frontiers | Frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 22 July 2024
DOI 10.3389/frobt.2024.1375490

OPEN ACCESS

EDITED BY
Raul Sena Ferreira,
Continental, France

REVIEWED BY
Koorosh Aslansefat,
University of Hull, United Kingdom
Jordan Litman,
University of Maine at Machias, United States
Farhood Negin,
Continental, France

*CORRESPONDENCE
Pradyumna Tambwekar,
pradyumna.tambwekar@gatech.edu

RECEIVED 23 January 2024
ACCEPTED 28 May 2024
PUBLISHED 22 July 2024

CITATION
Tambwekar P and Gombolay M (2024), Towards reconciling usability and usefulness of policy explanations for sequential decision-making systems.
Front. Robot. AI 11:1375490.
doi: 10.3389/frobt.2024.1375490

Towards reconciling usability and usefulness of policy explanations for sequential decision-making systems

Pradyumna Tambwekar* and Matthew Gombolay

School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, United States

Safety-critical domains often employ autonomous agents which follow a sequential decision-making setup, whereby the agent follows a policy to dictate the appropriate action at each step. AI-practitioners often employ reinforcement learning algorithms to allow an agent to find the best policy. However, sequential systems often lack clear and immediate signs of wrong actions, with consequences visible only in hindsight, making it difficult to humans to understand system failure. In reinforcement learning, this is referred to as the credit assignment problem. To effectively collaborate with an autonomous system, particularly in a safety-critical setting, explanations should enable a user to better understand the policy of the agent and predict system behavior so that users are cognizant of potential failures and these failures can be diagnosed and mitigated. However, humans are diverse and have innate biases or preferences which may enhance or impair the utility of a policy explanation of a sequential agent. Therefore, in this paper, we designed and conducted human-subjects experiment to identify the factors which influence the perceived usability with the objective usefulness of policy explanations for reinforcement learning agents in a sequential setting. Our study had two factors: the modality of policy explanation shown to the user (Tree, Text, Modified Text, and Programs) and the "first impression" of the agent, i.e., whether the user saw the agent succeed or fail in the introductory calibration video. Our findings characterize a preference-performance tradeoff wherein participants perceived language-based policy explanations to be significantly more useable; however, participants were better able to objectively predict the agent's behavior when provided an explanation in the form of a decision tree. Our results demonstrate that user-specific factors, such as computer science experience ($p < 0.05$), and situational factors, such as watching agent crash ($p < 0.05$), can significantly impact the perception and usefulness of the explanation. This research provides key insights to alleviate prevalent issues regarding innapropriate compliance and reliance, which are exponentially more detrimental in safety-critical settings, providing a path forward for XAI developers for future work on policy-explanations.

KEYWORDS

explainable AI (XAI), human-factors, reinforcement learning, Interpretability, user study

1 Introduction

There is a widening chasm between AI-practitioners and consumers due to the ever-expanding breadth of Artificial Intelligence (AI) systems. This rift between end-user and technology leads to a decrease in trust and satisfaction in autonomous systems (Matthews et al., 2019). Humans understandably become suspicious towards these systems and are less tolerant to failures and mistakes (Robinette et al., 2017; Kwon et al., 2018; Das et al., 2021). Explainable AI (XAI) was thus proposed as a means for developers to engender greater confidence in these systems by enabling users to understand the inner-workings and decision making process of AI algorithms (Xu et al., 2019; Jacovi et al., 2021). As such, XAI systems have now been broadly deployed in various capacities such as for banking (Grath et al., 2018), healthcare (Pawar et al., 2020), robotics (Anjomshoae et al., 2019) among other domains.

Many AI systems within these domains leverage a sequential decision making setup, where the agent follows a policy which sequentially dictates the action it will take at any given state in the environment. Explainable AI for sequential decision making systems raises different challenges compared to single-interaction tasks, such as decision support (Chakraborti et al., 2017a). User experiences with sequential decision making systems often involve repeated interactions. Furthermore, many applications within sequential decision making involve human-supervisory control in which humans provide feedback or demonstrations to change an agent's behavior (Griffith et al., 2013; Ravichandar et al., 2020), which require the user to iteratively update their feedback based on the new behavior of the sequential agent. Explanations for such interactions need to take into account the continuously shifting mental models of users, and provide explanations in the context of the agent's behavior in various scenarios.

For users who work with these systems, the stakeholder may have varying degrees of understanding of the agent's behavior in different contexts. An inappropriate understanding of an AI agent's behavior can have a regressive effect on AI-safety through creating a false sense of security (Ghassemi et al., 2021) and encouraging blind compliance (Poursabzi-Sangdeh et al., 2021) to uninterpretable behavior. Analyzing the mental models (Gentner and Stevens, 2014) of end-users has become a popular method of gauging a user's understanding of an autonomous system. The role of an explanation is to reconcile any differences in the user's mental model of the system with the actual conceptual model of the system, which in the case of sequential decision making systems is the *policy*. Recent work has shown the importance of mental models in user interactions with a sequential decision making agent by utilizing a formulation called "critical states," wherein the actions at these states encapsulated the essence of the policy (Huang et al., 2018). The authors showed that by presenting the actions of an agent at these critical states, a user is able to better identify the quality of two policies. Anderson et al. (2020) similarly study mental models for sequential agents, through a qualitative measurement of the accuracy of a user's mental model of the agent and the information a user utilizes to make a prediction. In our work, we utilize a *post hoc* plan-prediction task, in which we measure how often a participant is able to correctly predict an agent's behavior for the next few actions.

To avoid liability and mistrust between human-stakeholders and their AI-partners, we need to promote "explanatory debugging" (Kulesza et al., 2015) of these systems, so that humans can adequately simulate an agent's behavior and debug any faults. In this paper, we present a user study in which we compare multiple modalities of policy explanations with regards to the simulatability (Belle and Papantonis, 2021) of a sequential decision making agent. Our study focuses on an interpretable architecture, called differentiable decision trees (DDT), which were originally developed by Suárez and Lutsko (1999), and recently adapted to reinforcement learning as policy learners (Silva and Gombolay, 2020). DDTs are of interest to us due to the "white-box" nature of the architecture wherein the explanation is derived faithfully from the decision making process of the agent. Through DDTs, the actual policy learnt by an agent can be distilled as a predicate-based decision tree (Silva et al., 2020). In our study, we analyze the utility of decision-tree-based policy explanations in relation to other policy-explanation modalities such as language or programs. Language explanations are formatted as a paragraph or set of sentences, and programs are a set of if-else statements. Crucially, the information is presented in a different format but is internally consistent with the decision making process of the agent. We developed a forward simulation protocol (Doshi-Velez and Kim, 2017) in which we tested a participant's ability to interpret four modalities of explanations for a self-driving car on a highway, to build on prior work on mental models in the same domain (Huang et al., 2018; Huang et al., 2019). Driving is an accessible domain, as people generally have a model of how a car should drive on a highway, making it an interesting task to measure how well an explanation is able to shift this model to the car's actual policy. Autonomous driving is also a safety-critical application and thus is a highly relevant domain for explanatory debugging due to the ethical and liability concerns involved (Stilgoe, 2019; Zablocki et al., 2021). Therefore, it is important to better understand the factors that influence the perception and ability to apply the explanations in these scenarios. Our work also seeks to unpack the relationship between perceived usefulness of an explanation and actual usefulness, which we define as how well a participant is able to apply the explanation towards predicting the behavior of the AI agent.

Through our analysis, of subjective and objective metrics of explanation usefulness, we seek to present a better understanding of how to connect users to the right explanation which suits their individual context and characteristics. We identify key demographic factors that elucidate when an XAI method is more or less helpful for an individual user. Our analysis highlights issues regarding a lack of internal evaluative consistancy of XAI modalities by demonstrating that users objectively better understand the underlying working of the self-driving car with the help of an explanation, but subjectively prefer a different modality because the first explanation was ill-fitting towards their distinct disposition Zhou et al. (2022). To summarize, our contributions are as follows,

1. We present a novel study design to compare multiple modalities of explanations through both subjective metrics of usability and acceptance as well as objective metrics of simulatability.
2. We conduct qualitative and quantitative analysis on data from 231 participants to elucidate individual preferences of explanation modality as well as highlight the effect of

situational or dispositional factors on the perception of the XAI agent.

3. Our results highlight a lack of consistency in evaluative preference of explanation modalities, by showing that although participants rated text-based explanations to be significantly more useable than the decision tree explanation ($p < 0.05$), the decision tree explanation was found to be significantly more useful for simulating the functionality of the self-driving car ($p < 0.001$).

2 Related work

2.1 Explainable AI methodologies

Explainable AI is a prominent area of research within artificial intelligence. The most prevalent explainability methods are model-based approaches, which seek to explain the black-box of a deep neural network. A popular preliminary approach was by visualizing the outputs and gradients of a deep neural network (Simonyan et al., 2013; Yosinski et al., 2015; Selvaraju et al., 2017; Ghaeini et al., 2018). These methods provided informative visualizations of neural network outputs and parameters in order to enable users to interpret the functionality of the network. However, it has been found that approaches that rely on visual assessment can sometimes be misleading, as they may be specific to unique data or modelling conditions, and can be highly susceptible to outlying outputs that contradict the explanation (Adebayo et al., 2018; Kindermans et al., 2019; Serrano and Smith, 2019). Prior work has also sought to transform uninterpretable deep networks into interpretable architectures or modalities such as decision trees (Humbird et al., 2018; Silva and Gombolay, 2020; Paleja et al., 2021), or bayesian rule lists (Letham et al., 2015), and generate explanations by exploiting the "white-box" nature of these architectures (Silva et al., 2020).

