Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction

Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay

Georgia Institute of Technology

ARTICLE HISTORY
Compiled August 15, 2022

ABSTRACT
Intelligent agents must be able to communicate intentions and explain their decision-making processes to build trust, foster confidence, and improve human-agent team dynamics. Recognizing this need, academia and industry are rapidly proposing new ideas, methods, and frameworks to aid in the design of more explainable AI. Yet, there remains no standardized metric or experimental protocol for benchmarking new methods, leaving researchers to rely on their own intuition or ad hoc methods for assessing new concepts. In this work, we present the first comprehensive (n=286) user study testing a wide range of approaches for explainable machine learning, including feature importance, probability scores, decision trees, counterfactual reasoning, natural language explanations, and case-based reasoning, as well as a baseline condition with no explanations. We provide the first large-scale empirical evidence of the effects of explainability on human-agent teaming. Our results will help to guide the future of explainability research by highlighting the benefits of counterfactual explanations and the shortcomings of confidence scores for explainability. We also propose a novel questionnaire to measure explainability with human participants, inspired by relevant prior work and correlated with human-agent teaming metrics.

KEYWORDS
Explainable Artificial Intelligence, Trust, Human-Agent Interaction, Metrics

1. Introduction

Computational agents (e.g., robots or virtual agents) must be able to communicate intentions and explain their decision-making processes to build trust, foster confidence, and improve team dynamics (Boies, Fiset, & Gill, 2015; Paleja, Ghuy, Arachchige, & Gombolay, 2021), and research is increasingly investigating how *explainability* is necessary for many human-agent interactions and domains (Doshi-Velez & Kim, 2017; Rudin et al., 2021). For agents to effectively interact with humans in human-agent teams, agents must be capable of communicating their intentions and explaining their decision-making process. Researchers have investigated explainability methods for agents to empower users to better understand the reasoning behind the agent's behavior (e.g., through natural language generation (DeYoung et al., 2019), decision-tree extraction (Silva, Gombolay, Killian, Jimenez, & Son, 2020), and counterfactuals (Karimi, Schölkopf, & Valera, 2021)). Yet, while researchers have recognized the

CONTACT Andrew Silva. Email: andrew.silva@gatech.edu

need for agents to explain their decisions, we hypothesize that not all explainability methods are equally effective at communicating information, and some methods may even inhibit human understanding and collaboration. Current progress in the field of explainable artificial intelligence (xAI) is hindered by a lack of standardized measurement by which to evaluate explainability methods and a lack of clear comparison across various xAI techniques.

As noted by recent surveys on xAI (Adadi & Berrada, 2018; Karimi, Barthe, Schölkopf, & Valera, 2020; Rudin et al., 2021), explainability research lacks consistent definitions and evaluations making it difficult to draw sound conclusions about the efficacy of explainability techniques (Rudin, 2019). Additionally, such inconsistencies often lead to conflicting takeaways. For example, Jain and Wallace (2019) published "Attention is not explanation" just five months before Wiegreffe and Pinter (2019) published "Attention is not not explanation." Such contradictions are often contingent on differences in definitions or expectations and leave researchers ill-informed on whether to pursue attention for explanations. The enthusiastic pace of progress in xAI is outpacing the ability of the community to settle these debates with rigorous empirical or analytical study.

What is critically needed to pursue and adopt the most beneficial xAI methods for human-agent teaming are standardized metrics and experimental protocols. In pursuit of such a goal, prior research has put forth automated xAI metrics, such as model stability or complexity (Rosenfeld, 2021) or natural-language benchmarks (DeYoung et al., 2019), but very few prior works in xAI involve user studies with human participants in their evaluations (Jain & Wallace, 2019; Karimi et al., 2020). When humans are involved, the work typically takes a narrow look at a single method or use-case, and therefore has limited implications for the field at large. While standardized agent evaluation task sets (Bedny & Karwowski, 2003) and surveys exist in the literature (Bartneck, Kulić, Croft, & Zoghbi, 2009; Hartson, Andre, & Williges, 2001; Jian, Bisantz, & Drury, 2000; Nomura, Suzuki, Kanda, & Kato, 2006), there is not an empirically-validated or agreed-upon survey to evaluate explainability of virtual or embodied agents deployed to untrained human users ("lay" users, as defined by Ribera and Lapedriza (2019)). To make progress on developing useful xAI that operates effectively alongside human users, machine-learning researchers must have access to shared, validated surveys and experimental procedures to benchmark different xAI techniques.

In this work, we present the first evaluation of a battery of approaches to xAI with human users in a large-scale user study (n=286). A visual overview of our study is presented in Figure 1. For the first time, our work enables objective and subjective evaluation of different xAI methods with real human users across axes of performance, efficiency, trust, social-perception, and compliance. Rather than relying on speculation for how humans might respond to different xAI approaches, we present a true comparison in a between-subjects user study.

