI modalities and domain. This study helps to determine which mode, if any, is superior for this task and enables us to clearly analyze the relationship between performance and task-preference for each xAI modality.

4.1 Study Conditions

The population study is a within-subjects design to study the effects of xAI modalities. Therefore, the conditions in this study are the xAI modalities themselves (Section 3.2), including (1) Language, (2) Feature Maps, and (3) Decision Tree explanations. Each of these conditions were chosen to reflect popular avenues of explainability within human-robot or human-AV interactions [29, 46, 75, 76].

4.2 Procedure

Upon arrival to the onsite location, participants complete consent forms and are briefed on their task. Participants are introduced to each of the xAI mechanisms employed in the study, the interface for directing the car, and a mini-map that will assist them for each task. They then complete the Negative Attitudes towards Robots Scale (NARS) [74], "Big-Five" personality [18], and demographic data surveys, used as controls in our statistical analyses. Participants then begin on eleven navigation tasks (Section 3.1).

Participants complete two practice tasks to become acquainted with the simulator, controls, and explanations. Pilot studies revealed that very little practice was required for the task, so two tasks was sufficient. In this practice phase, explanations are randomly sampled from any of the three mechanisms used in our study (Section 3.2), giving the participant equal practice with each modality.

After completing the practice phase, participants begin the main body of the study, which consists of nine navigation tasks. Each task uses a single xAI modality from start to finish, which helps to mitigate consecutive mistakes that may stem from swapping between xAI modalities. The agent rotates between modalities as tasks are completed, and the ordering of xAI modalities is included as a control in our statistical analyses. Participants conclude the study with a survey asking them to rank the three xAI modalities according to their preferences.

4.3 Results

The population study involved 30 participants (Mean age = 23.8, SD = 3.25; 70% Male).

RQ 1.1 – Comparing the sum across five Likert-items as the preference rank, an ANOVA for xAI modality rankings showed a significant difference across baselines ($F(2, 84) = 35.1, p < 0.001$). A Tukey-HSD revealed that language explanations ranked significantly higher than both feature-importance maps ($p < 0.001$) and decision trees ($p < 0.001$), and feature-importance maps ranked significantly higher than decision trees ($p < 0.001$) (Figure 3).

RQ 1.2 – Data for inappropriate compliance did not pass a Shapiro-Wilk test for normality, and we therefore applied a Friedman's test, which was significant ($\chi^2(2) = 12.23, p = 0.002$), with a post-hoc revealing that language explanations lead to significantly fewer instances of inappropriate compliance than feature-importance maps ($p = 0.002$) (Figure 3).

Data for consideration time were not normally distributed. We therefore applied a Friedman's test, which was significant ($\chi^2(2) = 24.47, p < 0.001$). A post-hoc revealed that both language and feature-importance explanations are significantly faster than decision-tree explanations ($p < 0.001$).

Finally, an ANOVA across explanation modalities for consecutive mistakes was significant ($F(2, 74) = 8.0309, p < 0.001$), and a post-hoc revealed that participants make significantly more consecutive mistakes with feature-importance maps ($p < 0.001$) and language explanations ($p = 0.001$) than with decision trees.

RQ 1.3 – After grouping participants by their preferred modality, we do not find any significantly different trends for performance (i.e., participants that favor feature-importance maps do not perform best with feature-importance maps).

4.3.1 Takeaways. A review of the results for the population study reveals that language explanations are both significantly preferred relative to feature-importance and decision-tree explanations, and result in higher task-performance than feature-importance explanations. We do find, in line with contemporary work [104], that decision trees result in significantly fewer consecutive mistakes, suggesting that it is easier for participants to reflect on why the previous decision was incorrect and to immediately update their mental model of the agent. However, across most metrics, language explanations are superior to both other modalities considered in this work. We therefore consider language explanations to be the "gold standard" for this task (i.e., an agent that presents only language explanations will be significantly preferred and yield significantly higher task-performance for most of the population).

5 PERSONALIZATION STUDY

The population study helped identify a "gold-standard" explanation for our domain, and revealed significant population-wide trends with regards to preference and performance. However, knowledge of the best xAI modality is not readily available for most domains, and there may be adverse effects of universally applying population-wide trends on an individual level [111]. Our personalization study therefore studies the effects of different personalization strategies on preference- and task-performance-maximization, including a novel adaptive personalization approach that balances between a participant's preferences and task-performance needs[1].

[1]The xAI modalities are adjusted slightly between the population and personalization studies, following feedback from some population-study participants that the explanations were too simplistic and easy to memorize. Each modality was therefore made slightly more complicated, and an additional 12 pilot participants verified that the new explanations fairly reflected the results of the population study, while increasing in complexity. Details and examples are available in the supplementary material.

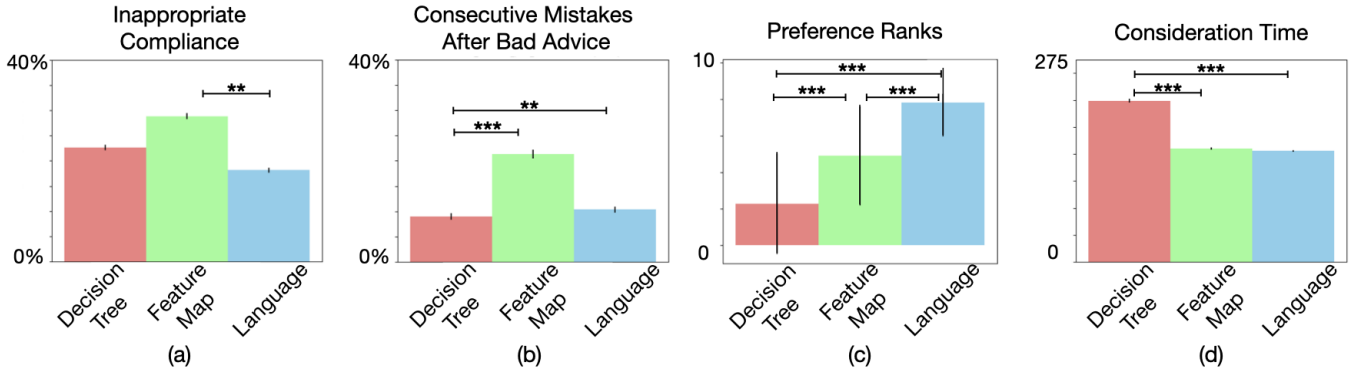


Figure 3: Visualized results from the population user study between decision trees, feature-importance maps, and language explanations. (a) Feature maps lead to significantly increased inappropriate compliance. (b) Both feature maps and language explanations lead to more consecutive mistakes (quantities are normalized by total number of mistakes). (c) Language is significantly preferred over decision trees and feature maps. (d) Decision trees are slower to parse (measured in seconds).

5.1 Adaptive Personalization Approach

The population study revealed that preference and performance do not necessarily align. This observation inspired the development of an adaptive personalization technique that can balance between a participant's preferences or performance-needs depending on the situation. To accomplish this personalization, our agent must model preferences or performance for each participant.

As discussed in Section 3.3, the agent tracks how often participants make mistakes (i.e., do not follow the optimal path), as well as the participant's feedback, for each xAI modality. Using these two metrics, the agent creates a preference distribution and a task-performance distribution for the participant. These distributions are largely driven by *negative* interactions with the agent (i.e., negative feedback or mistaken turns). The decision to focus on negative feedback owes to pilot studies, which revealed that negative interactions are rarer and more meaningful than positive interactions.

5.1.1 Producing Sampling Distributions. First, the agent counts all interactions for each modality and stores the resulting values in a vector, $\vec{x}$ (e.g., counting the total number of language, feature-map, and decision-tree interactions). The agent then separates this out into two additional quantities– the total number of negative interactions, $\vec{x_-}$ and the total number of positive interactions $\vec{x_+}$. The basis of the sampling distribution is then computed as $\vec{x_-} * \frac{\vec{x}}{\vec{x_+}}$. In other words, the total number of negative interactions for each modality, smoothed by the ratio of total-to-positive interactions. This normalized quantity will be high for modalities where interactions are more often negative, and low for xAI modalities that have far more positive than negative interactions.

The agent tallies the number of modalities with at least one negative interaction, which normalizes the above quantity, smoothing the distribution if negative interactions occur in all modalities. Finally, this value is negated so that modalities with higher negative values will be sampled less frequently. The distribution over all xAI modalities is computed according to Equation 1.

$$\vec{v} = -\frac{\vec{x}_- * \frac{\vec{x}}{\vec{x}_+}}{\sum_{i=0}^{|\vec{x}|}(1 \text{ if } \vec{x}_{-,i} > 0)} \tag{1}$$

The resulting distribution, $\vec{v}$, is then normalized using a softmax function to produce a probability distribution. When $\vec{x}$ is the vector of task-performance interactions (i.e., correct and incorrect turns), we obtain a distribution for task-performance, $\vec{d_T}$. If $\vec{x}$ is instead a vector of feedback interactions, we obtain a distribution of the participant's preferences, $\vec{d_P}$. Sampling from $\vec{d_P}$ will maximize the likelihood of selecting an explanation that aligns with satisfying a participant's preferences, while sampling from $\vec{d_T}$ will maximize the likelihood of selecting an explanation that helps the participant to identify the optimal action. However, there is no notion of balancing between these two, potentially-competing, objectives.

5.1.2 Balancing Between Multiple Objectives. To achieve the balance between participant preference and task-performance, we must find a way to balance between $\vec{d_P}$ and $\vec{d_T}$ depending on the participant's progress in the task. To this end, we define a new distribution, $\vec{d_B}$, that balances between $\vec{d_P}$ and $\vec{d_T}$, using a trade-off parameter λ. The trade-off parameter should emphasize $\vec{d_P}$ if the participant is going to be correct (i.e., adhere to participant preferences if there is little risk of a mistake), and emphasize $\vec{d_T}$ if there is a high risk of the participant being incorrect (i.e., ignore preferences and maximize task-performance if a mistake is likely). $\vec{d_B}$ is therefore constructed according to Equation 2.

$$\vec{d_B} = \lambda * \vec{d_P} + (1 - \lambda) * \vec{d_T} \tag{2}$$

To obtain this trade-off parameter, λ, the agent requires an estimate of whether the participant is likely to make a mistake. To this end, the agent employs a neural network to predict which direction the participant will choose at each intersection. This network consumes state information from the environment (e.g., position, orientation, goal position, and nearest roadblock position), and is trained over data collected from pilot studies and the population study (Section 4). However, because participants often take different paths, this network must personalize to each participant. The network therefore maintains a unique embedding for each participant, following prior personalization research [77, 90, 96]. Similarly, the network maintains a unique embedding for each navigation task,

allowing for contextual adaptation in addition to individualized personalization, as in [98]. These embeddings are both passed into the network alongside state information. During deployment the main body of the network is frozen, and only the personal and contextual embeddings are updated as participants act in the domain.