Other researchers focus on generating human-centered explanations which describe the actions of an agent in human-understandable language. One such approach is rationale generation which present *post hoc* explanations which rationalize the actions taken by an agent in a human-understandable manner (Ehsan et al., 2019; Das et al., 2021). Susequent human-centered AI work builds on rationalizing individual actions, by also providing a set of suggested actions to enable the user to understand how to achieve their specified goal (Singh et al., 2023). In instances where data is presented in a format understandable to an end-user, an elegant solution is to highlight individual training examples which influence the model to expose the reasons behind a model's output. Prior work has enabled approaches to identify and visualize individually the effect of training examples on the hidden representations of a neural network, and have applied these methods towards explaining the network or understanding the source of bias (Koh and Liang, 2017; Silva et al., 2022a). Alternative data-based explainability methods have also provided methods to highlight the sections of the training example which provide a reasoning for an output (Mullenbach et al., 2018; DeYoung et al., 2020; Lakhotia et al., 2021). Finally, recent work seeks to adapt the explanation to the needs or preferences of the user. Such approaches modify the explanation by eliciting user-inputs (Lai et al., 2023) or by learning an embedding to encode a user's preferences or performance (Li et al., 2023; Silva et al., 2024).

2.2 Explainable Reinforcement Learning

Autonomous agents deployed in safety-critical settings, often follow a sequential decision making paradigm wherein the agent learns a policy to determine the appropriate action at each state. Due to the distinctive nature of a sequential decision making tasks, explanations in this domain have varying structures and properties. Explanations for sequential-decision making algorithms are broadly categorized as Explainable Reinforcement Learning (XRL). XRL approaches often seek to reconcile the inference capacity or the mental model (Klein and Hoffman, 2008) of a user. Inference reconciliation involves answering investigatory questions from users such as "Why not action a instead of a'?" (Madumal et al., 2020; Miller, 2021; Zahedi et al., 2024), or "Why is this plan optimal?" (Khan et al., 2009; Hayes and Shah, 2017). Other instance-based methods seek to provide the user with an explanation to elucidate the important features or a reward decomposition to enable a user to better understand or predict individual actions of a sequential decision making agent (Topin and Veloso, 2019; Anderson et al., 2020; Das et al., 2023a) Model reconciliation approaches format explanations to adjust the human's mental model of the optimal plan to more accurately align it with the actual conceptual model of the agent (Chakraborti et al., 2017b; Sreedharan et al., 2019). The last important category of XRL is policy summarizations or highlights (Amir et al., 2019; Huang et al., 2019; Lage et al., 2019; Sequeira and Gervasio, 2020). These approaches describe the functionality of an AI agent, through intelligently selected example trajectories or visualizations.

Within the set of XRL approaches, the format of explanations that our study focuses on are "global" policy explanations, wherein we explain the policy as a whole to the user rather than explaining at the action-level. A prevalent global explanation methodology is "policy trees," wherein the agent explains the policy of the user in the form of a tree. A popular methodology to generate policy trees is through distilling a learned policy into a soft decision-tree (Wu et al., 2018; Coppens et al., 2019). However, these distallation approaches have a critical flaw: the policy trees may not represent the actual policy of the agent, but merely an understandable approximation (Rudin, 2019). To resolve this issue, recent work utilize a differentiable decision tree Suárez and Lutsko (1999) to learn and visualize the actual policy of the RL-agent (Silva and Gombolay, 2020; Paleja et al., 2023). In this work, we analyze the usability and usefulness of these policy-trees as explanations, in the context of other "global" explanation baselines, for explaining policies of a self-driving car in a highway-driving domain. In this work, we do not present a novel XAI methodology. Rather, we seek to better understand the utility of policy trees towards user-preference and performance while working with a sequential decision-making agent, in order to safeguard from the dangers of innapropriate compliance with an autonomous agent.

2.3 Evaluating explainability

With a greater focus placed on XAI systems, facilitating a means of evaluating the effectiveness and usability of these approaches has become increasingly important. *Human-grounded evaluation* (Doshi-Velez and Kim, 2017) is a popular methodology

to evaluate the usefulness of proposed approaches within simulated interactions. Human-grounded evaluation seeks to understand the perception of XAI systems and the aspects of the user-experience which can be improved to facilitate smoother interactions with such autonomous agents (Booth et al., 2019; Ehsan et al., 2019; Tonekaboni et al., 2019; Madumal et al., 2020). A common practice in human-grounded evaluation is to leverage the principle of mental models (Klein and Hoffman, 2008), wherein researchers attempt to reconcile the differences between the mental model of a user with the conceptual model being explained to measure how well the XAI method explains the agent's model (Hoffman et al., 2018; Bansal et al., 2019). This is typically measured by a post-explanation task or description which attempts to understand how much the explanation has helped the user learn to better understand the AI agent's decisions (Madumal et al., 2020; Zhang et al., 2020; Kenny et al., 2021; Silva et al., 2022b; Brachman et al., 2023). Our user study employs a similar task prediction methodology which reconciles a user's understanding of the self-driving car by asking participants to predict the actions of the car before and after receiving an explanation to measure the effect of an explanation on the accuracy of their predictions. We incorporate confidence ratings to each prediction question to develop a weighted task prediction metric for each participant.

To subjectively evaluate the perception of an XAI methodology, researchers have primarily applied the Technology Acceptance Model (TAM) (Davis, 1989). Many prior XAI surveys have employed this model to study the willingness of an individual to accept an XAI agent, through metrics such as ease-of-use, usefulness, intention to use, etc. (Ehsan et al., 2019; Conati et al., 2021). Another popular avenue of studying acceptance is through the items of trust and satisfaction. In prior work, Hoffman et al. (2018) present a trust scale which predicts whether the XAI system is reliable and believable. Recent work also formalizes a new human-AI trust model and emphasizes why "warranted" trust is an important factor in XAI acceptance (Jacovi et al., 2021). In this paper, we follow these two lines of analysis by leveraging a validated survey which combines the TAM model for usability with trust to understand the participants' subjective perception of our XAI agents.

Finally, the last avenue of related work that needs to be covered is studies which pertain to the impact of personality factors on a user's interaction with an explainable system. Recent work highlights the effects of differing XAI modalities on human-AI teaming with respect to subjective and objective metrics (Silva et al., 2022b). Their results suggest that explainability alone does not significantly impact trust and compliance; rather adapting to users and "meeting users half-way" is a more effective approach for efficient human-AI teaming. Prior work has also investigated how factors such as need for cognition (Cacioppo et al., 1984), openness (Goldberg, 1990) and other personality traits impact design of explainable interfaces for recommender systems (Millecamp et al., 2019; Millecamp et al., 2020; Conati et al., 2021).

Contemporary work has also found that system, demographic and personality factors as well as the type of explanation provided can have an impact on the perceived fairness and subjective sense of understanding of an intelligent decision making system (Shulner-Tal et al., 2022a; Shulner-Tal et al., 2022b). Furthermore, contemporary work has shown that dispositional factors, such as a user's intuition regarding the various decision making pathways in a human-ai interaction, can explain some differences in reliance and usefulness of different types of explanation (Chen et al., 2023). In congruence with these work, we incorporate some important dispositional (computer science experience, learning style, etc.) and situational (car failure/success) factors into the design of our study, and seek to understand how these factors impact a user's ability to utilize an explanation.

3 Methodology

To analyze utility of different modalities of explanations describing the decision making process of a sequential AI-agent, we conducted a novel human-grounded evaluation Doshi-Velez and Kim (2017) experiment to see which explanation modality is the most helpful objectively for simulating/predicting an agent's behavior and subjectively for usability. Our study was conducted within the highway domain (Abbeel and Ng, 2004) (see Figure 1). In this environment, the car needs to navigate through traffic on a three-lane highway, where the traffic is always moving in the same direction. We chose this domain due to the easily understandable nature of the domain. Most participants would have had prior experience driving or being a passenger in a car on a highway, so they are likely to have an expectation of how to "properly" drive on a highway. This allows us to test whether we are accurately able to convey the car's decision making process, and consolidate the differences between the two. Through this study we attempt to not only understand more about explanation preferences and perceptions but also the dispositional (CS Experience, Video Game experience, learning style, etc.) and situational (success/failure) factors which influence these preferences. Specifically, our analysis sought to answer the following questions,

- **Q1**: Which explanation modality affords the greatest degree of simulatability in terms of understanding the decision making process of the car and accurately predicting the car's actions?
- **Q2**: How do individual explanation modalities impact subjective measures of usability and trust, and are these individual preferences consistent with the metric of explanation usefulness studied in Q1?
- **Q3**: Are there any interaction affects between dispositional and situational factors, e.g., computer science experience, learning style, success/failure, on the subjective and objective measures studied in this protocol?

3.1 Experiment design

The factors in our experiment were 1) Explanation Format and 2) Success-vs.-Failure video. Our experiment was a between-subjects study with a 2×4 study design.

Explanation Format–Our study compares policy trees with other "global" explanation modalities to provide an alternate means of presenting the same information in the tree. In general, explanations as policy trees can be generated by various methods. Differentiable decision trees can be initialized by users through a

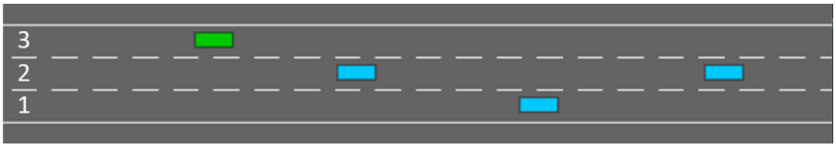


FIGURE 1
This diagram depicts the environment utilized in this study. The green car denotes the AI agent which is navigating through the highway and presenting explanations to the particpant for each action it takes.

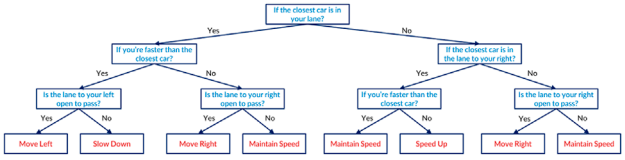
A

- If the car closest to you is in your lane, and you're going faster than it, try to move left when the left lane is clear.
- If you can't move left, then slow down.
- If the car closest to you is in your lane but you're not going faster than it, try to move into the lane on your right, when it's clear.
- If it's not clear, continue in your lane at the same speed you were at.
- If the car closest to you is in the lane to your right, and you are going faster than it, keep going at the same speed to pass the car that's in the lane to your right.
- If the closest car is to your left, and you are not going faster than it, try to move right when the lane to the right is clear.