Based upon our results, we conduct a post-hoc factor analysis on a composite xAI survey and find three potential dimensions of explainability we interpret as measuring transparency ($\alpha_1 = 0.83$), usability ($\alpha_2 = 0.82$), and simulatability ($\alpha_3 = 0.81$). We show that a composite scale comprised of these dimensions is correlated with measures of trust, perceptions of social competence, and performance. This new xAI survey offers the potential of a quantitative scoring mechanism for xAI agnostic to the particular technique being used, allowing for consistent evaluation of xAI across studies, techniques, and demographics. We conclude with proposed future work that will investigate the reliability and validity of this survey across multiple studies.

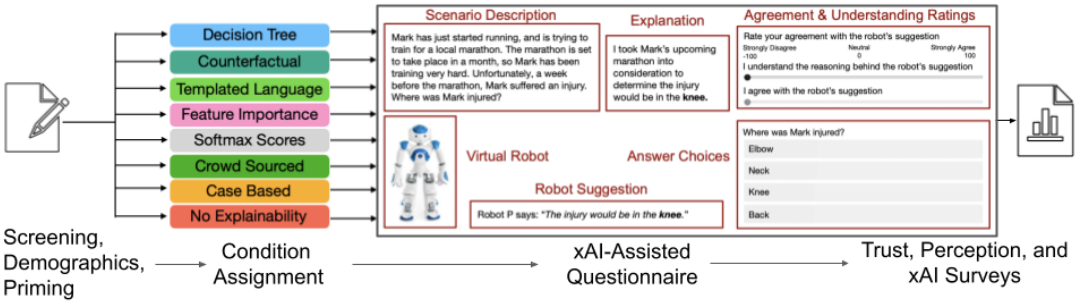


Figure 1.: A visual walkthrough of our study. Participants first complete consent forms, a screening task, and demographic surveys before beginning a priming task. The priming task prepares participants to consider usability and transparency of agent suggestions and explanations. Participants are then assigned a condition and receive instructions for their assigned explanations before beginning the main set of scenarios in our study. Each scenario includes a short paragraph of text, a question, an explanation from a virtual agent, two Likert items assessing per-scenario understanding and agreement, and four answer choices. After each question, participants are shown the correct answer and a running tally of their overall score. After completing all scenarios, participants complete a trust survey (Jian et al., 2000), social perceptions survey (Bartneck et al., 2009), and our xAI survey.

1.1. Contributions

The primary contribution in our work is the first large-scale evaluation of the objective and subjective effects of various forms of explainability on human-robot teaming. Our results show that explainability strongly correlates with trust ($p < 0.0001$), social competence ($p < 0.0001$), and performance ($p = 0.01$), and that counterfactual, language-based feature descriptions, and case-based explanations are rated as more explainable than probability scores ($p < 0.01$). We also contribute survey materials and study design insights for future work to build upon our user study, including a proposed explainability measurement survey to be verified in future work. Our results help to inform the design of explainability approaches in the future by revealing both the positive effects and potential risks of adopting different forms of xAI.

2. Related Work

In this section, we provide an overview of the literature that relates most closely to our study and point to surveys for interested readers to review the latest advances in xAI.

2.1. Explainability vs. Interpretability

With the rapid growth of work in xAI, debates over terminology persist. In our work, we are explicitly concerned with *explainable* machine learning – that is, we present explanations for model outputs from one of a set of popular approaches. Crucially, within xAI, explanations do not necessarily reflect the ground-truth decision-making of the model. Explanations may simply offer insight into how the decision was reached

(e.g., highlighting important features, presenting answer probabilities, etc.). Presenting model explanations lies in contrast to *interpretable* machine learning, where the model itself is easily read by a human (e.g., a small decision tree (Basak, 2004; Breiman, Friedman, Stone, & Olshen, 1984; Olaru & Wehenkel, 2003), rule list (Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2017; C. Chen & Rudin, 2017; Letham, Rudin, McCormick, Madigan, et al., 2015; Weiss & Indurkhya, 1995), or simple linear model (Caruana et al., 2015)). The distinction between the two is important (Adadi & Berrada, 2018; Lipton, 2018; Rudin, 2019), though still unsettled (Hase & Bansal, 2020). For more thorough surveys on the recent advances in xAI, we refer readers to Holzinger, Carrington, and Müller (2020), Linardatos, Papastefanopoulos, and Kotsiantis (2021), and Hoffman, Mueller, Klein, and Litman (2018). We specifically target *explainable* methods in this work, being a broader class of algorithms and techniques, though we include a decision-tree condition to compare to an *interpretable* technique.