At each intersection, this model predicts which direction the participant is going to choose, producing a probability distribution over the three possible directions, $\vec{y}$. If the agent predicts that the participant is going to go in the optimal direction, then the trade-off parameter, λ, is set to $argmax(\vec{y})$ (i.e., set to the logit value for the optimal direction). Otherwise, λ is set to $1 - argmax(\vec{y})$.

5.2 Study Conditions

The personalization study compares five xAI-selection strategies:

- **Balanced personalization** – explanations are drawn from $\vec{d}_B$, as described in Section 5.1.
- **Preference maximization** – explanations are drawn from $\vec{d}_P$, only conditioning on participant preferences.
- **Task-performance maximization** – explanations are drawn from $\vec{d}_T$, only conditioning on participant task-performance.
- **Random explanations** – explanations are randomly selected from the three available modalities.
- **Language-only explanations** – all explanations use the language condition, known a-priori to yield the best performance and match most people's preferences (Section 4.3.1).

5.3 Procedure

Following a pilot study, the personalization study is designed as a set of within-subjects experiments. In this study, participants work with two personalization strategies (Section 5.2). Each experiment compares balanced-personalization to one other approach for choosing explanations. This design helps to control for variance across participants, which we observed to be high. The ordering of these two strategies is counter-balanced across all participants.

The personalization study begins by following the same procedures as in the population study (Section 4.2). After being briefed on the task and the xAI modalities, participants complete a single training task and then begin two calibration tasks. All three tasks (one training and two calibration) randomly cycle through all available xAI modalities, giving participants exposure to the various explanations they will receive. Over the two calibration tasks, the agent begins to gather feedback and observations on the participants behavior to create $\vec{d}_P$ and $\vec{d}_T$. The agent also uses these tasks to learn personal and contextual embeddings for the personalization network (Section 5.1.2).

After calibration tasks, the participants complete three navigation tasks with the first selection strategy, and then stop to complete a set of surveys on trust [49], perceptions of social competence [7], perceived workload [38], and explainability [97]. Participants then provide free-form text on how well they thought that the agent conformed to their preferences. After this set of questions, participants resume the navigation tasks, completing an additional three tasks with the second selection strategy. Finally, participants complete the same set of questionnaires a second time.

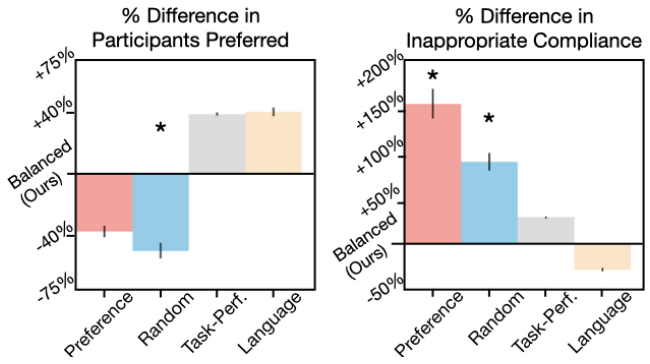


Figure 4: Comparisons relative to balanced-personalization. (Left) Percent of participant preferences for personalization modes, showing significant preference for balanced personalization over no personalization. (Right) Rates of inappropriate compliance, showing that balanced-personalization leads to significantly lower inappropriate compliance than preference-maximization or no personalization.

5.4 Results

The personalization study involved 60 participants (Mean age=24.87, SD=6.75; 53% Male). While a subset of significant results are presented here, all statistical test details and pairwise condition comparisons are presented in full supplementary material.

RQ 2.1 – Comparing balanced personalization to random explanations, a wilcoxon signed rank test for preference rankings showed a significant difference across conditions ($W = 4, p = 0.04$). Additionally, comparing binary feedback ratings for preference maximization and balanced-personalization, a Friedman's test revealed significantly higher feedback ratings for preference-maximization ($\chi^2(1) = 5.444, p = 0.02$). Interestingly, we find that, while preference-maximization leads to significantly higher positive binary feedback during the study, it appears to result in lower retrospective preference ratings after the study (Figure 4).

We do not find statistically significant differences between either feedback data or text-responses for the task-performance agent vs. the balanced-personalization agent. We therefore find evidence to answer **RQ 2.1**– balanced personalization is significantly preferred over random explanations.

RQ 2.2 – Comparing balanced-personalization to random explanations, a wilcoxon signed rank test reveals significantly higher inappropriate compliance using random explanations ($W = 25, p = 0.03$). We also observed significantly higher inappropriate compliance in the preference-maximization agent compared to balanced-personalization ($W = 21, p = 0.02$) (Figure 4). Similarly, participants took significantly more steps above optimal performance with a preference-maximization agent compared to a balanced-personalization agent ($W = 31, p = 0.04$). We therefore answer **RQ 2.2**– balanced personalization yields significantly higher performance relative to preference-maximization or random explanations.

RQ 2.3 – We find no statistically significant differences between the balanced-personalization agent and language-only agent along the performance or preference metrics. We therefore find that balanced personalization is not worse along task-performance and

preference metrics when compared to the a-priori known best xAI modality for our domain.

5.4.1 Takeaways. Reviewing the results of the personalization study, we find that balanced personalization is significantly superior to no personalization (i.e., random explanations) along the axes of both preference and task-performance. Similarly, balanced personalization leads to significantly fewer mistakes than preference maximization. We find no significant differences between task-performance maximization, suggesting that task-performance is of paramount importance for participants in our study [110]. In other words, our study found that receiving explanations the participants did not prefer to use (e.g., seeing mostly decision trees even when asking to stop receiving them) did not register as the agent not conforming to participant preferences, so long as the participant was perceived to be making progress on the task. This trend could potentially be due to "experienced accuracy", wherein a participant's experience or perception of an agent's accuracy affects their interactions with the system [65, 112].

Finally, we find no statistically significant differences between language-only explanations (known to be best before the study) and a balanced-personalization agent. This finding implies that balanced personalization will not under-perform the best xAI mode for a new domain, while avoiding the need to run a population study. Deploying balanced personalization can significantly reduce the overhead for deploying xAI to new domains while also ensuring that participants receive xAI modalities that match their needs (e.g., feature-importance maps for participants that cannot use language).

6 DISCUSSION AND LIMITATIONS

In the population study, we find significant differences between the three xAI modalities, decision trees, language, and feature-importance maps, examined in this work when considering participant preference and task-performance. Despite these trends, we find an interesting counterexample, in which decision-tree explanations are significantly better for identifying mistaken decision-making processes for consecutive errors (Section 4.3). This result echoes prior work [104], finding that language explanations were significantly preferred by untrained participants, but that participants were better at modeling agent behavior when using decision trees.

In the personalization study, we confirm that balanced personalization yields significant performance improvements relative to preference maximization or no personalization. Furthermore, while participants provide significantly more negative responses to a balanced-personalization agent, they retrospectively perceived the balanced-personalization to do a *better* job conforming to their preferences. Similarly, we find that a task-performance maximization agent receives very positive retrospective perceptions of personalization (40% over balanced-personalization, shown in Figure 4), despite *never* considering $\vec{d}_P$ when making decisions. Together, these findings suggest that participants in the personalization study prized task performance over preference-accommodation. Until a satisfactory level of performance is met, accommodating preferences may not be perceived as important or useful, as participants seem to fixate on optimally completing the task rather than on engaging with an agent that listens to their feedback. This finding echoes prior work on the effects of performance on trust [110], and underscores the importance of personalizing xAI for task performance. Applying a balanced personalization approach, as introduced in this work, we can achieve the benefits of maximizing task-performance at crucial junctions (e.g., if the user is likely to make a mistake or if their mental model appears to be incorrect) while *also* accommodating user preferences.

While our work was conducted on a simulated task, our findings generalize more broadly to any domain in which a human might need to work with concept-based explanations [22], such as feature-maps, decision trees, or counterfactual explanations. Using such explanations, a mistake is often identified by the inclusion of an errant feature, as in this research. We show that humans have diverse preferences and experiences when working with such explanations, and that adaptive personalization can enhance human-robot interactions that rely on xAI for decision-verification [40].

Limitations – These studies were conducted primarily with university students on a driving simulator, rather than on an autonomous vehicle with a broader population. Additionally, we observe some results with large effect sizes but no statistical significance (Figure 4), which may be due the sample size in our studies.

Personalization agents in this work also assumed access to ground truth information in the domain (e.g., optimal turns) or explicit preference feedback, which may be challenging to obtain the real-world. While such ground-truth directions could come from external sources (GPS navigation) and feedback data could come from observations of humans [5], these external data sources are not used in our work. Finally, personalization in this work was confined to selecting xAI modalities that match a participant's preferences, but did not extend to adapting the explanations themselves.

7 CONCLUSION

To be useful in the real world, digital agents and robots must be able to personalize to a diverse population of users, even without prior knowledge of the best way to interact or the most popular interaction modality for a given task. In this work, we have studied the differences between three popular explainability techniques: language explanations, feature-importance maps, and decision trees, in the context of a simulated AV study. We have presented an approach to personalization that balances subjective human-preference with objective task-performance. A separate user study confirmed that such a balanced personalization approach yields significantly improved task-performance relative to a preference maximization agent, and is not worse than an agent that uses explanations which are known a-priori to maximize preference scores and task-performance in the population. Our study is the first to personalize explanations to task-performance, and we show that personalization must consider task-performance to be successful. We discuss the implications of our results, including the need for high-performance agents before considering preference maximization, and the need to carefully balance adherence to preferences with task-performance.