B

- First, figure out which lane the nearest car to you is in.
  - If the nearest car is in your lane, check if you are going faster than it.
    - If you're faster than the car in front of you, check whether you can safely make a left. If you can't, slow down.
    - If you're slower than the car in front of you, check whether you can safely make a right. If not, maintain current speed.
  - If the nearest car is in the lane to your right, speed up until you are faster than it.
  - Finally, if the nearest car is in the lane to your left, try to move right if it is safe.
    - If not, stay at the current speed until the opportunity to move right arises.

C



D

```
def agent_algorithm(game_state):
    current_lane = get_current_lane()
    left_lane = get_left_lane()
    right_lane = get_right_lane()
    act = {'left': 0, 'right': 1, 'maintain speed': 2, 'slow down': 3, 'speed up': 4}
    if closest_car_lane() == current_lane:
        if faster_than_closest():
            if left_lane_open():
                agent.set_action(act['left'])
            else:
                agent.set_action(act['slow down'])
        else:
            if right_lane_open():
                agent.set_action(act['right'])
            else:
                agent.set_action(act['maintain speed'])
    else:
        if closest_car_lane() == right_lane:
            if faster_than_closest():
                agent.set_action(act['maintain speed'])
            else:
                agent.set_action(act['speed up'])
        else:
            if right_lane_empty():
                agent.set_action(act['right'])
            else:
                agent.set_action(act['maintain speed'])
```

FIGURE 2
This figure depicts the four policy explanations shown to participants corresponding to each baseline. **(A)** Basic Text: A language description generated using a template from the decision-tree policy, **(B)** Modified Text: A simplied version of the language description presented in an easy-to-understand manner, **(C)** Decision Tree: A decision tree describing the exact policy of the self-driving car, **(D)** Program: Pseudo-code of the decision making process of the car.

graphical user-interface (Silva and Gombolay, 2020), or through language descriptions of the policy (Tambwekar et al., 2023). Once these DDTs are trained, these models can be discretized to present a discrete policy tree to the user as an explanation (Silva et al., 2020). In our work, for our RL-agent's policy, we select a policy tree for our approach from a dataset of lexical decision trees in prior work (Tambwekar et al., 2023), which included 200 human-specified policies for a car in the highway domain. The policy we chose was a complete decision tree of depth three, which corresponded to the largest possible policy in this dataset. The complete set of modalities we utilize in our study, all stem from the selected policy tree. The selection of these modalities was motivated by the principles of "explanatory debugging" proposed in prior work (Kulesza et al., 2013; Kulesza et al., 2015). These works discuss balancing "soundness" and "completeness" of an explanation with the need to maintain comprehensibility. By choosing these baselines, we seek to understand the comprehensibility of four explanation modalities that are all perfectly sound and complete. We provide a description and rationale for the selection of each modality below:

Explanation Modality 1: Tree—The first explanation modality is a decision tree which represents the policy of the self-driving car. Decision trees have become a popular method of explaining decisions for human-AI teaming scenarios (Paleja et al., 2021; Wu et al., 2021; Tambwekar et al., 2023). Differential decision trees have been proven to be an interpretable method of representing

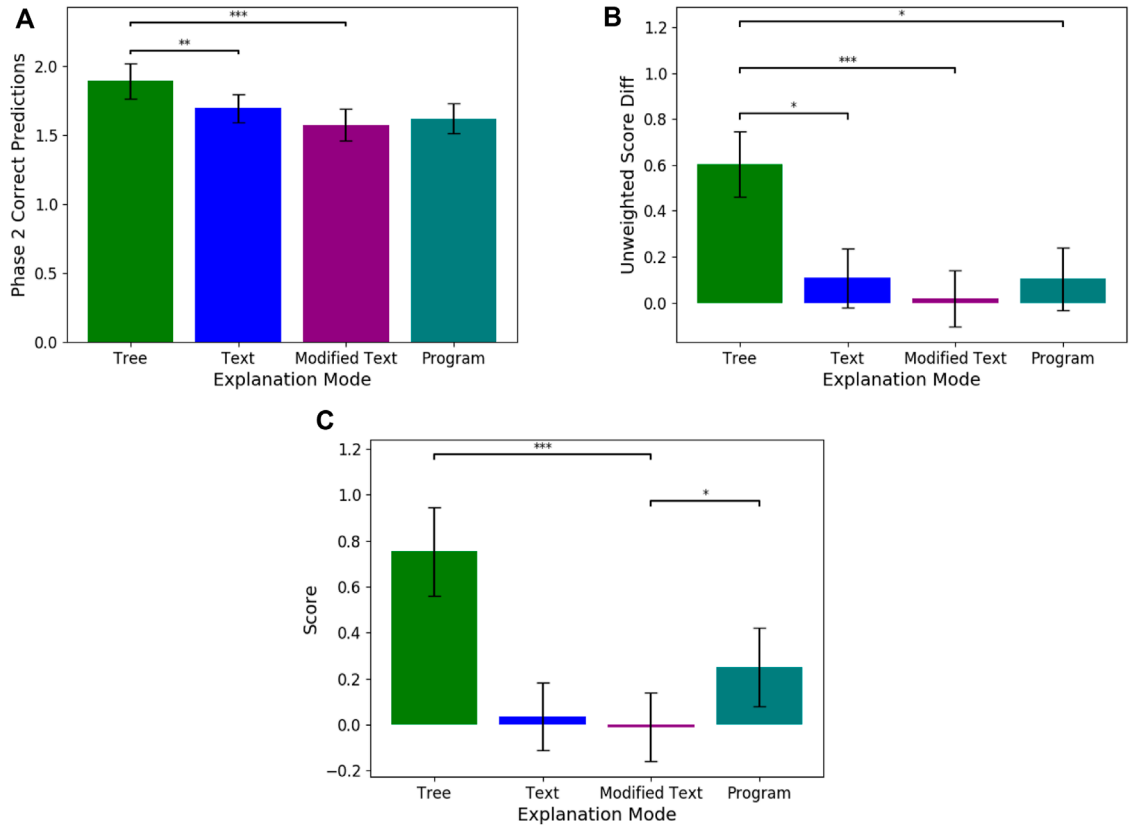


FIGURE 3
These graphs plot the means and standard errors for three objective evaluation metrics i.e., **(A)** Phase 2 Correct Predictions, **(B)** Unweighted Score Diff, and **(C)** Score, across the four explanation modalities. Significant differences between modalities is noted in the graphs. The score is computed through the equation presented in Eq. 1.

a policies that can be employed towards generating "white-box" explanations for users that actually represent the underlying behavior (Silva and Gombolay, 2020; Paleja et al., 2022; Custode and Iacca, 2023). The decision tree explanation seeks to represent explanations generated by these differentiable decision trees. Further details regarding the functionality of differentiable decision trees can be found in the appendix.

Explanation Modality 2: Basic Text—A policy description generated from the decision-tree policy using a simple text-grammar. Language has also shown to be an effective means of explaining the actions of an sequential decision making agent (Hayes and Shah, 2017; Ehsan et al., 2019; Das et al., 2021). Therefore, we wanted to ascertain whether a user could better interpret the information in the decision tree when presented in a text paragraph.

Explanation Modality 3: Modified Text—A modified version of the text description, presented in a format which is easier to parse, with simplified language and indentation. The modified text explanation seeks to improve comprehensibility of the text explanation by simplifying and rephrasing details of the original text explanation and additional formatting. By including this modality, we hope to understand whether our expectation of the comprehensibility of an explanation is reflective in improvements of actual user-comprehension.

Explanation Modality 4: Program—A set of *if-else* statements encoding the logic of the decision tree. The choice of program/rule-based explanations was to cater to scenarios wherein explainable systems are utilized to assist domain experts who wish to debug agent behavior. As computer science experience was one of the factors we were studying, we were interested in determining whether participants with CS experience were able to process the same information better as psuedo-code compared to the other modalities.

The specific explanations provided to participants in our study are shown in Figure 2. Our study follows a between-subjects study design, therefore each participant only received a single explanation modality.

Success/Failure–Prior work has studied how the nature of an explanation sways a user's ability to tolerate the agent failing (Ehsan et al., 2019). In this study, we analyze the opposite relationship, i.e., how does seeing the agent succeed or fail impact their perception of the explanation. A user's predisposition has been known to impact a user's interaction with an intelligent agent (Chen et al., 2023; Clare et al., 2015). Through showing the participant a video of the car succeeding or failing, our goal was to measure whether this had any discernible impact on the user's predisposition such that it affected the way they interacted with the policy explanation in our study. At the start of the experiment,