2.2. Evaluating Explainability

The question of how to appropriately evaluate xAI research is crucial and has thus garnered much attention. Automated metrics, such as ROAR for feature importance (Hooker, Erhan, Kindermans, & Kim, 2018), ERASER for natural-language explanations (DeYoung et al., 2019), or model-agnostic measures such as stability and complexity (Rosenfeld, 2021), attempt to approximate human understanding with a benchmark or to score a method on how internally consistent the method is. However, such approximations have never been thoroughly tested or empirically validated with humans.

As opposed to employing automatic metrics, one could evaluate xAI on a strict case-by-case basis by considering the deployment domain, users, model performance (Saragih & Morrison, 2021), robustness, and more (Sokol & Flach, 2020). While such a thorough evaluation may be preferable when possible, it is prohibitively expensive to run such an evaluation on every model for every deployment domain. The Explanation Satisfaction scale (Hoffman et al., 2018) measures the utility of an explanation (either from a human or an xAI technique) as determined by experts in the field of xAI. Crucially, this scale is not designed for an untrained population. What we need is a tool to enable general comparison of how untrained human users perceive and use xAI.

While we are unaware of any prior work that has performed a thorough comparison of a multitude of xAI techniques on a large, untrained population, prior research has empirically evaluated individual xAI techniques and use-cases with human users in narrow cases (Hase & Bansal, 2020; Hutton, Liu, & Martin, 2012; Nguyen, 2018; Poursabzi-Sangdeh, Goldstein, Hofman, Wortman Vaughan, & Wallach, 2021; Tintarev & Masthoff, 2012). Researchers have primarily examined whether or not human users rate xAI as helpful, and such research has produced mixed results (Hutton et al., 2012; Nguyen, 2018). Research with a limited sample population of computer science students found that saliency measures (i.e., feature importance) helped improve a user's understanding of decisions (Hase & Bansal, 2020). Earlier research with crowd-sourced users found similar results depending on the difficulty of the task, with harder tasks leading to the perception of less-useful explanations (Hutton et al., 2012; Nguyen, 2018). Similarly, prior research has found that users tend to like explanations from recommender systems given in the form of natural language (Tintarev & Masthoff,

2012). Most surprisingly, prior research on explainability and compliance has found that users were *more* likely to agree with a decision-making tool if the tool provided an explanation – even if the tool was incorrect (Poursabzi-Sangdeh et al., 2021). This result runs counter to the intuition that more explainable methods will reduce human over-trust. Generalizing the result of prior work Poursabzi-Sangdeh et al. (2021), our large-scale study reveals that xAI makes no change in human compliance with a virtual agent, while examining a broader set of xAI techniques. Our research drives at the perceived utility of explanations, human compliance, performance, trust, and social perceptions.

2.3. Human-Centric Explainability

As explanations often involve interaction with a human user, there is also prior work on how to frame explanation research around the human in the loop (Ehsan & Riedl, 2020). Research on human preferences has found that humans typically prefer simpler explanations, only allowing for explanations to grow complex when all of the components of the explanation are highly probable (Lombrozo, 2006, 2007). Miller (2019) provides a set of key insights and common themes relating to human explanations and properties of explanations. Explanations in human-human contexts often establish a common ground or knowledge-base from which to make decisions or justify behaviors. Miller (2019) highlights various types of explanations that may be applied to algorithmic explanation (e.g., Aristotle's *Four Causes* model (Lloyd & Lloyd, 1996)) and how these mechanisms might be leveraged in different scenarios. Wang, Yang, Abdul, and Lim (2019), Liao, Gruen, and Miller (2020), and, Schoonderwoerd, Jorritsma, Neerincx, and van den Bosch (2021) approach explainability with an eye towards design, developing frameworks, question-banks, or undergoing full case-studies to assist in the development of algorithms for explainability. Similarly, Lage et al. (2018) provides explanation design insights following a large-scale user study on the effects of explanation length, complexity, and repetition on subjective preferences and human-user accuracy, finding that shorter and simpler explanations were preferred. Related work has also examined how concepts such as fairness, accountability, and transparency relate to explainability (Shin, 2021), finding that causability plays a role in human trust.

In our work, we instead approach explainability through the lens of subjective usability and preference. Specifically, we ask the question: Given various forms that an explanation may take (e.g., language expressions (DeYoung et al., 2019), decision trees (Weiss & Indurkhya, 1995), feature importance maps (Ribeiro, Singh, & Guestrin, 2016), etc.), which form is considered the easiest to use, interpret, and trust?

2.4. Explainability in our study

Our work compares seven broad categories of xAI methods including case-based reasoning, decision trees, feature importance, probability scores, counterfactuals, natural-language explanations, and crowd-sourced explanations. These seven were chosen as overarching categories of xAI to broadly capture the scope of current xAI research. Our seven conditions enable us to compare different modalities that may be used for presenting an explanation (e.g., highlighting input features vs. presenting a decision-tree), as well as comparing different forms of presenting the same information (e.g., presenting percent-likelihoods for each answer vs. presenting a natural-language

sentence that includes an answer probability). Below, we present more detail on each of the conditions in our study, and a visual example of each condition is given in Figure 2.