ACKNOWLEDGEMENTS

This work was supported by NSF award CNS 2219755, NASA Early Career Fellowship 80NSSC20K0069, and a gift from Konica Minolta.

REFERENCES

[1] Anna M. H. Abrams, Pia S. C. Dautzenberg, Carla Jakobowsky, Stefan Ladwig, and Astrid M. Rosenthal-von der Pütten. 2021. A Theoretical and Empirical Reflection on Technology Acceptance Models for Autonomous Delivery Robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 272–280. https://doi.org/10.1145/3434073.3444662

[2] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–37.

[3] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818* (2019).

[4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[5] Agnes Axelsson and Gabriel Skantze. 2023. Do You Follow? A Fully Automated System for Adaptive Robot Presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 102–111. https://doi.org/10.1145/3568162.3576958

[6] Nils Axelsson and Gabriel Skantze. 2019. Modelling Adaptive Presentations in Human-Robot Interaction using Behaviour Trees. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Satoshi Nakamura, Milica Gasic, Ingrid Zuckerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino (Eds.). Association for Computational Linguistics, Stockholm, Sweden, 345–352. https://doi.org/10.18653/v1/W19-5940

[7] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009).

[8] 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

[9] Roel Boumans, René Melis, Tibor Bosse, and Serge Thill. 2023. A Social Robot for Explaining Medical Tests and Procedures: An Exploratory Study in the Wild. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 263–267. https://doi.org/10.1145/3568294.3580085

[10] Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, et al. 2023. Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 220–239.

[11] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the origins of bias in word embeddings. *arXiv preprint arXiv:1810.03611* (2018).

[12] Rich Caruana, Hooshang Kangarloo, JD Dionisio, Usha Sinha, and David Johnson. 1999. Case-based explanation of non-case-based learning methods.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association.

[13] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[14] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2021. Generate natural language explanations for recommendation. *arXiv preprint arXiv:2101.03392* (2021).

[15] Letian Chen, Sravan Jayanthi, Rohan Paleja, Daniel Martin, Viacheslav Zakharov, and Matthew Gombolay. 2022. Fast Lifelong Adaptive Inverse Reinforcement Learning from Demonstrations. In *Proceedings of the 6th Conference on Robot Learning (CoRL), 2022*.

[16] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. 2021. Exploiting Shared Representations for Personalized Federated Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). 2089–2099.

[17] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503.

[18] Andrew J. Cooper, Luke D. Smillie, and Philip J. Corr. 2010. A confirmatory factor analysis of the Mini-IPIP five-factor model personality scale. *Personality and Individual Differences* 48, 5 (2010), 688–691. https://doi.org/10.1016/j.paid.2010.01.004

[19] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for Robot Failures: Generating Explanations That Improve User Assistance in Fault Recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 351–360. https://doi.org/10.1145/3434073.3444657

[20] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for Robot Failures: Generating Explanations That Improve User Assistance in Fault Recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 351–360. https://doi.org/10.1145/3434073.3444657

[21] Devleena Das, Sonia Chernova, and Been Kim. 2023. State2Explanation: Concept-based explanations to benefit agent learning and user understanding. In *Proceedings of the Conference on Neural Information Processing Systems*.

[22] Devleena Das, Been Kim, and Sonia Chernova. 2023. Subgoal-Based Explanations for Unreliable Intelligent Decision Support Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 240–250. https://doi.org/10.1145/3581641.3584055

[23] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. 2020. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461* (2020).

[24] 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).

[25] Canh T Dinh, Nguyen H Tran, and Tuan Dung Nguyen. 2020. Personalized federated learning with moreau envelopes. *arXiv preprint arXiv:2006.08848* (2020).

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

[27] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J. Gershman, David O'Brien, Kate Scott, Stuart Shieber, Jim Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. 2017. *Accountability of AI Under the Law: The Role of Explanation*. SSRN Scholarly Paper ID 3064761. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3064761

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

[29] 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.

[30] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948* (2020).

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

[32] Paul D. S. Fink, Anas Abou Allaban, Omoruyi E. Atekha, Raymond J. Perry, Emily S. Sumner, Richard R. Corey, Velin Dimitrov, and Nicholas A. Giudice. 2023. Expanded Situational Awareness Without Vision: A Novel Haptic Interface for Use in Fully Autonomous Vehicles. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 54–62. https://doi.org/10.1145/3568162.3576975

[33] Naomi T. Fitter, Megan Strait, Eloise Bisbee, Maja J. Mataric, and Leila Takayama. 2021. You're Wigging Me Out! Is Personalization of Telepresence Robots Strictly Positive?. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 168–176. https://doi.org/10.1145/3434073.3444675

[34] Leilani H Gilpin, Andrew R Paley, Mohammed A Alam, Sarah Spurlock, and Kristian J Hammond. 2022. " Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI. *arXiv preprint arXiv:2207.00007* (2022).

[35] Amar Halilovic and Felix Lindner. 2023. Visuo-Textual Explanations of a Robot's Navigational Choices. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 531–535. https://doi.org/10.1145/3568294.3580141

[36] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. 2020. Lower bounds and optimal algorithms for personalized federated learning. *arXiv preprint arXiv:2010.02372* (2020).

[37] Filip Hanzely and Peter Richtárik. 2020. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516* (2020).

[38] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human*

Mental Workload, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

[39] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491

[40] Bradley Hayes and Julie A. Shah. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 303–312. https://doi.org/10.1145/2909824.3020233

[41] Robert R Hoffman, John W Coffey, Kenneth M Ford, and Mary Jo Carnot. 2001. Storm-lk: A human-centered knowledge model for weather forecasting. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 1. 752–752.

[42] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[43] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257.

[44] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz* (2020), 1–6.

[45] Fang-I Hsiao, Jui-Hsuan Kuo, and Min Sun. 2019. Learning a multi-modal policy via imitating demonstrations with mixed behaviors. *arXiv preprint arXiv:1903.10304* (2019).

[46] 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.

[47] Amanda Hutton, Alexander Liu, and Cheryl Martin. 2012. Crowdsourcing evaluations of classifier interpretability. In *2012 AAAI Spring Symposium Series*.

[48] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3543–3556. https://doi.org/10.18653/v1/N19-1357

[49] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000).

[50] Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. 2019. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488* (2019).

[51] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050* (2020).

[52] 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*. 353–362.

[53] Joongheon Kim, Seunghoon Park, Soyi Jung, and Seehwan Yoo. 2021. Spatio-temporal split learning. In *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S)*. IEEE, 11–12.

[54] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

[55] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2020. Generating and understanding personalized explanations in hybrid recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–40.

[56] Johannes Maria Kraus, Julia Merger, Felix Gröner, and Jessica Pätz. 2023. 'Sorry' Says the Robot: The Tendency to Anthropomorphize and Technology Affinity Affect Trust in Repair Strategies after Error. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 436–441. https://doi.org/10.1145/3568294.3580122

[57] Alyssa Kubota, Emma I. C. Peterson, Vaishali Rajendren, Hadas Kress-Gazit, and Laurel D. Riek. 2020. JESSIE: Synthesizing Social Robot Behaviors for Personalized Neurorehabilitation and Beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/3319502.3374836

[58] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *arXiv preprint arXiv:2301.09656* (2023).

[59] Seong Hee Lee, Nicholas Britten, Avram Block, Aryaman Pandya, Malte F. Jung, and Paul Schmitt. 2023. Coming In! Communicating Lane Change Intent in Autonomous Vehicles. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 394–398. https://doi.org/10.1145/3568294.3580113

[60] Seong Hee Lee, Vaidehi Patil, Nicholas Britten, Avram Block, Aryaman Pandya, Malte F. Jung, and Paul Schmitt. 2023. Safe to Approach: Insights on Autonomous Vehicle Interaction Protocols with First Responders. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 399–402. https://doi.org/10.1145/3568294.3580114

[61] Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581* (2019).

[62] Jamy Li, Rebecca Currano, David Sirkin, David Goedicke, Hamish Tennent, Aaron Levine, Vanessa Evers, and Wendy Ju. 2020. On-Road and Online Studies to Investigate Beliefs and Behaviors of Netherlands, US and Mexico Pedestrians Encountering Hidden-Driver Vehicles. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 141–149. https://doi.org/10.1145/3319502.3374790

[63] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems* 41, 4 (2023), 1–26.

[64] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems* 2 (2020), 429–450.

[65] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[66] Zhaojiang Lin, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. *arXiv preprint arXiv:1905.10033* (2019).

[67] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2021), 18.

[68] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2493–2500.

[69] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2020. What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 173–182.

[70] Martijn Millecamp, Sidra Naveed, Katrien Verbert, and Jürgen Ziegler. 2019. To explain or not to explain: The effects of personal characteristics when explaining feature-based recommendations in different domains. In *Proceedings of the 6th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, Vol. 2450. CEUR; http://ceur-ws. org/Vol-2450/paper2. pdf, 10–18.

[71] Dylan Moore, Rebecca Currano, Michael Shanks, and David Sirkin. 2020. Defense Against the Dark Cars: Design Principles for Griefing of Autonomous Vehicles. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 201–209. https://doi.org/10.1145/3319502.3374796

[72] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695* (2018).

[73] Dong Nguyen. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1069–1078. https://doi.org/10.18653/v1/N18-1097

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

[75] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2022), 10142–10162. https://doi.org/10.1109/TITS.2021.3122865

[76] 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*.

[77] Rohan Paleja, Andrew Silva, Letian Chen, and Matthew Gombolay. 2020. Interpretable and Personalized Apprenticeship Scheduling: Learning Interpretable Scheduling Policies from Heterogeneous User Demonstrations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6417–6428.

[78] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors* 52, 3 (2010), 381–410. https://doi.org/10.1177/0018720810376055 arXiv:https://doi.org/10.1177/0018720810376055 PMID: 21077562.

[79] Jeyoung Park, Jeeyeon Kim, Da-Young Kim, Juhyun Kim, Min-Gyu Kim, Jihwan Choi, and WonHyong Lee. 2022. User Perception on Personalized Explanation by Science Museum Docent Robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 973–975. https://doi.org/10.1109/HRI53351.2022.9889654

[80] Matthias Paulik, Matt Seigel, Henry Mason, Dominic Telaar, Joris Kluivers, Rogier van Dalen, Chi Wai Lau, Luke Carlson, Filip Granqvist, Chris Vandevelde, et al. 2021. Federated Evaluation and Tuning for On-Device Personalization: System Design & Applications. *arXiv preprint arXiv:2102.08503* (2021).