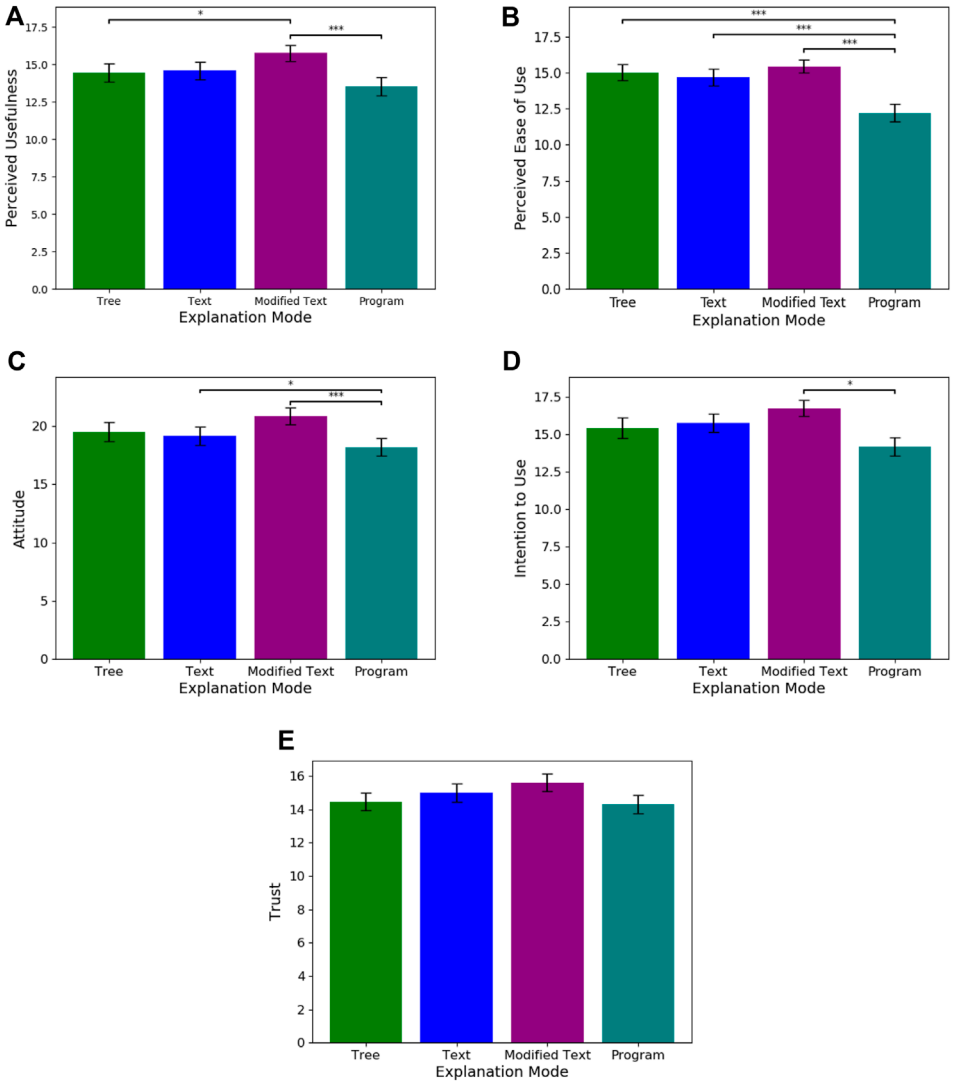


FIGURE 4
These graphs plot the means and standard errors for each subjective evaluation metric i.e., **(A)** Perceived Usefulness, **(B)** Perceived Ease of Use, **(C)** Attitude, **(D)** Intention to Use, and **(E)** Trust, across the four explanation modalities. Significant differences between modalities is noted in the graphs..

each participant was shown a 1-min video of a simulation of the car in the highway domain to help the participant build a mental model the AI's behavior. Each participant was shown one of two videos, depending on whether they were assigned to the success or failure condition. In the "failure" video, the car crashed at the end of the video, and, in the "success", the car successfully reaches a "finish line." We have provided screenshots in the appendix to depict what the participant sees at the end of each video. Both these videos were generated using the same policy for the agent. Our aim was to measure whether watching the car succeed or crash subjectively influenced participants' perception of any XAI modality or objectively impaired their ability to apply the explanation. A similar analysis of "first impressions" of the agent was done in a parallel study from contemporary work, wherein the authors found that lower decision accuracies for participants with the "failure" condition (Vered et al., 2023). Our study differs, from this prior work, in testing the impact of success/failure in the context of explanatory debugging for reinforcement learning policies rather than "single-move" explanations.

3.2 Metrics

In this section we describe the metrics we employ to subjectively gauge perception of each explanation, with regards to usability and trust, and objectively evaluate a user's ability to simulate the decision making of the car. To measure usability, we adapted a survey, which incorporated trust into the TAM model (Davis, 1989), from prior work on human-evaluation of e-service systems (Belanche et al., 2012). This survey included questions on *usability*, *ease of use*, *attitude*, *intention to use*, and *trust*. The TAM model is the predominant measurement method for the perception of any

TABLE 1 This table provides the specific items of the usability and trust questionnaire employed in this study. This survey was previously used to measure perception of "e-services." In our study, we replaced all references of "e-services" to "explainable agent."

Category	Questions
Perceived Usefulness	Using this explainable agent would be useful for me
	Using this explainable agent will improve my effectiveness
	Using this explainable agent will improve my performance
Perceived Ease-of-use	With this explainable agent, it would be easy to get the information I need
	Learning to operate with this explainable agent would be easy
	This explainable agent would be easy to use
Attitude	Using this explainable agent is an idea I like
	Using this explainable agent would be a pleasant experience
	Using this explainable agent is a good idea
	Using this explainable agent is a wise idea
Trust	I trust this explainable agent
	This explainable agent is reliable
	This explainable agent is trustworthy
Intention to use	When I will need it, I will intend to use this explainable agent rather than an agent with no explanation
	When I will need it, I predict I would use this explainable agent rather than an agent with no explanation
	When I will need it, I would like to use this explainable agent rather than an agent with no explanation

technology due to its strong correlation with technology adoption, and has recently been widely employed towards explainable-AI (Ehsan and Riedl, 2019; Ehsan et al., 2021; Bayer et al., 2022; Panagoulias et al., 2024). Our choice of survey for this study was dictated by the desire to integrate Trust into the acceptance factors measured by the standard TAM measurement survey. We replaced references to "e-service" in the original survey with "explainable agent" for this user study. The complete survey utilized in this study can be viewed in Table 1. Note that we did not employ the frequently utilized trust scale for XAI proposed in Hoffman et al. (2018), because our study did not satisfy the assumptions required as per the authors, i.e., "the participant has had considerable experience using the XAI system." In our case, participants were interacting with the XAI agent for the first time for only 20 min, therefore we ascertained that this scale was not applicable.

To measure objective simulatability, we computed a prediction score using participants' predictions before and after receiving an explanation for the car's actions. We asked participants four prediction questions where they predicted the next sequence of actions the car will take. Using their answers to these questions, we compute a task prediction score as shown in Equation 1,

$$score = \sum_{i=1}^{4} c_{a,i} \times \delta_{a,i} - \sum_{i=1}^{4} c_{b,i} \times \delta_{b,i} \quad (1)$$

In this formula, the δ parameters represent whether or not the participant was able to predict the car's actions correctly. $\delta_{a,i}$ is assigned a value of +1 if the ith question was answered correctly *after* receiving an explanation and −1 otherwise. $\delta_{b,i}$ similarly represents the correctness of the participant's prediction for the i^{th} question before receiving an explanation. $c_{b,i}$ represents the confidence rating for question, i, *before* receiving an explanation, and $c_{a,i}$ represents the confidence rating for the ith question *after* receiving an explanation. Confidence ratings are obtained by asking participants how confident they are in their prediction of the car's next sequence of actions, on a 5-item scale from "Not confident at all" to "Extremely confident." To compute $c_{b,i}$ and $c_{a,i}$, we assign numeric values to a participants confidence rating uniforming between 0.2 and 1, in increments of 0.2 (i.e., Not confident = 0.2, Slightly confident = 0.4, Moderately confident = 0.6, Very confident = 0.8, Extremely confident = 1). When combined, c and δ represent a weighted prediction score. The score variable represents the difference between the weighted prediction scores across four different prediction questions. Unlike prior performance measurements, which measure compliance or correctness in isolation, our weighted score metric enables us to incorporate confidence such that a participant is rewarded for having higher confidence in their correct predictions and *vice versa*. We also measure the unweighted score, the number of correct answers after receiving an explanation, and weighted number of

correct answers after receiving an explanation to track simulatability performance.

3.3 Procedure

This entire procedure was approved under a minimum risk exempt-protocol by our institute's IRB (Protocol H21040). This experiment was conducted online via Amazon Mechanical Turk. Our study began with a demographics survey about age, gender, education and experience with computer science and video games. Computer Science and video game experience were measured on a 4-point, self-reported scale from "very inexperienced" to "very experienced." Participants were also asked to answer short surveys regarding their orientation towards things or people (Graziano et al., 2012) and learning style (visual-vs.-verbal (Mayer and Massa, 2003)). The rest of our study is divided into three phases.

In Phase 1 of the study, the participant first received a 1-min video of the car driving on a highway, in which the car would either reach the finish line (success) or crash into another car (failure) at the end of the video. The purpose of this video is to enable the participant to build a mental model of how the self-driving car interacts with the world so as to improve their ability to predict behaviors in other scenarios. The participant could reference this video as many times as they needed throughout the study to help understand the car's behavior as it was provided at the top of each page in the study. Participants could scroll to the top of the page for each question to rewatch the video if they needed to. Next, the participant would be asked to complete *four* prediction tasks. For each prediction task, the participant was shown a unique, 8-s video of the car driving on the virtual highway. These prediction videos were selected from recordings of the driving agent to best represent the different types of behavior of the car, i.e., slow down and switch lanes, maintain the same speed in the same lane, overtake from the left, etc. Based on this video of the car, participants were asked to predict the next set of actions of the agent from a set of five options (including an option for none of the above), by utilizing their inferred mental model of the car. We chose to ask participants to predict the next *set* of actions ("move right and speed up," "maintain speed and crash into the car ahead," etc.), as this required participants to perform multiple predictions of the car's behavior for each question thereby providing a more accurate measure of how well they understood the car's behavior. Each scenario involved the car executing a policy on a different part of the highway. Each prediction question was accompanied with a 5-point confidence rating (Not confident - Extremely confident).

In Phase 2, participants would perform the same tasks as in Phase 1, with the exception being that participants also received an additional explanation for the actions of the car in one of the four formats specified earlier. Conducting the same prediction tasks with and without an explanation allowed us to directly analyze the impact of an explanation on the perception of the explainable agent. Finally, in the third and final phase, participants were asked to complete our usability and trust survey to subjectively evaluate the explanation modality they worked with.