A case-based explanation (Barnett et al., 2021; Caruana, Kangarloo, Dionisio, Sinha, & Johnson, 1999; Koh & Liang, 2017) shows training data that closely resemble testing samples. Case-based explanations help end-users to understand output decisions by relating the current input to a known data point and drawing a connection to the previous, known label for such a data point. Seeing labeled training examples that look like a given testing sample (Klein, 1993), the user may achieve greater understanding of why the model produced a certain classification.

We also consider decision trees for explainability (Agarwal & Das, 2020; Bastani, Pu, & Solar-Lezama, 2018; Craven & Shavlik, 1995; Murthy, 1998; Silva et al., 2020; Wu et al., 2018). A decision tree is a graphical flow-chart showing a cascade of "True/False" checks that lead to a decision. Each decision node includes a check against the input data, which end-users may use to manually verify the output of the system. By showing an *interpretable* flow-chart to an end user, the user is empowered to assess and understand the decision.

We next examine attention/saliency mechanisms for xAI (Jain & Wallace, 2019; Ribeiro et al., 2016; Suau, Zappella, & Apostoloff, 2020; Wiegreffe & Pinter, 2019) by using a feature-importance based explanation (Caruana et al., 2015; Štrumbelj & Kononenko, 2014). These approaches show users the features of an input sample that were the most important for a classification. The user can gain a better understanding of a decision by ensuring that the features are reasonable or consistent with their own expectations. Crucially, such models only reveal correlations between features and predictions; they do not imply or predict causal relationships.

Our work also examines counterfactual explanations (Karimi et al., 2020, 2021; Verma, Dickerson, & Hines, 2020; Wachter, Mittelstadt, & Russell, 2017). A counterfactual explanation works by telling a user how a decision would be different if perturbations were made to an input sample. Based on a counterfactual scenario, a user can infer how the original decision was made, though claims around the true explainability of current black-box counterfactual methods remain contested (White & Garcez, 2021).

We also include explanations via natural language, which is an active area of research dedicated to providing textual descriptions of classifications (H. Chen, Chen, Shi, & Zhang, 2021; DeYoung et al., 2019; Ehsan & Riedl, 2020). Often, this work involves gathering a large corpus of annotated explanations or images, which can then be leveraged to learn a generative language model that produces natural-language sequences to explain given model input samples and output predictions (Mishra & Rzeszotarski, 2021). In our work, the natural language explanations are produced and vetted by researchers and pilot participants to ensure quality and consistency.

Finally, we investigate probability (i.e., confidence) scores presented in the form of "crowd-sourced" explanations (i.e., a natural-language sentence presenting the percentages of experts that voted on an answer) or as a table of answer probabilities for the human to interpret. Such a modality offers explainability by showing the uncertainty of the model for a given input sample. Prior research on using confidence scores as explanations (van der Waa, Schoonderwoerd, van Diggelen, & Neerincx, 2020; Zhang, Liao, & Bellamy, 2020) shows that such explanations improve user trust and confidence. However, such results are contentious, as confidence scores can vary significantly due to small perturbations in samples (Hogan & Kailkhura, 2018; Kailkhura, Gallagher, Kim, Hiszpanski, & Han, 2019), suggesting that user trust may

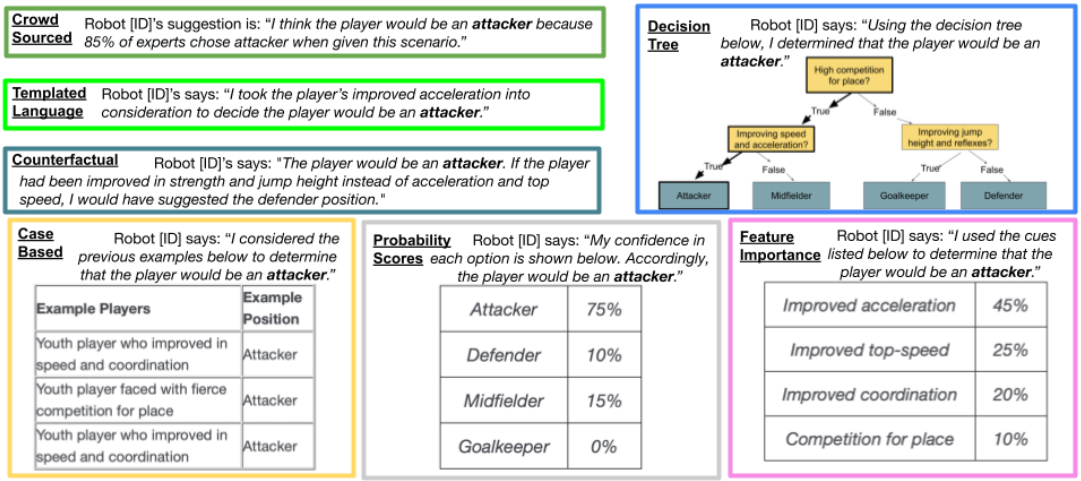


Figure 2.: An example for each of the xAI conditions in our study.

be misplaced.