[81] 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*.

[82] Samantha Reig, Michal Luria, Janet Z. Wang, Danielle Oltman, Elizabeth Jeanne Carter, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2020. Not Some Random Agent: Multi-Person Interaction with a Personalizing Service Robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 289–297. https://doi.org/10.1145/3319502.3374795

[83] 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*.

[84] Amelie Sophie Robrecht, Markus Rothgänger, and Stefan Kopp. 2023. A Study on the Benefits and Drawbacks of Adaptivity in AI-generated Explanations. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, Würzburg, Germany.

[85] 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).

[86] Ognjen Rudovic, Nicolas Tobis, Sebastian Kaltwang, Björn Schuller, Daniel Rueckert, Jeffrey F Cohn, and Rosalind W Picard. 2021. Personalized Federated Deep Learning for Pain Estimation From Face Images. *arXiv preprint arXiv:2101.04800* (2021).

[87] 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* (Ingolstadt, Germany) *(MuC '21)*. Association for Computing Machinery, New York, NY, USA, 325–329. https://doi.org/10.1145/3473856.3474018

[88] Florian Schröder, Sonja Stange, and Stefan Kopp. 2023. Resolving References in Natural Language Explanation Requests about Robot Behavior in HRI. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 772–774. https://doi.org/10.1145/3568294.3579981

[89] Mariah L Schrum, Erin Hedlund-Botti, and Matthew Gombolay. 2022. Reciprocal MIND MELD: Improving Learning From Demonstration via Personalized, Reciprocal Teaching. In *Proceedings of the 6th Conference on Robot Learning (CoRL), 2022*.

[90] Mariah L Schrum, Erin Hedlund-Botti, Nina Moorman, and Matthew C Gombolay. 2022. MIND MELD: Personalized Meta-Learning for Robot-Centric Imitation Learning. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 157–165.

[91] 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.

[92] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Enhancing Fairness Perception–Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. *International Journal of Human–Computer Interaction* (2022), 1–28.

[93] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24, 1 (2022), 2.

[94] Andrew Silva, Rohit Chopra, and Matthew Gombolay. 2022. Cross-Loss Influence Functions to Explain Deep Network Representations. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, 1–17. https://proceedings.mlr.press/v151/silva22a.html

[95] 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.

[96] Andrew Silva, Katherine Metcalf, Nicholas Apostoloff, and Barry-John Theobald. 2022. FedEmbed: Personalized Private Federated Learning. *arXiv preprint arXiv:2202.09472* (2022).

[97] 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.

[98] Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2023. FedPerC: Federated Learning for Language Generation with Personal and Context Preference Embeddings. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, Dubrovnik, Croatia.

[99] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

[100] Sonja Stange and Stefan Kopp. 2020. Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 619–627. https://doi.org/10.1145/3319502.3374802

[101] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665.

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

[103] Aviv Tamar, Khashayar Rohanimanesh, Yinlam Chow, Chris Vigorito, Ben Goodrich, Michael Kahane, and Derik Pridmore. 2018. Imitation learning from visual data with multiple intentions. In *International Conference on Learning Representations*.

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

[105] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[106] Paul Voigt and Axel Von dem Bussche. 2017. The EU General Data Protection Regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017).

[107] Tong Wang, Cynthia Rudin, Finale Velez-Doshi, Yimin Liu, Erica Klampfl, and Perry MacNeille. 2016. Bayesian rule sets for interpretable classification. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE.

[108] Rebecca Wiczorek and Dietrich Manzey. 2014. Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human factors* 56, 7 (2014), 1209–1221.

[109] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 11–20. https://doi.org/10.18653/v1/D19-1002

[110] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 408–416.

[111] Dewei Yi, Jinya Su, Cunjia Liu, Mohammed Quddus, and Wen-Hua Chen. 2019. A machine learning based personalized system for driving state recognition. *Transportation Research Part C: Emerging Technologies* 105 (2019), 241–261. https://doi.org/10.1016/j.trc.2019.05.042

[112] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

8 ADDITIONAL DOMAIN DETAILS

Before and after each study, participants complete several surveys including: demographics, negative attitudes towards robots [74], mini-IPIP personality survey [18], and experience with driving, robots, and decision trees. Following the completion of a portion of the studies in this work, participants complete a robot trust survey [49], the NASA-TLX survey of perceived workload [38], a survey on perceptions of anthropomorphism and social competence of the digital agent [7], and a survey on the perceived explainability of a digital agent [97].

9 OVERALL STUDY FLOW

We provide an overview figure for our study flow in Figure 5. In our population study, participants first complete consent forms and are briefed on the task. They then fill out pre-surveys and demographic information, before training on the simulator for two tasks. After the training phase, participants being the main body of the study, rotating through each of the available conditions for a total of 3 tasks with each xAI modality (e.g., language, then feature-importance, then decision-trees, then repeat the cycle two more times). After their final task, participants provide their preference rankings for each of the xAI modalities, and are finally debriefed on their experience.

In the personalization study, participants also begin with consent forms, briefing, pre-survey, and demographic surveys. Participants then complete one training task and two calibration tasks, in which the driving agent rotates through each xAI modality for every intersection. After the second (and final) calibration task, participants are randomly assigned to either the adaptive personalization strategy or to a baseline condition. Participants complete three tasks with this strategy, then fill out a preference survey. After the preference survey, participants resume the study with the other personalization strategy (either adaptive or baseline). After completing three more tasks, participants redo the preference survey for the second personalization strategy, and are then debriefed.

10 PROVIDING INCORRECT SUGGESTIONS AND EXPLANATIONS

In both studies, incorrect suggestions were provided as the exact *opposite* of the correct direction, and participants were warned of this information at the beginning of the study. We opted to make incorrect suggestions point in the opposite direction (as opposed to a random incorrect direction) so that participants could, in theory, always take the optimal route (e.g., if the agent is incorrect and says "go left", the participant knows that going "right" is optimal).

Because there are often three available directions, choosing an "opposite" is possible for only two out of three options (i.e., left and right). When the correct direction is simply "straight", we reduce the number of options by arbitrarily disabling one direction. In other words, the agent randomly masks out "right" or "left", thereby only presenting two options to the participant. The incorrect suggestion is whichever direction was not masked ("left" or "right"), and the correct direction is to go in the only other option available (i.e., "straight"). Participants are told how to handle this situation during the briefing (Appendix 18).

In practice, we found that many participants did pick up on these rules, and did not struggle to know how to handle an explanation that they perceived to be incorrect (though they did not always understand when or why explanations were incorrect). Commonly, participants struggled when they knew a suggestion was incorrect, but they wanted to accept the suggestion because the direction itself seemed to be correct from their viewpoint. For example, consider the situation where the goal is immediately to the participant's left, but a construction site lay between the participant and the goal. A participant that wants to get to the goal as quickly as possible will want to turn left. If the digital assistant incorrectly suggests "left" and provides an invalid explanation, the participant may recognize that they *should not* go left, but they may turn left anyway, simply because they already expected they should go that way (and often, they hoped that "this time it will be right,"). Similarly, if the agent *correctly* suggested an alternate direction, participants may recognize that they *should* comply (i.e., take a less direct path, such as going "right" in the above example), but still end up going the wrong way, simply because they hope that their more direct path will work.

11 INCORRECT EXPLANATIONS

Incorrect explanations were signalled by the inclusion of "red-herring" features, as told to participants (Section 18). These included: the weather, the radio, the sky, traffic, rush hour, or the president's motorcade. Participants are explicitly told most of these (they are not explicitly told about the president's motorcade, though they do see it in the training and calibration tasks, so they are able to learn that rule before beginning the main task). They are also explicitly told that any explanation considering "external factors" (i.e., not pertaining to the road or to construction sites and car crashes) is an incorrect explanation. An example of an incorrect decision tree explanation is given in Figure 7.

Examples of incorrect explanations in our work included:

- "I think we should turn left because the president's motorcade is in town."
- "I think we should continue straight because it didn't rain last night."
- "I think we should turn right because clear skies could impair our cameras."

12 CORRECT EXPLANATIONS

Correct explanations were signalled by failing to include "red-herring" features. In particular, if the explanation pertained only to the "shortest path" or "optimal route", then it was correct. Additionally, if the explanation contained only information about the goal or obstacles on the path to the goal (e.g., construction sites or car crashes), then it was correct. For feature-importance maps, which did not necessarily contain any of this information, explanations were correct as long as they did not highlight the sky. An example of a correct decision tree explanation is given in Figure 8.

Examples of correct explanations in our work included:

- "I think we should turn left because there is a construction project in our path if we turn right."
- "I think we should continue straight because it is the shortest path to the goal."
- "I think we should turn right because we will hit a pile-up if we continue straight."