Our study design relates to that of another study conducted by Huang et al. (2018), wherein they establish the importance of "critical states," in engendering an more representative mental model of the self-driving car's policy. In this work, the authors show that by providing examples of what the car will do in these critical states, a user is more likely to identify the superior policy between two choices. Despite establishing that critical states help build a mental model, they do not directly test whether the explanation makes the participants more likely to be able to interpret and predict the car's actions. In our study, we directly focus on the "explanatory debugging" capabilities of different kinds of policy explanations which are all equally sound and complete. By doing so, we hope to add to the existing literature in this field and better understand how the comprehensibility of a policy explanation changes with the presentation and format of the explanation.

4 Results

Our analysis was conducted on data from 231 participants, recruited from mechanical turk (54% identified as Male, 46% identified as Female, and $< 1\%$ identified as Non-binary/other). Out of all responses collected, we only included responses in our final dataset from participants who had submitted the survey once. To the best of our knowledge, all responses that were from repeat or malicious responders were filtered out. A total of 46 participants reported having some degree of computer science experience. The average time taken for our survey was 18 min and participants were paid $4 for completing our study (which equates to $13.34 per hour).

We created a multivariate regression model with the explanation mode, success/failure and demographics values as the independent variable, with the dependent variable being the subjective or objective metric being studied. Each regression model was checked to meet normality, via the Kolgomogorov-Smirnov Test and homoscedasticity via the Breusch-Pagan Test, and we applied non-parametric tests to analyze models that did not pass these assumptions (Table 2). Models were checked to meet normality and homoscedasticity assumptions. Omnibus tests were performed before pairwise comparisons were made. We used multivariate linear regression with AIC as our occam's razor for modelling covariates and interaction effects. We chose linear regression over its non-parametric alternatives as linear regression is a straightforward approach which effectively reveals the salient relationships between the independent and dependent variables.

4.1 Research question 1

For Q1 (understanding the simulatability of each individual modality), we compared how each explanation mode affected the task prediction performance using our objective metrics (See Figure 3). We find that tree explanations were significantly more beneficial for predicting more questions correctly in phase 2 when compared to modified (Estimate = −1.257, SE = 0.374, $p < 0.001$) and basic text (Estimate = −1.138, Standard Error (SE) = 0.3548, $p < 0.01$). After taking into account confidence ratings, the usage of tree explanations still significantly improved weighted number of correct answers in Phase 2 as compared to the modified text explanations

TABLE 2 This table details the independent variable, dependent variable and covariates for each model. We have also listed down the assumptions of the ANOVA test and transforms applied. Note that for the last column, we report the *p*-values for Breusch-Pagan test, which is a heteroscedasticity test. Therefore $p > 0.05$ implies that the models pass the homoscedasticity assumption.

DV	Transform	Significant covariates	Normality	Heteroscedasticity
Usefulness	boxcox	Explanation, education	$p < 0.05$	$p > 0.05$
Ease of Use	N/A	Explanation, education	$p < 0.05$	$p > 0.05$
Attitude	boxcox	Explanation, education, success	$p < 0.05$	$p > 0.05$
Intention to Use	boxcox	Explanation	$p > 0.05$	N/A
Trust	boxcox	N/A	$p > 0.05$	N/A
Weighted Correct Answers Phase 2	N/A	Explanation, Weighted Correct Phase 1	$p < 0.05$	$p > 0.05$
Correct Questions in Phase 2	N/A	Explanation, Correct Questions in Phase 1	$p < 0.05$	$p > 0.05$
Score	N/A	Explanation	$p < 0.05$	$p > 0.05$
Unweighted Score	N/A	Explanation, gender	$p < 0.05$	$p > 0.05$

(Estimate = −0.571, SE = 0.167, $p < 0.001$), and the basic text-based explanations (Estimate = −0.480, SE = 0.160, $p < 0.01$). For the score metric described earlier, both trees (Estimate = 2.517, SE = 0.558, $p < 0.001$) and programs (Estimate = 1.414, SE = 0.555, $p < 0.05$) significantly improved the participant's score when compared to the modified text baseline. **These results imply that users are able to more accurately simulate and understand an agent's decisions using trees.**

4.2 Research question 2

With respect to Q2 (understanding the perceived usability of individual modalities), we found that modified text was rated to be significantly more useful than both the program (Estimate = 47.6434, SE = 12.484, $p < 0.001$) and the tree (Estimate = 29.098, SE = 12.470, $p < 0.05$) baselines (See Figure 4). For ease of use, the tree, text, and modified text baselines were rated significantly higher than the program explanation ($p < 0.001$). For the metric of intention to use, a Wilcoxon signed rank test showed that modified text was preferred to program ($p < 0.05$). These results suggest an inconsistency between the subjective and objective evaluation metrics for the decision tree and program vs. text-based modalities. **Although they were found to be less useful for accurately predicing the actions of the car, participants perceived text-based explanations as significantly more useable than decision trees and programs.** We applied a non-parametric Wilcoxon-Signed Rank Test to analyze trust, however, none of the explanation modalities were found to significantly impact trust.

4.3 Research question 3

In relation to Q3 (understanding the effect of individual factors on the subjective and objective measures of each modality), we found that participants with low CS experience have significantly improved relative prediction scores when using the modified text explanation as compared to the tree (Estimate = −2.1597, SE = 0.611, $p < 0.001$) or program (Estimate = −1.436, SE = 0.609, $p < 0.05$) explanations. For usefulness, higher CS experience significantly decreases the relative advantage of text over program for both the basic (Estimate = −14.12, SE = 6.335, $p < 0.05$) and modified text (Estimate = −19.69, SE = 6.597, $p < 0.01$) modalities. Similarly, with respect to attitude and ease-of-use, high-CS experience was found to significantly decrease the preference of modified text ($p < 0.01$) and text ($p < 0.05$) explanations relative to programs. **Overall, our results showed that with respect to simulatability and usability, increasing CS experience negatively impacts the text-based explanations compared to the program or tree explanations.**

4.4 Additional results -

We note that we did not find self-reported learning-style preference (visual vs. verbal) to be a significant influencing factor for either the subjective or objective measures we studied. Next we studied whether success and failure were found to significantly influence XAI perception. With respect to success, we found that watching the car succeed in the priming video–as opposed to failing, i.e., crashing–significantly improved a participant's attitude (Estimate = 11.986, SE = 4.64, $p < 0.05$). However, a similar effect was not observed for the other dependent subjective variables, i.e., ease-of-use, usefulness and intent-to-use. This implies that although, on average, participants felt that working with the XAI agent that failed was "unpleasant", it did not impact their usability. This may indicate that better care needs to be taken to appease end-users in situations where they work with agents that frequently fail. Unlike in the case of attitude, success/failure was not found to affect the score of a participant, i.e., watching the car fail did not affect the participants ability to understand the explanation.

From the results of an ANOVA test on a linear regression model, success was found to be extremely important. However, our trust model did not satisfy the normality assumptions of our parametric linear regression test. Prior work has shown that an F-test can be robust to the normality assumption (Cochran, 1947; Glass et al., 1972; Blanca et al., 2017). Therefore, while we cannot conclude that success significantly impacts trust, it does appear to be a important factor with respect to trust.

5 Discussion

In our study, we found inconsistency in human preferences of explanation modalities with respect to subjective and objective metrics. Participants found language-based explanations to be significantly more useful ($p < 0.001$) even though participants performed better according to our objective metrics when using the tree-based explanation ($p < 0.001$). Prior work has often reported a significant difference between the performance of a stakeholder with and without an explanation. Explanations have been shown to improve situational awareness (Paleja et al., 2021), task-accuracy (Das et al., 2023a) and error-avoidance (Das et al., 2023b). However, a stakeholder's ability to utilize an explanation to improve their "performance" is more nuanced than a simple binary relationship (Poursabzi-Sangdeh et al., 2021). Explanations are not universally benefitial; Sometimes, providing an explanation begets over-reliance in the intelligent system leading to instances of inappropriate compliance (Ehsan and Riedl, 2021; Silva et al., 2022b). A contemporary study with neurologists showed that more "explainable" methodologies may disrupt or hamper a neurologists decision-making processes (Gombolay et al., 2024). Our findings augment these prior works by further motivating the need for human-centered or user-centered perspectives to explainability which consider a user's situational or dispositional factors (Ehsan and Riedl, 2020; Liao et al., 2020; Dhanorkar et al., 2021). Participants' preference towards using modes of explanation which objectively perform poorer on task performance metrics is a clear indicator that explanations need to consider the individual dispositions of the potential end-user to engender adoption.

When explanations are ill-fitting of an individual's dispositional or situational circumstances, users may be unable or unwilling to utilize the explanation to understand the decision making of the car. For example, we found, through our post-survey feedback, that participants with little or no programming experience were often discouraged and confused by the program-based explanation. One participant stated that the explanation was counter-productive in that it made the participant "second guess [their] initial choice," and further stated that "If it was supposed to be reassuring and confirming, it was not." Another participant stated that the nature of the program-based explanation made it "functionally useless" to the task assigned. Other participants took issue with the nature or structure of the explanation. One participant stated in reference to the modified-text explanation that, "Some of the sentences could have been combined and just said left or right instead of having a statement for each." Another participant stated that they were "better with visual learning," and, therefore, preferred to go by their initial assumptions based on the video rather than use the text explanation, thereby ignoring the explanation altogether.

Humans create mental models for systems they interact with Hoffman et al. (2018), that encapsulate their understanding of how the agent functions. These mental models often contain misconceptions or misinterpretations, and it is the job of the explanation to satisfactorily consolidate the user's mental model. In order to effectively do so, the explanation needs to be presented to the user in a manner which caters to their unique socio-technical disposition Sokol and Flach (2020). As seen by our findings, a simple factor such as computer science experience can significantly affect a user's ability to employ an explanation to understand functionality of the car. One participant's response encapsulates this sentiment regarding mental models: After receiving the decision tree the participant stated, the explanation "was generally helpful in that it helped [the participant] focus on the other car that was the biggest factor in the AI's decision making." This indicated that the participant was able to apply the explanation to improve their mental model of the car's behavior, by identifying the factors in the environment that influence the car's decisions. Another participant stated that the modified-text explanation "really helped" because "it showed how to see the car and how it would interact with the world around it." In both these situations, the user was more open to adopting the explanation because the explanation was able to satisfactorily fill in the gaps in their mental model of the car, by helping participants perceive how the car may be processing the information available in the environment to make decisions.