3. Overview

Our research seeks to answer questions about the effects of using different classes of xAI on trust, performance, perceptions of social competence or intelligence, compliance, and efficiency when such methods are deployed as decision-aids for untrained humans. While there is a vast landscape of xAI research and dozens of methods that could all be compared and contrasted, we are interested in human perceptions and performance with different classes of xAI. To ensure relevant, high-quality explanations, we wizard-of-oz (WoZ) (Kelley, 1984) all explanations in the study and conduct multiple iterative pilot studies with our explanations. Specifically, we investigate the following research questions

(1) **RQ1** – What is the relationship between agent explainability and human-rated subjective metrics (i.e., user trust or social impressions)?
(2) **RQ2** – What is the relationship between agent explainability and objective task performance (i.e., accuracy and efficiency)?
(3) **RQ3** – Are there significant differences in perceived explainability across the classes of xAI in our study?

We note that our research questions are specifically around the relationships between explainability and objective/subjective metrics, and we do not explicitly investigate the causal relationship between explainability and the metrics in our study.

To answer these questions, we conduct a between-subjects user study in which participants must answer a set of multiple-choice questions with the aid of a virtual agent assistant. Our task involves answering a set of multiple-choice questions, where each question relates to a short paragraph. The user receives an answer suggestion from a virtual agent and an explanation drawn from one of the following conditions:

- **Templated Language** – Natural language citing the most relevant feature in the question.

- **Counterfactual** – Natural language describing the second-likeliest answer and how the scenario should change to produce the second-likeliest answer.
- **Decision Tree** – A graphic flow-chart with three "True/False" checks leading to a classification.
- **Probability Scores** – Probability scores for each answer choice.
- **Crowd-Sourced** – The percentages of experts that selected each answer choice.
- **Case-Based** – Three short examples of prior scenarios and their associated answers.
- **Feature Importance** – Relevance scores for each feature in the scenario.
- **No Explanations** – No explainability added.

4. Materials and Methods

We conducted a 1×8 between-subjects user study to answer our research questions.

4.1. Pilot Studies

Before beginning our full study, we conducted several iterative pilot studies to refine our study design. Our pilot studies involved a total of 54 participants, running different versions of the study over time. Through our pilot studies, we learned to increase total compensation for our study, add a screening quiz, modify explanations, and improve the explanation introduction portion of our study, after observing that our study took longer than expected and garnered several low-effort responses (Buchanan & Scofield, 2018). Throughout, we iterated on instructions to improve completion rates.

4.2. Pre-Study

Participants first completed a consent form and then received a brief set of instructions for the task. These instructions included an example scenario and an introduction to the virtual agent and the rating scales that would be used to judge agent advice. After receiving instructions, participants were given a five-question quiz on the instructions they received, and any participant that did not answer all five questions correctly was removed from the study. This quiz served to screen participants who might have confounded our results (Buchanan & Scofield, 2018). After passing the instructions quiz, we first collected demographic information from participants and asked them to complete the negative attitudes toward agents (NARS) questionnaire (Nomura et al., 2006) to measure whether such data might confound our results.

4.3. Scenarios

Our study consisted of showing participants a set of scenarios and then asking multiple-choice questions. We show an example of one such scenario, with a **Templated Language** explanation, in Figure 1. The scenario includes a short description that provides background information about an imaginary person, ending in a question. The participant is prompted to respond to two Likert items and to answer the question, and participants assigned to an xAI condition see an explanation placed between the agent suggestion and the two Likert items.

The questions in our study generally require the participant to infer something about that person's preferences, future decisions, or past actions. While some scenarios are quite simple, others are more challenging or require specific areas of prior knowledge. This range in difficulty promoted varied reliance on the agent throughout the study. The scenarios were manually generated as commonsense reasoning questions and refined through pilot studies and testing. All scenarios are included in the appendix.

4.4. Agent-based Decision-support

Central to our study was the assistance that participants received from a virtual agent and the xAI method used. To offset any adjustment in the study, our instructions indicated that the virtual agent was changed out for every answer, and the graphic for the agent was cycled between different photos of a NAO robot. The NAO robot is a 25-DoF humanoid robot from Softbank Robotics, shown in Figure 1. Additionally, the agents were all referred to using a different ID to indicate that the agents were not consistent across scenarios. For several pre-specified questions, the agent suggested the wrong answer rather than the correct answer to measure inappropriate compliance with agent advice.