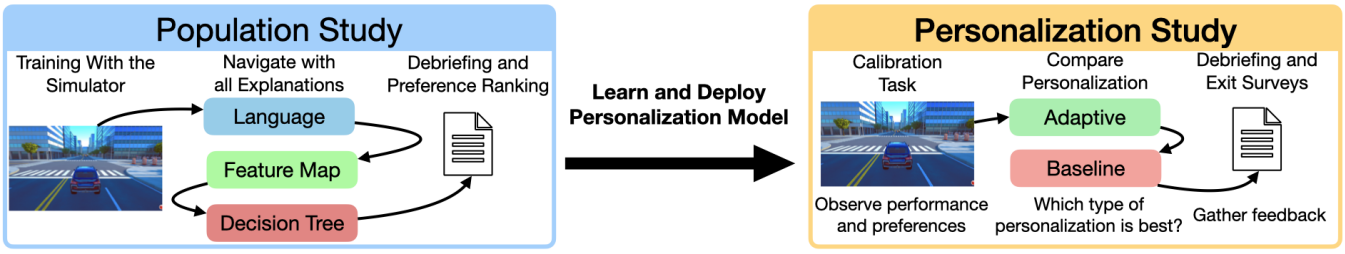
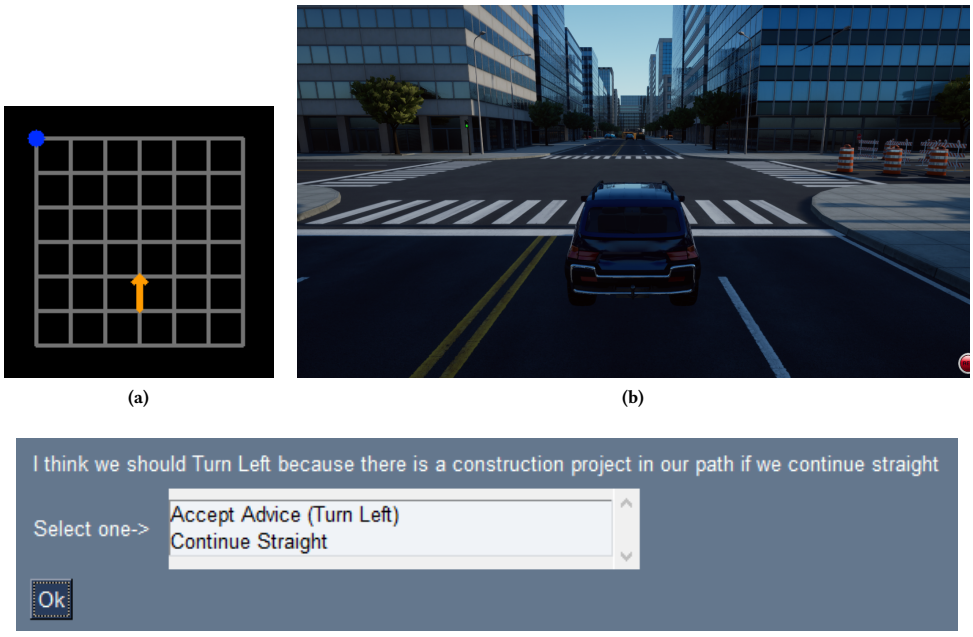


Figure 5: We conduct two user studies, beginning with a population study (left) in which all participants work with three xAI modalities, revealing significant differences across the population. We then use this data to build an adaptive personalization model that is deployed in a set of personalization studies to compare various personalization approaches.



(a) (b) (c)

Figure 6: Our custom self-driving domain created using Unreal and the AirSim [91] simulator. 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.

13 EXPLANATION CHANGES BETWEEN STUDIES

After the conclusion of the population study, we opted to change the content of some of the explanation modalities. This is because a few participants mentioned that they found it easier to memorize all *correct* explanations, rather than learning the rule to identify *incorrect* explanations (particularly for the language modality). However, the goal of our work is to study when xAI enables users to identify errant decision-making from a digital assistant, not to study how easy it is to memorize all possible correct explanations (which would be impractical in a real-world setting).

To make the explanations more complex, we added several language templates and rephrases, thereby greatly increasing the number of possible language explanations (from 6 up to 47). We also added one new decision node and 2 new leaf nodes to the decision tree. For reference, the original decision tree is presented in Figure 9, and the updated tree is presented in Figure 10. Finally, for feature-importance explanations, we modulated the brightness (i.e. importance) of buildings and trees in the image. Rather than being set to a static color, the color and brightness was changed to be randomly sampled, with a low set of values defined for correct explanations (Figure 15), and a high set of values defined for incorrect explanations (Figure 18). We conducted a pilot study with 12 additional participants using these new explanations, which showed that there were no significantly different trends from the

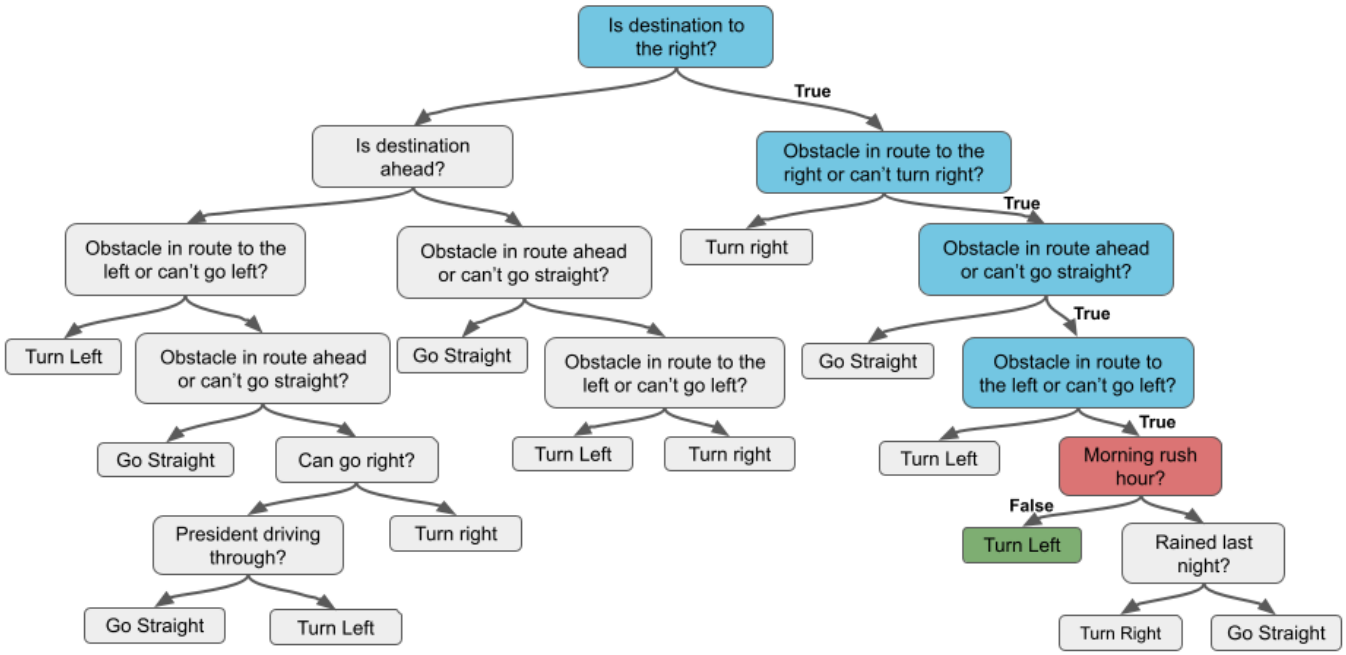


Figure 7: A decision tree explanation that the participant should ignore. Note that the highlighted path (i.e., the decision suggestion) includes a red-herring feature (rush-hour traffic).

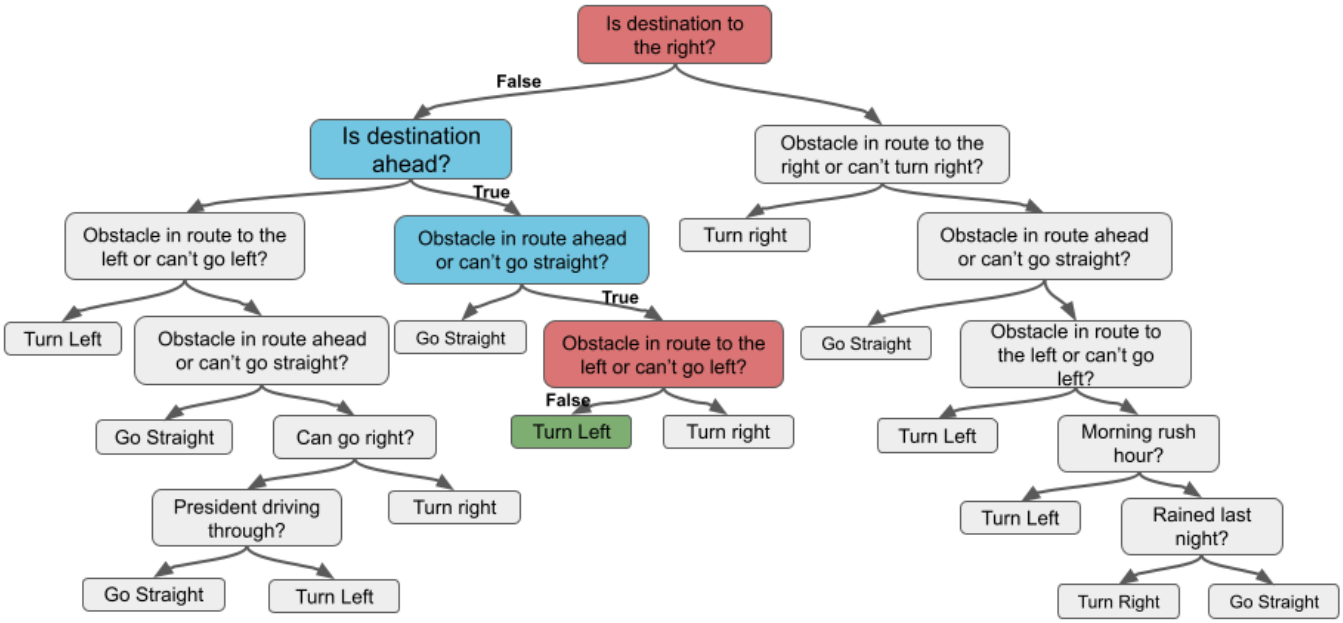


Figure 8: A decision tree explanation that the participant should adhere to. Note that the highlighted path (i.e., the decision suggestion) considers only relevant details (path to the goal and obstacles) and ignores red-herring features.

population study. Upon examining the results and debriefing participants, we found that no participant was able to memorize correct explanations after these changes.