Overall, our results support the position that researchers should design personalized XAI interfaces which can cater to the social needs of the end-users interacting with these systems. We do not claim to be the first to show that Personalized XAI is necessary, which has already been shown in recent work (Millecamp et al., 2019; Millecamp et al., 2020). However, these works are restricted to recommendation/tutoring systems. A highly relevant recent study developed a personalized explainable-AI methodology such that the AI-assistant can present the users with explanations that balance their preferences and performance (Silva et al., 2024). Crucially, they showed that a balanced personalization method lead to significantly fewer instances of innapropriate compliance than personalizing based on preference alone. Our analysis identifies key demographics factors which can be integrated into such personalized xAI methodologies, such as computer science experience, and highlights the importance of these factors with respect to subjective perception and objective use of XAI modalities.

6 Limitations and future work

Firstly, our study follows a human-grounded evaluation structure we performed our analysis on for a simulated self-driving car. Therefore, it is important to acknowledge that while these results provide a comprehensive initial estimate, they may vary when this study is replicated on the real task. It should also be emphasized that in addition to real-world transfer, an additional limitation pertains to the generalizability of our findings beyond our preliminary experimental setup. We expect our results to generalize to other sequential decision-making domains as these results fundamentally concern the formatting of policy explanations widely employed in recent work, however, every domain/application possesses unique

intricacies which may affect the trends found in this study. Future work could benefit from a similar study in a real setting, i.e., application grounded evaluation, which accurately reproduces the experience of receiving explanations in the real world and measures whether our results will generalize to differing experimental setups. Interesting domains to conduct a similar study, to both understand the generalizability of our results and real-world transfer, would be in a state-of-the-art self-driving car simulator (Schrum et al., 2024) or an in-home robot setting (Szot et al., 2021; Patel et al., 2023).

Secondly, this study does not consider longitudinal human-adaptation to XAI systems. It may be possible that although a participant initially does not prefer to use an explanation after brief interactions with the agent, a period of time to adapt to the explanation modality may alter their preference. In future work, it would be interesting to setup the study as a multi-day experiment where the participants work with the explainable agent on a series of subtasks (one per day) to measure whether there are any longintudinal factors which impact their behavior. These subtasks could increase in difficulty to account for the user's adaptation to the system to ensure that the user still needs explanations.

Another important limitation to consider is the participant's level of immersion. Viewing the simulation of the car in our environment may not be enough for the participant to recognize the consequences of working with a self-driving car. In a situation where the stakes are more obvious, participants may perceive explanations differently. This may have contributed towards the lack of significance for the trust model. Since participants may not have been immersed/understood the potential real-world consequences, their internal model for trust may have been independent to the explanation provided. However, we believe that our study still contributes novel insights that provide a stepping stone towards a future application-grounded analysis. Future work could leverage attentional or physiological measurements of immersion to understand whether immersion correlates with a participant's perception of preference or their performance on the task (Hagiwara et al., 2016; Hammond et al., 2023).

Our analysis also found that computer science experience influences end-user perception and preference in our sequential domain. However, a relevant limitation of our approach is that CS experience was measured on a self-reported scale. Therefore, a participant's inherent biases and experiences may impact their own ratings of their computer science experience. In future work, we hope to incorporate a quantitative measurement of CS experience in the form of a short competency quiz. Furthermore, beyond CS experience, there may be additional experiential or dispositional factors which may impact our dependent variables. One important dispositional factor that we did not consider in our study is epistemic curiosity, or a user's "general desire for knowledge" (Hoffman et al., 2023). The two categories of curiosity are I-type curiosity, which is triggered by subjective feelings of situational interest, and D-type curiosity which is triggered by violated expectations or missing information (Litman, 2019). Recent work has found that individual differences in these curiosity types have an impact in the relative utility and value of information in various organizational settings (Lievens et al., 2022). These two dimensions of curiosity dovetail well with our paradigm of subjective and objective perception of XAI methods, and therefore, are items we hope to incorporate in future work. Another important cognitive factor to incorporate could be an individual's "Need for Cognition," which measures an individual's tendency to engage in or enjoy effortful cognitive activity (Cacioppo et al., 1984). There may be a correlation between a user's "Need for Cognition" and their ability to effectively process a given explanation, therefore making it a relevant factor to include in future studies. In future work, we aim to leverage the insights from this study to develop a personalized XAI methodology. In such a case, the XAI interface could evaluate an individual's disposition and demographic factors to recommend a type of explanation. Then the user can specify any additional properties they would like within the explanation and adaptably modify the type of explanations it receives from the agent. We hypothesize that such an approach would truly give rise to human-centered explainability and bridge the gap between stakeholders and AI technology.

7 Conclusion

Explainable AI must have a stakeholder-focus to engender long-term adoption. Simply unraveling the internal mechanisms of an Artificial Intelligence agent is insufficient if it is not presented in a way the end-user can easily understand. To produce user-centered XAI approaches, we need to better understand what influences XAI perception. In this paper, we present a novel user-study which studies subjective user-preference towards disparate XAI modalities, for a sequential decision-making system such as a self-driving car, and how situational (e.g., watching the car succeed or fail) and dispositional factors (e.g., computer science experience) influence this perception. We show that computer science experience can reduce an individual's preference towards the text-based modalities, as well as how watching the car fail (crash into another car) worsens their attitude towards the XAI agent. Our findings also highlight an important internal inconsistency in explanation preference. Text-based explanations were perceived to be more useable according to our subjective survey, however, decision tree explanations were found to be more useful in terms of more accurately predicting the car's actions. XAI developers need to balance the tradeoff between willingness to adopt and usefulness, as the perceived usability varies based on an individuals specific intrinsic and situational criteria. We hope that this work promotes a wider study of personalized XAI approaches which curate explanations to fit the particular needs and circumstances of individual stakeholders.

Data availability statement

The datasets presented in this article are not readily available because the data collected through our study is confidential as per our IRB protocol. Requests to access the datasets should be directed to PT, pradyumna.tambwekar@gatech.edu.

Ethics statement

The studies involving humans were approved by Georgia Institute of Technology Institutional Review Board. The studies were

conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

PT: Conceptualization, Data curation, Investigation, Methodology, Writing - original draft, Visualization, Writing–review and editing, MG: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by MIT Lincoln Laboratory under grant 7000437192, the NASA Early Career Fellowship under grant 80HQTR19NOA01-19ECF-B1, the National Science Foundation under grant CNS-2219755, the Office of Naval Research under grant N00014-23-1-2887, and a gift by Konica Minolta, Inc. to the Georgia Tech Research Foundation. Konica Minolta, Inc. was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2024.1375490/full#supplementary-material

References

Abbeel, P., and Ng, A. Y. (2004). "Apprenticeship learning via inverse reinforcement learning," in Proceedings of the twenty-first international conference on machine learning (ACM), Banff, Alberta, Canada, 1.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* 31. doi:10.5555/3327546.3327621

Amir, O., Doshi-Velez, F., and Sarne, D. (2019). Summarizing agent strategies. *Aut. Agents Multi-Agent Syst.* 33, 628–644. doi:10.1007/s10458-019-09418-w

Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., et al. (2020). Mental models of mere mortals with explanations of reinforcement learning. *ACM Trans. Interact. Intell. Syst.* 10, 1–37. doi:10.1145/3366485

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019). "Explainable agents and robots: results from a systematic literature review," in 18th international conference on autonomous agents and multiagent systems (AAMAS 2019), Montreal, QC, May 13–17, 2019 (International Foundation for Autonomous Agents and Multiagent Systems), 1078–1088.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. (2019). "Updates in human-ai teams: understanding and addressing the performance/compatibility tradeoff," *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 2429–2437. doi:10.1609/aaai.v33i01.33012429

Bayer, S., Gimpel, H., and Markgraf, M. (2022). The role of domain expertise in trusting and following explainable ai decision support systems. *J. Decis. Syst.* 32, 110–138. doi:10.1080/12460125.2021.1958505

Belanche, D., Casaló, L. V., and Flavián, C. (2012). Integrating trust and personal values into the technology acceptance model: the case of e-government services adoption. *Cuad. Econ. Dir. Empres.* 15, 192–204. doi:10.1016/j.cede.2012.04.004

Belle, V., and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Front. Big Data* 39, 688969. doi:10.3389/fdata.2021.688969

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., and Bendayan, R. (2017). Non-normal data: is anova still a valid option? *Psicothema* 29, 552–557. doi:10.7334/psicothema2016.383

Booth, S., Muise, C., and Shah, J. (2019). Evaluating the interpretability of the knowledge compilation map: communicating logical statements effectively. *IJCAI*, 5801–5807. doi:10.24963/ijcai.2019/804

Brachman, M., Pan, Q., Do, H. J., Dugan, C., Chaudhary, A., Johnson, J. M., 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, Sydney, NSW, Australia, 220–239.

Cacioppo, J. T., Petty, R. E., and Feng Kao, C. (1984). The efficient assessment of need for cognition. *J. personality Assess.* 48, 306–307. doi:10.1207/s15327752jpa4803_13

Chakraborti, T., Sreedharan, S., and Kambhampati, S. (2017a). Balancing explicability and explanation in human-aware planning. arXiv Preprint arXiv:1708.00543. doi:10.24963/ijcai.2019/185

Chakraborti, T., Sreedharan, S., Zhang, Y., and Kambhampati, S. (2017b). Plan explanations as model reconciliation: moving beyond explanation as soliloquy. arXiv Preprint arXiv:1701.08317. doi:10.5555/3171642.3171666

Chen, V., Liao, Q. V., Wortman Vaughan, J., and Bansal, G. (2023). "Understanding the role of human intuition on reliance in human-ai decision-making with explanations," Proceedings of the ACM on human-computer interaction, October 2023 7, 1–32. doi:10.1145/3610219

Clare, A. S., Cummings, M. L., and Repenning, N. P. (2015). Influencing trust for human–automation collaborative scheduling of multiple unmanned vehicles. *Hum. factors* 57, 1208–1218. doi:10.1177/0018720815587803

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics* 3, 22–38. doi:10.2307/3001535

Conati, C., Barral, O., Putnam, V., and Rieger, L. (2021). Toward personalized Xai: a case study in intelligent tutoring systems. *Artificial Intell.* 298, 103503. doi:10.1016/j.artint.2021.103503

Coppens, Y., Efthymiadis, K., Lenaerts, T., Nowé, A., Miller, T., Weber, R., et al. (2019). "Distilling deep reinforcement learning policies in soft decision trees," in nternational joint conference on artificial intelligence, Macao, China, August, 2019, 1–6.