Throughout the study, the virtual agent offered explanations for its suggestions to the participant. These explanations were all created via WoZ and validated in our pilot studies. When the agent suggested the wrong answer, the explanation supported the wrong answer (i.e., the explanation and the wrong answer were internally consistent).

4.5. Priming Task

After completing our pre-study forms, participants advanced to the priming task, involving five scenarios without agent explanations. Of these initial five scenarios, the agent suggested the wrong answer once, showing participants that they could not rely on the agent's suggestion without considering whether the suggestion might be true/false with the aid of the xAI provided. After the first set of scenarios, participants completed our new xAI survey, developed to measure user-rated explainability of an agent's suggestions. By tasking participants with a set of scenarios and the new xAI survey *before* they were assigned a condition, we primed users to consider how useful or transferable the agent's explanations might be for the remainder of the study.

4.6. Condition Assignment

Participants were randomly assigned to one of our conditions. For all conditions other than **Nothing**, participants were given a brief walkthrough of how their condition would work. For example, in the **Decision Tree** condition, participants were introduced to the concept of a flow-chart as a decision-aid and were given an example for how one might be applied and how it could be interpreted. This introduction provided a high-level overview of how to read agent explanations, as many xAI approaches (e.g., **Feature Importance**, **Decision Tree**, **Case Based**, and **Probability Scores**) do not use natural language, and, therefore, could be unintelligible to novice users without some level of introduction. However, we did not provide in-depth explanations of how these methods manifest explanations, as the purpose of our experiment is to evaluate how *untrained* participants would rate different explainability measures. Examples from the introduction to each condition

are given in Figure 2.

4.7. Primary Task

Once the participants completed the priming task with five scenarios, they began the primary task of the study, which was comprised of fourteen scenarios. The agent offered incorrect suggestions on the fifth, seventh, eighth, tenth, and twelfth questions. We fixed the ordering of all questions and answer suggestions to control for any randomization effects on participant ratings at the end of the study. In total, participants answered twenty total questions (i.e., one for instructions, five for priming, and fourteen for the main body of the study) and the agent only offered incorrect advice six times total; thus, the agent was correct more often than it was incorrect (i.e., correct 70% of the time). If the agent had never been incorrect (or never been correct), we would not have been able to study participant compliance or reliance with the agent's suggestions. We skewed the agent to be correct for 70% of the available scenarios, as prior work suggests that a less accurate agent may have been discounted entirely (Wiczorek & Manzey, 2014; Yang, Unhelkar, Li, & Shah, 2017).

4.8. Follow-up

After completing all twenty questions (one introductory, five priming, and fourteen primary), the participants completed a trust in automation survey (Jian et al., 2000) and the Godspeed survey (Bartneck et al., 2009) to provide us with metrics for the effects of xAI on trust and perception of the agent. Finally, participants completed our post-trial xAI survey (after initially completing it for the priming questions) and were then given the opportunity to enter free-response text before completing the study.

4.9. xAI for Human-agent Interaction Survey Development

Our work leverages a novel xAI survey to measure human-rated explainability of agent explanations and suggestions. We created a 30-question survey with items intended to measure simulatability, transparency, and usability of the agent's explanations. Questions in the survey are targeted at these three primary axes after prior work identified simulatability, transparency, and usability as important metrics for explainability (Holzinger et al., 2020; Sokol & Flach, 2020). Questions in our xAI survey were inspired by guidelines introduced in prior work (Hoffman et al., 2018; Sokol & Flach, 2020) and prior surveys on usability (Brooke, 1996) and causality (Holzinger et al., 2020).

As prior work has already established questionnaires to evaluate topics such as usability (Brooke, 1996) and explanation faithfulness (Hoffman et al., 2018; Sokol & Flach, 2020), we aggregated and extended existing questions rather than generating an entirely new set of questions via interviews (Nomura et al., 2006) or a word-elicitation process (Jian et al., 2000). Questions in our xAI include questions from prior work as well as new questions specifically designed to re-test questions in prior work (e.g., by including negations of existing questions). All items in the survey are rated on a seven-point scale from "Strongly Disagree" to "Strongly Agree," and the final explainability score is calculated as the sum of all items in the questionnaire (adding the inverted value for negative items). The full 30-question survey is given below, and citations to relevant prior work are given for each question.