14 TASK ORDERINGS

In the population study, participants were required to complete nine navigation tasks, three with each explanation modality. Participants rotated between explanation modalities, such that every third task

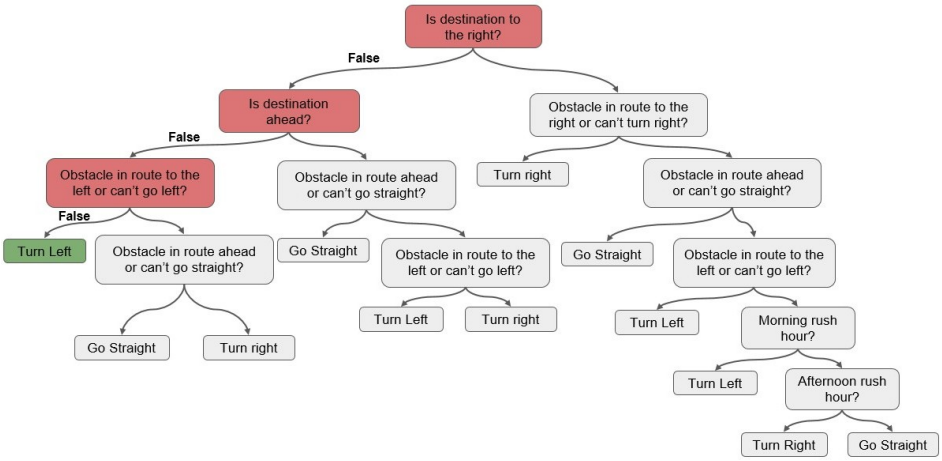


Figure 9: Decision tree explanation from the population study.

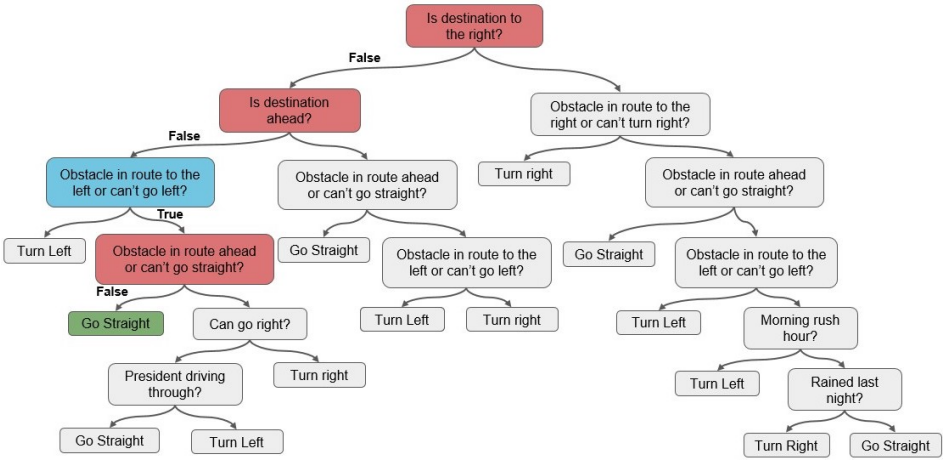


Figure 10: Decision tree explanation from the personalization study, with one decision node and 2 leaf nodes added.

was completed with the same explanation modality. We enumerated all six possible orderings of explanation modes (e.g., (1) decision-tree, (2) feature-importance, (3) language, or (1) feature-importance, (2) language, (3) decision tree, etc.) and distributed participants evenly across all orderings, such that each ordering received five participants.

In the personalization study, we included a total of six "test" tasks (i.e., not training or calibration). All participants went through the same six test tasks. We constructed a Latin square to order tasks for participants, running the studies with balanced orderings. We also balance the ordering of which explanation selection strategy is shown first or second, resulting in a study with 12 participants for each comparison. An additional 12 participants (for a total of 24) are recruited for the task-performance maximization vs. balanced personalization study (i.e., $\vec{d}_T$ vs. $\vec{d}_B$), following the results of a power-analysis.

14.1 Statistical Test Details

We performed a repeated-measures multivariate analysis to compute the effects of different conditions (explanation modality in the population study, personalization approach in the personalization study) on various metrics (explanation preference ranking, inappropriate compliance, etc.). Across various tests, the condition is modeled as a fixed-effect covariate, participant ID is a random effect covariate. We use the AIC metric to determine which additional covariates should be included for each test, considering task ordering, condition ordering, and different resposnes from a demographic data pre-survey (e.g., race, gender, age, robotics experience, etc.). We then apply an analysis of variance (ANOVA) to identify significance across baselines, and further employ 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. If the data do not pass normality assumptions, we apply a non-parametric Friedman's test with a Nemenyi's All-Pairs Comparisons post-hoc. If the data do not pass homoscedasticity assumptions, we proceed with a wilcoxon signed rank test. Finally

for binary data and count data, we apply a wilcoxon signed rank test.

15 ADDITIONAL RESULTS

We present full pairwise comparisons between conditions in this section, showing the results when controlled for variance across participants in different conditions.

16 POPULATION STUDY STATISTICAL ANALYSES

An ANOVA for explanation modality rankings yielded a significant difference across baselines ($F(2, 84) = 35.1, p < 0.001$). Data for inappropriate compliance did not pass a Shapiro-Wilk test for normality, and we therefore applied a Friedman's test, which was significant ($\chi^2(2) = 12.23, p = 0.002$). Data for correct-non-compliance was also not normally distributed, and so we again applied a Friedman's test, which was significant ($\chi^2(2) = 12.23, p = 0.002$). An ANOVA across explanation modalities for consecutive mistakes was significant ($F(2, 74) = 8.0309, p < 0.001$). Finally, data for consideration time were not normally distributed, and we therefore applied a Friedman's test, which was significant ($\chi^2(2) = 24.47, p < 0.001$).

17 PERSONALIZATION STUDY STATISTICAL ANALYSES

17.0.1 Balanced personalization vs. language-only explanations. A wilcoxon signed rank test for preference rankings did not yield a significant difference across conditions ($W = 14, p = 0.825$). A Friedman's test over feedback data was not significant ($\chi^2(1) = 1.6, p = 0.206$). A wilcoxon signed rank test for inappropriate compliance was not significant ($W = 8, p = 0.8688$), nor did a wilcoxon signed rank test for steps above optimal ($W = 37, p = 0.3769$). Finally, an ANOVA across personalization approaches for consideration time was not significant ($F(1, 11) = 4.6718, p = 0.054$).

17.0.2 Balanced personalization vs. task-performance-based personalization. A wilcoxon signed rank test for preference rankings did not yield a significant difference across conditions ($W = 35, p = 0.9585$). A Friedman's test over feedback data was not significant ($\chi^2(1) = 2.5789, p = 0.108$). A wilcoxon signed rank test for inappropriate compliance was not significant ($W = 99, p = 0.1463$), as did a wilcoxon signed rank test for steps above optimal ($W = 97, p = 0.314$). Finally, an ANOVA across personalization approaches for consideration time was not significant ($F(1, 23) = 0.066, p = 0.799$).

17.0.3 Balanced personalization vs. preference-based personalization. A wilcoxon signed rank test for preference rankings did not yield a significant difference across conditions ($W = 3, p = 0.117$). We applied a Friedman's test for feedback data, which revealed a significant difference ($\chi^2(1) = 5.444, p = 0.0196$). A wilcoxon signed rank test for inappropriate compliance found significance ($W = 21, p = 0.01776$), as did a wilcoxon signed rank test for steps above optimal ($W = 31, p = 0.03818$). Finally, an ANOVA across personalization approaches for consideration time was not significant ($F(1, 11) = 0.1634, p = 0.694$).

17.0.4 Balanced personalization vs. random explanations. A wilcoxon signed rank test for preference rankings did yield a significant difference across conditions ($W = 4, p = 0.0363$). A Friedman's test over feedback data was not significant ($\chi^2(1) = 0.3173, p = 0.317$). A wilcoxon signed rank test for inappropriate compliance found significance ($W = 25, p = 0.03275$), though a wilcoxon signed rank test for steps above optimal did not find significance ($V = 20.5, p = 0.1531$). Finally, data for consideration time did not pass normality assumptions, and a Friedman's test was not significant ($\chi^2(1) = 0.33, p = 0.564$).

17.0.5 Participant Recruitment. We recognize that our results occasionally include large effects that are not statistically significant, and that such results may become statistically significant with a larger sample population. Our study featured over 100 participants (including pilots), but a power analysis of our preference-ranking results revealed that we would need over 60 participants for a significant effect in the preference-based personalization vs. balanced-personalization comparison. Extrapolating to the rest of our work, we would have required over 240 participants for the study. We leave such a thorough investigation to future work.

18 PARTICIPANT BRIEFING

Thank you for participating in our study! Before we get started I need you first to fill out this consent and data release form, please take your time to review it and let me know if you have any questions.

<Participant completes consent form>

Thank you! So today you are going to be helping to guide a self-driving car through a simulated city. The car will handle all of the actual control, and you will be responsible for commanding the car where to go at each intersection in the city. You will be given navigational assistance from the self-driving car in the form of on-screen prompts, so the car will tell you which way to go in order to get to the goal as fast as possible. The simulator will pause while the agent thinks about which way to go, and it takes about 4-5 seconds at each intersection for the agent to produce a suggestion. The agent will also present you with short explanations for why it's giving you a directional suggestion. You can decide at each intersection whether you want to go with what the agent suggests. The car may occasionally not allow you to travel in a direction that seems open. This happens if the car has not yet fully mapped a given turn.

These explanations can come in three different forms, either written descriptions, decision trees the car uses, or feature importance maps. And for reference, here is an example of what each of those looks like:

<Participant is shown example written description reading: "You should bring lunch to work today because the restaurants in your area will be closed for a holiday.">

In the written descriptions, you'll see a sentence explaining why one choice is better than another.

<Participant is shown an example feature importance map, shown in Figure 15.>

With the feature importance map, you'll see a highlighted image with relevant elements of the image highlighted in different colors, such as the outlines of trees and buildings next to the road. The green blob highlights the direction of the best path, whereas the other red blobs highlight the other possible directions that were not chosen due to obstacles or car crashes.