Custode, L. L., and Iacca, G. (2023). Evolutionary learning of interpretable decision trees. *IEEE Access* 11, 6169–6184. doi:10.1109/access.2023.3236260

Das, D., Banerjee, S., and Chernova, S. (2021). "Explainable ai for robot failures: generating explanations that improve user assistance in fault recovery," in ACM/IEEE interanational conference on human-robot interaction, Boulder, CO, United States, March 2021, 351–360. doi:10.1145/3434073.3444657

Das, D., Chernova, S., and Kim, B. (2023a). State2explanation: concept-based explanations to benefit agent learning and user understanding. *Adv. Neural Inf. Process. Syst.* 36, 67156–67182. doi:10.5555/3666122.3669057

Das, D., Kim, B., and Chernova, S. (2023b). "Subgoal-based explanations for unreliable intelligent decision support systems," in Proceedings of the 28th international conference on intelligent user interfaces, Syndney Australia, March 2023, 240–250.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340. doi:10.2307/249008

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., et al. (2020). "ERASER: a benchmark to evaluate rationalized NLP models," in Proceedings of the 58th annual meeting of the association for computational linguistics (Online: Association for Computational Linguistics), 4443–4458.

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., and Li, Y. (2021). "Who needs to know what, when? Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle," in Proceedings of the 2021 ACM designing interactive systems conference, Virtual Event, United States, June 2021, 1591–1602.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv Preprint arXiv:1702.08608. doi:10.48550/arXiv.1702.08608

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., et al. (2021). The who in explainable ai: how ai background shapes perceptions of ai explanations. arXiv Preprint arXiv:2107.13509. doi:10.48550/arXiv.2109.12480

Ehsan, U., and Riedl, M. (2019). On design and evaluation of human-centered explainable ai systems. Glasgow'19.

Ehsan, U., and Riedl, M. O. (2020). Human-centered explainable ai: towards a reflective sociotechnical approach. arXiv Preprint arXiv:2002.01092.

Ehsan, U., and Riedl, M. O. (2021). *Explainability pitfalls: beyond dark patterns in explainable ai*. arXiv Preprint arXiv:2109.12480.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," in IUI '19.

Gentner, D., and Stevens, A. L. (2014). *Mental models*. 1st Edn. Psychology Press. doi:10.4324/9781315802725

Ghaeini, R., Fern, X. Z., and Tadepalli, P. (2018). Interpreting recurrent and attention-based neural models: a case study on natural language inference. arXiv Preprint arXiv:1808.03894.

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health* 3, e745–e750. doi:10.1016/s2589-7500(21)00208-9

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev. Educ. Res.* 42, 237–288. doi:10.3102/00346543042003237

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *J. Pers. Soc. Psychol.* 59, 1216–1229. doi:10.1037//0022-3514.59.6.1216

Gombolay, G. Y., Silva, A., Schrum, M., Gopalan, N., Hallman-Cooper, J., Dutt, M., et al. (2024). Effects of explainable artificial intelligence in neurology decision support. *Ann. Clin. Transl. Neurol.* 11 (5), 1224–1235. doi:10.1002/acn3.52036

Grath, R. M., Costabello, L., Van, C. L., Sweeney, P., Kamiab, F., Shen, Z., et al. (2018). Interpretable credit application predictions with counterfactual explanations. arXiv Preprint arXiv:1811.05245.

Graziano, W. G., Habashi, M. M., Evangelou, D., and Ngambeki, I. (2012). Orientations and motivations: are you a "people person," a "thing person," or both? *Motivation Emotion* 36, 465–477. doi:10.1007/s11031-011-9273-2

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., and Thomaz, A. L. (2013). Policy shaping: integrating human feedback with reinforcement learning. *Adv. Neural Inf. Process. Syst.* 26.

Hagiwara, M. A., Backlund, P., Söderholm, H. M., Lundberg, L., Lebram, M., and Engström, H. (2016). Measuring participants' immersion in healthcare simulation: the development of an instrument. *Adv. Simul.* 1, 17–19. doi:10.1186/s41077-016-0018-x

Hammond, H., Armstrong, M., Thomas, G. A., and Gilchrist, I. D. (2023). Audience immersion: validating attentional and physiological measures against self-report. *Cogn. Res. Princ. Implic.* 8, 22. doi:10.1186/s41235-023-00475-0

Hayes, B., and Shah, J. A. (2017). "Improving robot controller transparency through autonomous policy explanation," in 2017 12th ACM/IEEE international conference on human-robot interaction (HRI), Vienna, Austria, March 2017 (IEEE), 303–312.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable ai: challenges and prospects. arXiv Preprint arXiv:1812.04608. doi:10.48550/arXiv.1812.04608

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023). Measures for explainable ai: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Front. Comput. Sci.* 5, 1096257. doi:10.3389/fcomp.2023.1096257

Huang, S. H., Bhatia, K., Abbeel, P., and Dragan, A. D. (2018). "Establishing appropriate trust via critical states," in 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), Madrid, Spain, October 2018 (IEEE), 3929–3936.

Huang, S. H., Held, D., Abbeel, P., and Dragan, A. D. (2019). Enabling robots to communicate their objectives. *Aut. Robots* 43, 309–326. doi:10.1007/s10514-018-9771-0

Humbird, K. D., Peterson, J. L., and McClarren, R. G. (2018). Deep neural network initialization with decision trees. *IEEE Trans. on neural Netw. learning Syst.* 30, 1286–1295. doi:10.1109/tnnls.2018.2869694

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). "Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in ai," in Virtual event, canada, proceedings of the 2021 ACM conference on fairness, accountability, and transparency, Virtual Event, Canada, March 2021, 624–635.

Kenny, E. M., Ford, C., Quinn, M., and Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in xai user studies. *Artif. Intell.* 294, 103459. doi:10.1016/j.artint.2021.103459

Khan, O., Poupart, P., and Black, J. (2009). "Minimal sufficient explanations for factored markov decision processes," in Proceedings of the international conference on automated planning and scheduling, Thassaloniki, Greece, October 2009, 19, 194–200. doi:10.1609/icaps.v19i1.13365

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., et al. (2019). "The (un) reliability of saliency methods," in *Explainable AI: interpreting, explaining and visualizing deep learning* (Springer), 267–280.

Klein, G., and Hoffman, R. R. (2008). Macrocognition, mental models, and cognitive task analysis methodology. *Naturalistic Decis. Mak. macrocognition*, 57–80.

Koh, P. W., and Liang, P. (2017). "Understanding black-box predictions via influence functions," in *Proceedings of the 34th international conference on machine learning. (PMLR), vol. 70 of Proceedings of machine learning research*. Editors D. Precup, and Y. W. Teh, 1885–1894.

Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. (2015). "Principles of explanatory debugging to personalize interactive machine learning," in Proceedings of the 20th international conference on intelligent user interfaces, 126–137.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). "Too much, too little, or just right? ways explanations impact end users' mental models," in IEEE symposium on visual languages and human centric computing, San Jose, CA, United States, September 2013, 3–10.

Kwon, M., Huang, S. H., and Dragan, A. D. (2018). "Expressing robot incapability," in Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction, Chicago, IL, United States, February 2018, 87–95.

Lage, I., Lifschitz, D., Doshi-Velez, F., and Amir, O. (2019). "Exploring computational user models for agent policy summarization," in Proceedings of the 28th international joint conference on artificial intelligence, Macao, China, August 2019, 28, 1401–1407. doi:10.5555/3367032.3367231

Lai, V., Zhang, Y., Chen, C., Liao, Q. V., and Tan, C. (2023). "Selective explanations: leveraging human input to align explainable ai," *Proceedings of the ACM on Human-Computer Interaction* 7, 1–35. doi:10.1145/3610206

Lakhotia, K., Paranjape, B., Ghoshal, A., Yih, S., Mehdad, Y., and Iyer, S. (2021) "FiD-ex: improving sequence-to-sequence models for extractive rationale generation," in Proceedings of the 2021 conference on empirical methods in natural language processing, Online and Dominican Republic, November 2021. (Online and Punta Cana, Dominican Republic: Association for Computational Linguistics), 3712–3727.

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: building a better stroke prediction model. *Ann. Appl. Statistics* 9, 1350–1371. doi:10.1214/15-aoas848

Li, L., Zhang, Y., and Chen, L. (2023). Personalized prompt learning for explainable recommendation. *ACM Trans. Inf. Syst.* 41, 1–26. doi:10.1145/3580488

Liao, Q. V., Gruen, D., and Miller, S. (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, Honolulu, HI, United States, April 2020, 1–15.

Lievens, F., Harrison, S. H., Mussel, P., and Litman, J. A. (2022). Killing the cat? A review of curiosity at work. *Acad. Manag. Ann.* 16, 179–216. doi:10.5465/annals.2020.0203

Litman, J. (2019). *Curiosity: nature, dimensionality, and determinants*. Cambridge University Press.