(1) The explanations were detailed enough for me to understand. (Holzinger et al., 2020)
(2) I understood the explanations within the context of the question. (Holzinger et al., 2020; Shin, 2021)
(3) The explanations provided enough information for me to understand. (Holzinger et al., 2020)
(4) I understood how the agent arrives at its answer. (Brooke, 1996; Hoffman et al., 2018)
(5) I was able to use the explanations with my knowledge base. (Holzinger et al., 2020)
(6) I would be able to repeat the steps that the agent took to reach its conclusion.
(7) I think that most people would learn to understand the explanations very quickly. (Brooke, 1996; Hoffman et al., 2018; Holzinger et al., 2020)
(8) I would not understand how to apply the explanations to new questions. (Hoffman et al., 2018)
(9) I would not be able to recreate the process by which the agent generated its answers.
(10) I understand why the agent used specific information in its explanation. (Hoffman et al., 2018; Holzinger et al., 2020)
(11) I understood the agent's reasoning. (Brooke, 1996; Hoffman et al., 2018; Shin, 2021)
(12) I could have applied the agent's reasoning to new problems, even if the agent didn't give me suggestions.
(13) The explanations were actionable, that is, they helped me know how to answer the questions. (Hoffman et al., 2018)
(14) I believe that I could provide an explanation similar to the agent's explanation.
(15) I would need more information to understand the explanations. (Holzinger et al., 2020)
(16) I had trouble using the explanations to answer the question. (Brooke, 1996)
(17) I believe that the explanations would not help most people in answering the question. (Hoffman et al., 2018)
(18) The explanations were an important resource for me to answer the question. (Hoffman et al., 2018)
(19) I do not think most people would provide similar explanations as the agent's explanation.
(20) I think that most people would be able to interpret the explanation of the agent. (Brooke, 1996)
(21) Most people would be able to accurately reproduce the agent's decision-making process.
(22) Most people would not be able to apply the agent's explanations to the questions. (Hoffman et al., 2018)
(23) I could not follow the agent's decision-making process. (Holzinger et al., 2020)
(24) I could easily follow the explanation to arrive at an answer to the question. (Brooke, 1996)
(25) The explanations were useful. (Brooke, 1996)
(26) I am able to follow the agent's decision-making process step-by-step.
(27) The explanations were not relevant for the questions I was given.
(28) I understand how the agent's decision-making process works.
(29) I could apply the explanations to the questions I was given.
(30) I could not figure out how the agent arrived at its suggestions.

In the remainder of this work, we present results using the full 30-question survey as a measure of explainability, comparing our results with surveys in the literature measuring complementary phenomena (i.e., trust-in-automation (Jian et al., 2000) and Godspeed (Bartneck et al., 2009)) and examining the relationship between participant-rated explainability and objective/subjective metrics.

4.10. Procedure

We recall that participants first completed pre-study consent forms and a briefing of the task, showing them one introductory scenario. Participants were then screened to ensure high-quality responses, with failures on the screening task being removed from the study. After finishing the screening task, participants provided demographic data and then began a priming task of five scenarios that prepared participants to consider the usability, transparency, and simulatability of agent suggestions and explanations. Following the priming task, participants were randomly assigned one of eight possible conditions and provided instructions for their assigned condition (e.g., participants in the "Decision Tree" condition were taught how to read and interpret decision trees). Finally, participants began the main body of the study and completed fourteen scenarios of varying difficulty with the assistance of a virtual agent. Upon completion of all scenarios, participants rated the agents on trustworthiness (Jian et al., 2000), intelligence and likeability (Bartneck et al., 2009), and explainability.

4.11. Measures

We seek to quantify the relationship between explainability and trust, task performance, and social perceptions of agents (i.e., is the agent "kind," "amicable," and "socially intelligent?") and to determine which approaches to xAI will provide the greatest objective benefits to human-agent team fluency. Using the following metrics, we can effectively capture both objective task performance and subjective impressions of the virtual agent and explainability condition. To answer our research questions, we employ the following metrics:

- **M1** (RQ2) Completion time – how long it takes participants to complete the primary task of the survey.
- **M2** (RQ2) Accuracy – how many questions the participant answers correctly.
- **M3** (RQ2) Compliance – how frequently the participant agrees with the agent's suggestion.
- **M4** (RQ1) Social Competence – how participants perceive the agent as a social agent according to the Godspeed questionnaire rating the agent on scales relating to kindness, friendliness, intelligence, etc. (Bartneck et al., 2009).
- **M5** (RQ1) Trust – participant's trust in the agent as measured by the trust-in-automation (Jian et al., 2000) survey.
- **M6** (RQ1) Explainability – participant's self-rated understanding and explainability as measured by the full xAI survey introduced above.