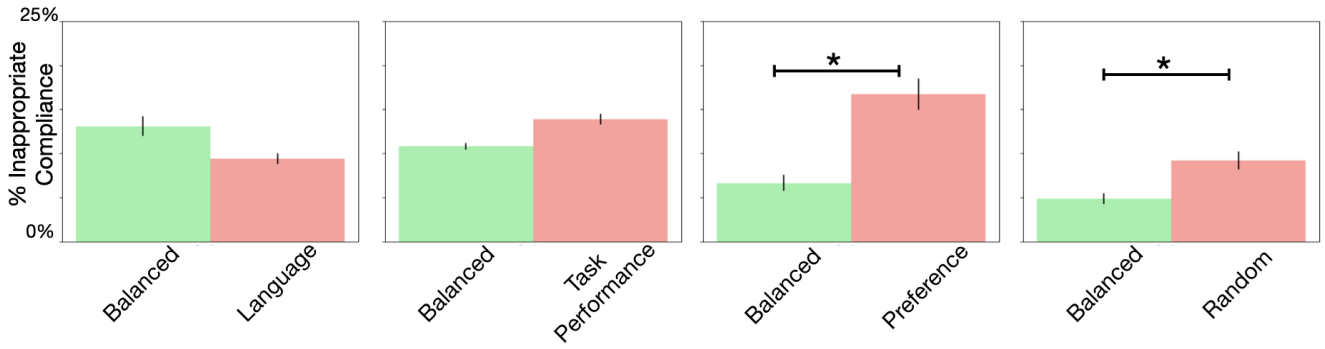


Figure 11: Visualized percent of inappropriate compliance across all condition-comparisons in the personalization study. Balanced personalization helps participants identify errant decision suggestions significant more than preference maximization or no personalization at all.

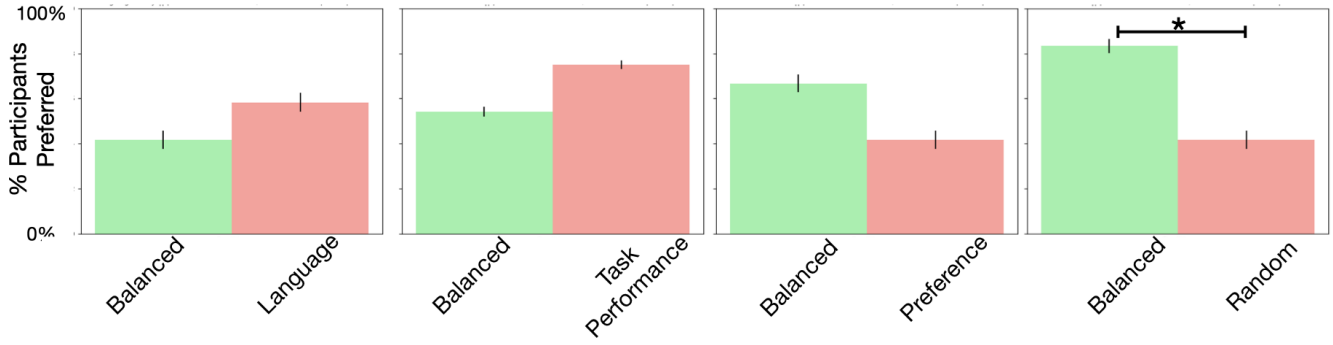


Figure 12: Visualized preference comparison scores across all conditions in the personalization study. Balanced personalization is significantly more preferred than no personalization at all.

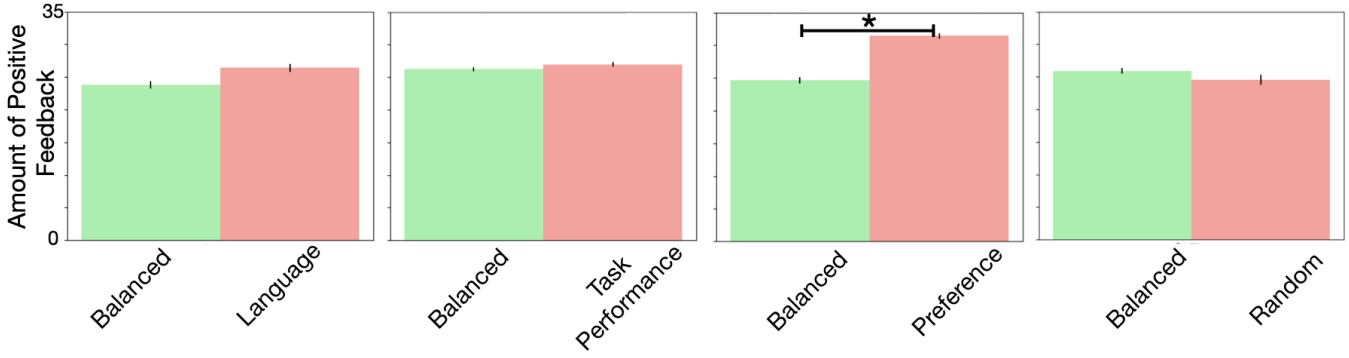


Figure 13: Visualized binary preference feedback across all condition-comparisons in the personalization study. Preference maximization results in significantly more "Yes" responses than balanced-personalization.

<Participant is shown an example decision tree, depicted in Figure 16.>

In the decision tree, you'll see a flowchart with true/false checks that lead to a decision, where a "true" check is labeled as true and highlighted in blue, and a "false" check is labeled with false and highlighted in red.

<Participant is shown an example mini-map, given in Figure 17.>

You will also have access to a mini-map at each intersection, which will look like this. The arrow indicates your current position and heading, and the blue circle is your destination.

Before we begin, could you please fill out the following pre-study survey for me, and please stop when the survey asks you not to advance further:

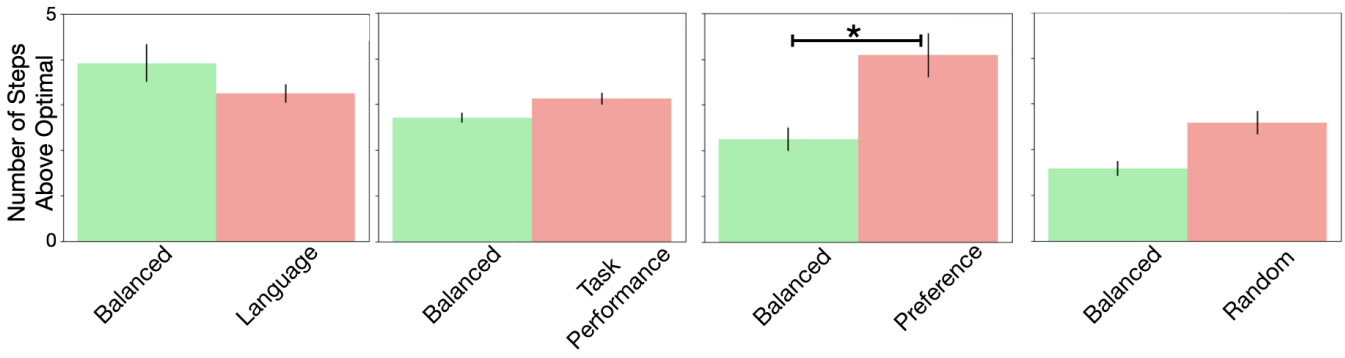


Figure 14: Visualized steps above optimal across all condition-comparisons in the personalization study. Balanced personalization helps participants reach the goal in significantly fewer steps than preference maximization.

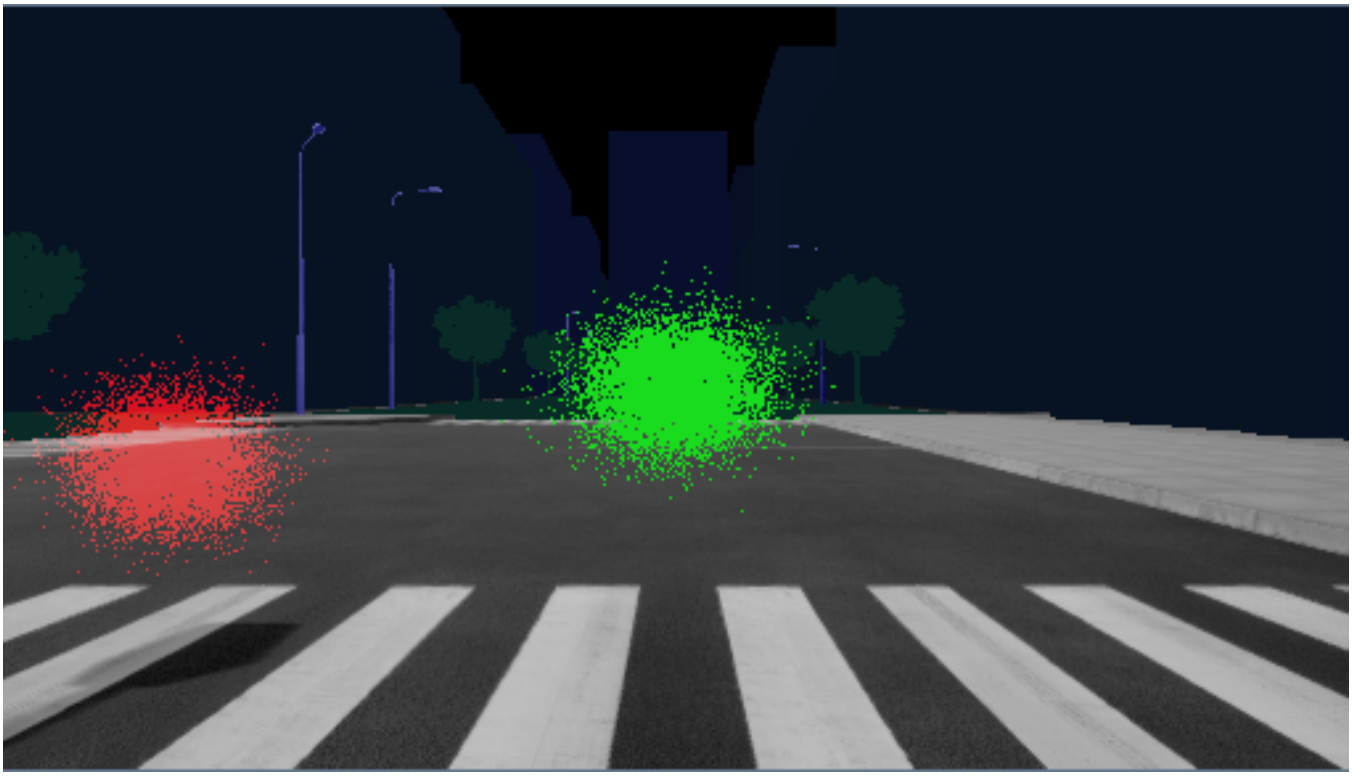


Figure 15: Feature importance map example shown to participants during the instructions phase of the study.