Madumal, P., Miller, T., Sonenberg, L., and Vetere, F. (2020). "Explainable reinforcement learning through a causal lens," in Thirty-fourth AAAI conference on artificial intelligence, New York, New York, United States, April 2020, 34, 2493–2500. doi:10.1609/aaai.v34i03.5631

Matthews, G., Lin, J., Panganiban, A. R., and Long, M. D. (2019). Individual differences in trust in autonomous robots: implications for transparency. *IEEE Trans. Hum. Mach. Syst.* 50, 234–244. doi:10.1109/thms.2019.2947592

Mayer, R. E., and Massa, L. J. (2003). Three facets of visual and verbal learners: cognitive ability, cognitive style, and learning preference. *J. Educ. Psychol.* 95, 833–846. doi:10.1037/0022-0663.95.4.833

Millecamp, M., Htun, N. N., Conati, C., and Verbert, K. (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, Genoa, Italy, July 2020, 173–182.

Millecamp, M., Naveed, S., Verbert, K., and Ziegler, J. (2019). "To explain or not to explain: the effects of personal characteristics when explaining feature-based

recommendations in different domains," in Proceedings of the 24th international conference on intelligent user interfaces, Marina del Ray, California, March 2019, 2450, 10–18. doi:10.1145/3301275.3302313

Miller, T. (2021). Contrastive explanation: a structural-model approach. *Knowledge Eng. Rev.* 36, e14. doi:10.1017/S0269888921000102

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). "Explainable prediction of medical codes from clinical text," in Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers), June 2018 (New Orleans, Louisiana: Association for Computational Linguistics), 1101–1111. doi:10.18653/v1/N18-1100

Paleja, R., Chen, L., Niu, Y., Silva, A., Li, Z., Zhang, S., et al. (2023). Interpretable reinforcement learning for robotics and continuous control. arXiv Preprint arXiv:2311.10041. doi:10.15607/RSS.2022.XVIII.068

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., and Gombolay, M. (2021). The utility of explainable ai in *ad hoc* human-machine teaming. *Adv. Neural Inf. Process. Syst.* 34, 610–623.

Paleja, R., Niu, Y., Silva, A., Ritchie, C., Choi, S., and Gombolay, M. (2022). Learning interpretable, high-performing policies for continuous control problems. arXiv Preprint arXiv:2202.02352. doi:10.15607/RSS.2022.XVIII.068

Panagoulias, D. P., Virvou, M., and Tsihrintzis, G. A. (2024). A novel framework for artificial intelligence explainability via the technology acceptance model and rapid estimate of adult literacy in medicine using machine learning. *Expert Syst. Appl.* 248, 123375. doi:10.1016/j.eswa.2024.123375

Patel, M., Prakash, A., and Chernova, S. (2023). Predicting routine object usage for proactive robot assistance. arXiv Preprint arXiv:2309.06252.

Pawar, U., O'Shea, D., Rea, S., and O'Reilly, R. (2020). "Explainable ai in healthcare," in 2020 international conference on cyber situational awareness, data analytics and assessment (CyberSA), June 2020 (IEEE), 1–2.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021) "Manipulating and measuring model interpretability," in Proceedings of the 2021 CHI conference on human factors in computing systems, Virtual, May 2021. New York, NY: Association for Computing Machinery.

Ravichandar, H., Polydoros, A. S., Chernova, S., and Billard, A. (2020). Recent advances in robot learning from demonstration. *Annu. Rev. Control Robotics Auton. Syst.* 3, 297–330. doi:10.1146/annurev-control-100819-063206

Robinette, P., Howard, A. M., and Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situations. *IEEE Trans. Hum. Mach. Syst.* 47, 425–436. doi:10.1109/thms.2017.2648849

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi:10.1038/s42256-019-0048-x

Schrum, M. L., Sumner, E., Gombolay, M. C., and Best, A. (2024). Maveric: a data-driven approach to personalized autonomous driving. *IEEE Trans. Robotics* 40, 1952–1965. doi:10.1109/tro.2024.3359543

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization," in 2017 IEEE international conference on computer vision (ICCV), Venice, Italy, October 2017, 618–626.

Sequeira, P., and Gervasio, M. (2020). Interestingness elements for explainable reinforcement learning: understanding agents' capabilities and limitations. *Artif. Intell.* 288, 103367. doi:10.1016/j.artint.2020.103367

Serrano, S., and Smith, N. A. (2019). Is attention interpretable?. arXiv Preprint arXiv:1906.03731.

Shulner-Tal, A., Kuflik, T., and Kliger, D. (2022a). Enhancing fairness perception–towards human-centred ai and personalized explanations understanding the factors influencing laypeople's fairness perceptions of algorithmic decisions. *Int. J. Hum. Comput. Interact.* 39, 1455–1482. doi:10.1080/10447318.2022.2095705

Shulner-Tal, A., Kuflik, T., and Kliger, D. (2022b). Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics Inf. Technol.* 24, 2. doi:10.1007/s10676-022-09623-4

Silva, A., Chopra, R., and Gombolay, M. (2022a). "Cross-loss influence functions to explain deep network representations," in *Proceedings of the 25th international conference on artificial intelligence and statistics. (PMLR), vol. 151 of Proceedings of machine learning research*. Editors G. Camps-Valls, F. J. R. Ruiz, and I. Valera, 1–17.

Silva, A., and Gombolay, M. (2020). "Neural-encoding human experts' domain knowledge to warm start reinforcement learning" in The 23rd international conference on artificial intelligence and statistics, Online, August 2020.

Silva, A., Gombolay, M., Killian, T., Jimenez, I., and Son, S.-H. (2020). "Optimization methods for interpretable differentiable decision trees applied to reinforcement learning (Online: PMLR)," *Proceedings Machine Learning Research* 108, 1855–1865.

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., and Gombolay, M. (2022b). Explainable artificial intelligence: evaluating the objective and subjective impacts of xai on human-agent interaction. *Int. J. Hum. Comput. Interact.* 39, 1390–1404. doi:10.1080/10447318.2022.2101698

Silva, A., Tambwekar, P., Schrum, M., and Gombolay, M. (2024). "Towards balancing preference and performance through adaptive personalized explainability," in Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction, Boulder, CO, United States, March 2024, 658–668.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv Preprint arXiv:1312.6034. doi:10.48550/arXiv.1312.6034

Singh, R., Miller, T., Lyons, H., Sonenberg, L., Velloso, E., Vetere, F., et al. (2023). Directive explanations for actionable explainability in machine learning applications. *ACM Trans. Interact. Intell. Syst.* 13, 1–26. doi:10.1145/3579363

Sokol, K., and Flach, P. (2020). One explanation does not fit all: the promise of interactive explanations for machine learning transparency. *KI-Künstliche Intell.* 34, 235–250. doi:10.1007/s13218-020-00637-y

Sreedharan, S., Olmo, A., Mishra, A. P., and Kambhampati, S. (2019). Model-free model reconciliation. arXiv Preprint arXiv:1903.07198.

Stilgoe, J. (2019). Self-driving cars will take a while to get right. *Nat. Mach. Intell.* 1, 202–203. doi:10.1038/s42256-019-0046-z

Suárez, A., and Lutsko, J. F. (1999). Globally optimal fuzzy decision trees for classification and regression. *IEEE Trans. on Pattern Analysis Machine Intelligence* 21, 1297–1311. doi:10.1109/34.817409

Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., et al. (2021). "Habitat 2.0: training home assistants to rearrange their habitat," in *Advances in neural information processing systems*. Editors M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Glasgow, United Kingdom: Curran Associates, Inc), 34, 251–266.

Tambwekar, P., and Gombolay, M. (2023). Towards reconciling usability and usefulness of explainable ai methodologies. arXiv Preprint arXiv:2301.05347.

Tambwekar, P., Silva, A., Gopalan, N., and Gombolay, M. (2023). Natural language specification of reinforcement learning policies through differentiable decision trees. *IEEE Robot. Autom. Lett.* 8, 3621–3628. doi:10.1109/LRA.2023.3268593

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019). "What clinicians want: contextualizing explainable machine learning for clinical end use," in Machine learning for healthcare conference, Ann Arbor, MI, United States, August 2019 (PMLR), 359–380.

Topin, N., and Veloso, M. (2019). "Generation of policy-level explanations for reinforcement learning," in Proceedings of the thirty-third AAAI conference on artificial intelligence and thirty-first innovative applications of artificial intelligence conference and ninth aaai symposium on educational advances in artificial intelligence, Honolulu, Hawaii, United States, January 2019, 33, 2514–2521. doi:10.1609/aaai.v33i01.33012514

Vered, M., Livni, T., Howe, P. D. L., Miller, T., and Sonenberg, L. (2023). The effects of explanations on automation bias. *Artif. Intell.* 322, 103952. doi:10.1016/j.artint.2023.103952

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., and Doshi-Velez, F. (2018). "Beyond sparsity: tree regularization of deep models for interpretability." Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence, New Orleans, Louisiana, United States, February 2018.

Wu, M., Parbhoo, S., Hughes, M. C., Roth, V., and Doshi-Velez, F. (2021). Optimizing for interpretability in deep neural networks with tree regularization. *J. Artif. Intell. Res.* 72, 1–37. doi:10.1613/jair.1.12558

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). "Explainable ai: a brief survey on history, research areas, approaches and challenges," inCCF international conference on natural language processing and Chinese computing, Dunhuang, China, October 2019 (Springer), 563–574.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv Preprint arXiv:1506.06579. doi:10.48550/arXiv.1506.06579

Zablocki, É., Ben-Younes, H., Pérez, P., and Cord, M. (2021). Explainability of vision-based autonomous driving systems: review and challenges. arXiv Preprint arXiv:2101.05307. doi:10.1007/s11263-022-01657-x

Zahedi, Z., Sengupta, S., and Kambhampati, S. (2024). "'Why didn't you allocate this task to them?'negotiation-aware task allocation and contrastive explanation generation," Thirty-eighth AAAI conference on artificial intelligence, Vancouver, Canada, March 2024, 38, 10243–10251. doi:10.1609/aaai.v38i9.28890

Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in Proceedings of the 2020 conference on fairness, accountability, and transparency, Barcelona, Spain, January 2020, 295–305.

Zhou, Y., Ribeiro, M. T., and Shah, J. (2022). Exsum: from local explanations to model understanding. arXiv Preprint arXiv:2205.00130. doi:10.18653/v1/2022.naacl-main.392