<Participant takes pre-study surveys>

Great, thank you! So now we will begin the actual study! As you are helping the car to navigate through the city, you will go through 9 navigation tasks. The first will be a training task for you to become acclimated to working with the car and the self-driving agent, then we'll do 2 calibration tasks for the agent to learn to accommodate your behaviors. After that, you'll navigate with one agent for 3 tasks, then we'll stop to do a couple of surveys. Finally, you'll resume the task with a new agent for 3 final tasks, and we'll conclude with a final set of surveys. For each task, the car will reset to a new location in the city, and the goal location will move.

There are a few important things to bear in mind:

First, the city is littered with construction projects and car crashes that can block certain routes, and these car crashes and construction projects can move around when the car resets.

Second, the agent is not perfect and will sometimes make mistakes. The agent is good at estimating obstacles on your shortest path to the goal. However, when it focuses on external factors such as the time of day, the sky, rush hour traffic, the weather, etc., that means the agent is malfunctioning. If this happens, the agent is

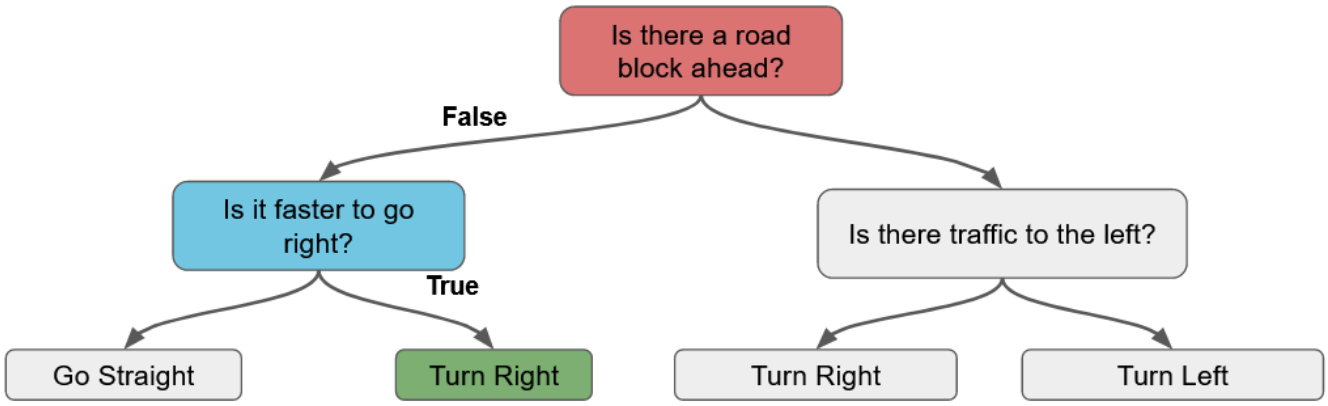


Figure 16: Decision Tree example shown to participants during the instructions phase of the study.

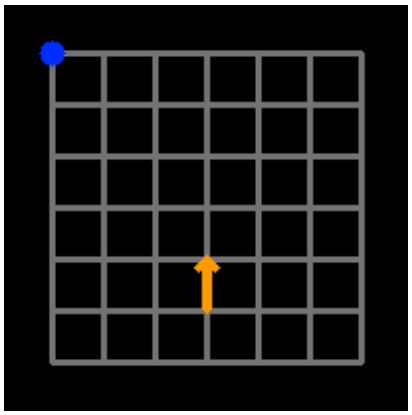


Figure 17: Mini-map example shown to participants during the instructions phase of the study.

giving you the wrong direction! In such cases, you should do the opposite of what the agent says, to the extent possible. Even when the explanation is wrong the map is correct. Note that this can happen regardless of the explanation, whether it is language, decision tree or feature map. Remember, if it is looking at external criteria such as time of day, rush hour traffic, the sky, the music on the radio, or the weather, it is wrong.

Here are examples of when each is wrong.

<Participant is shown language incorrect language explanation reading: "You should turn left because the radio is set to NPR.">

So we see here the language refers to the radio, which is an external factor so the correct decision is to turn right.

<Participant is shown an incorrect feature importance map, shown in Figure 18.>

Here we see the feature importance map is looking at the sky, again that is an external factor so you would not follow the agent's suggestion here.

<Participant is shown an incorrect decision tree, shown in Figure 19.>

And here we see that the decision tree is considering the weather, which again is an external factor, so you would go right instead of left. Of note, just because this node is incorrect, doesn't mean the entire tree is wrong, so if you were to see an explanation using the other half of this tree, for example, you could trust that suggestion. But you would not trust it if the decision used an external feature.

Third, you will be timed during the main body of the task, and we'd like for you to complete each task as quickly as possible. But, the timer will be paused while you make up your mind at each intersection, so you aren't penalized for taking time to make up your mind on which way you want to go. When you make a choice, please be mindful that you cannot undo it! Once you click "OK", the car will start to drive on, so be sure you choose the direction you want to go.

Finally, you have a maximum of 20 total interactions per task. So if you cannot reach the goal within 20 intersections, the task will end and immediately progress to the next one.

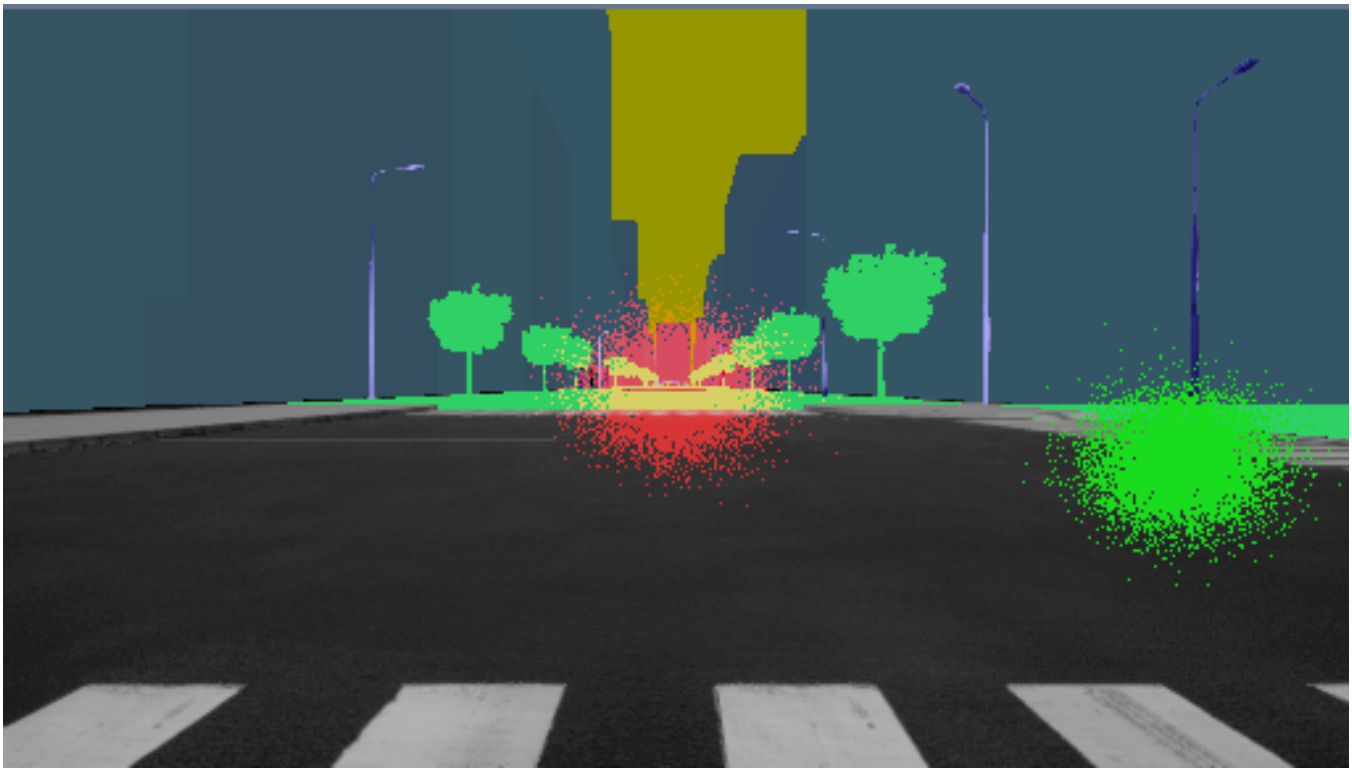


Figure 18: Incorrect feature importance map example shown to participants during the instructions phase of the study.

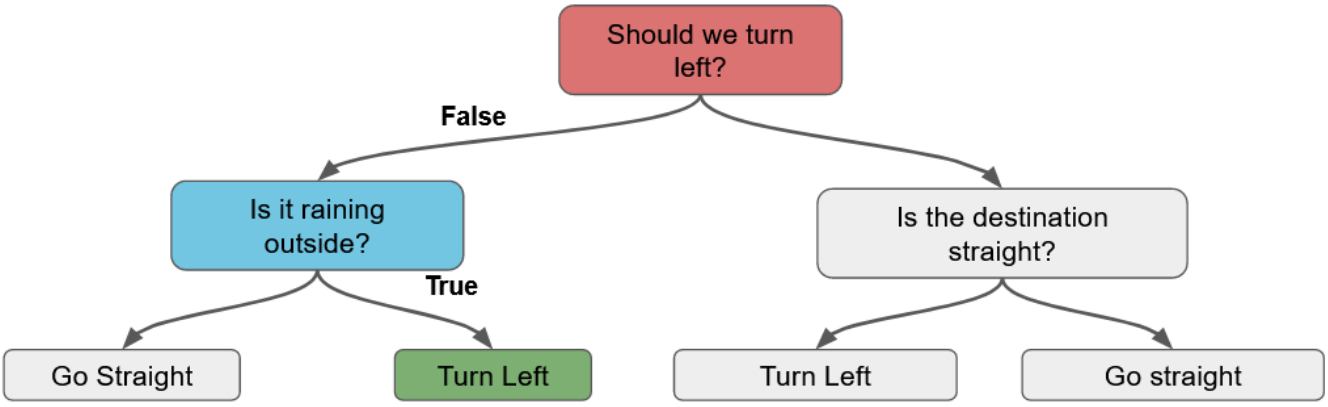


Figure 19: Incorrect decision tree example shown to participants during the instructions phase of the study.