4.12. Participants

We recruited a total of 340 participants for our pilot studies and final study from Amazon Mechanical Turk (Paolacci, Chandler, & Ipeirotis, 2010). Our study was

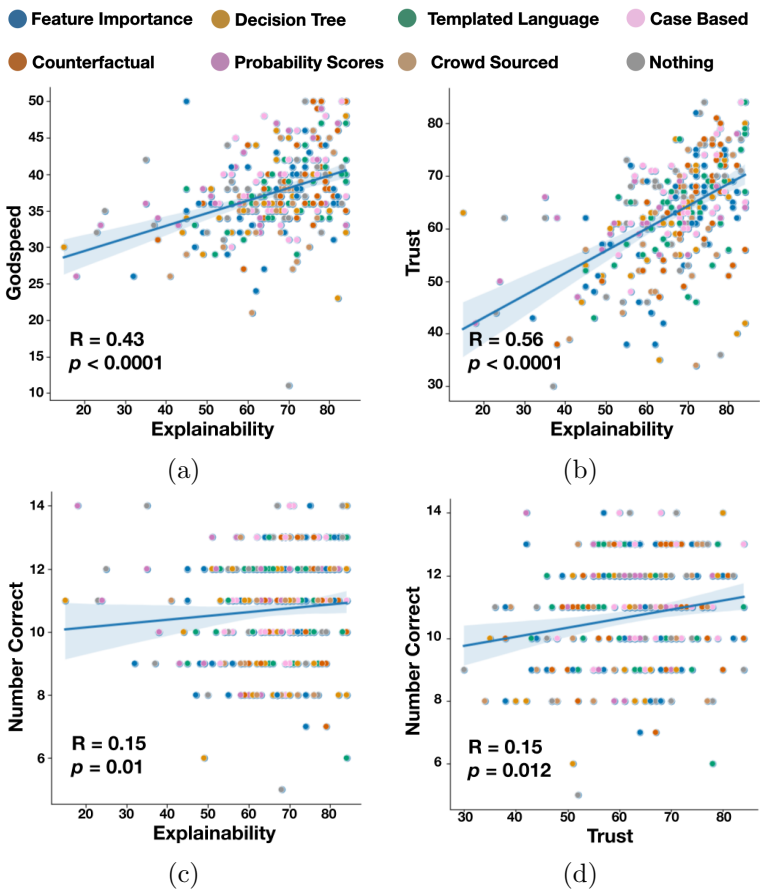


Figure 3.: A depiction of our explainability results correlated to: (a) social perception, (b) trust, and (c) accuracy and of trust correlated to accuracy (d). We find trust, accuracy, and social perceptions were statistically significantly correlated with explainability as measured by our xAI survey, both lending support for use of our survey as a measure of explainability and addressing **RQ1** and **RQ2**. We also observe that trust and accuracy are statistically significantly correlated. Each dot represents a data point with the regression line and confidence intervals drawn for each correlation.

approved by an IRB[1] and participants were compensated $5.00. After our pilot studies, our final study included 286 participants (Mean age: 43.0; SD: 10.7; 52% Female). The study took approximately 25 minutes.

5. Results

In this section, we review and discuss key results from our final study. We tested all data for normality and homoscedasticity, and if parametric assumptions failed we applied a non-parametric test.

[1]Our study was approved by the Georgia Institute of Technology IRB under Protocol H20522.

Condition	xAI	Trust	Social Competence
Templated Language	170 (18.3)	63.5 (8.96)	37.2 (4.12)
Counterfactual	177 (24.2)	64.7 (10.5)	39.1 (5.34)
Decision Tree	164 (34.7)	61.0 (10.4)	36.8 (5.54)
Probability Scores	145 (42.3)	61.3 (7.9)	37.8 (5.37)
Crowd Sourced	166 (25.4)	60.5 (11.8)	35.2 (6.12)
Case Based	172 (18.3)	66.2 (7.5)	39.5 (5.19)
Feature Importance	167 (30.3)	61.8 (11.1)	36.9 (5.34)
Nothing	158 (35.3)	63.1 (12.1)	37.4 (6.37)

Table 1.: In this Table we report the mean and (standard deviation) for explainability according to the our xAI survey scores, trust (Jian et al., 2000), and social competence (Bartneck et al., 2009) for each of the conditions in our study.

5.1. Significant Findings

We summarize our significant findings here and provide average variables from our analyses (Table 1) before providing deeper analysis on each research question further below.

- Participant trust is correlated with agent explainability ($\rho = .56$, $p < 0.0001$) **(RQ1)**.
- Social competence of the agent is correlated with explainability ($\rho = .43$, $p < 0.0001$) **(RQ1)**.
- Question-answering accuracy is correlated with explainability ($\rho = 0.15$, $p = 0.01$) **(RQ2)**.
- Accuracy is correlated with trust ($\rho = 0.16$, $p = 0.012$) **(RQ2)**.
- **Probability Scores** are rated as significantly less explainable than **Counterfactual** ($p < 0.001$), **Case Based** ($p < 0.001$), **Templated Language** ($p < 0.001$), **Feature Importance** ($p < 0.01$), **Crowd Sourced** ($p = 0.025$), and **Decision Tree** ($p = 0.039$) explanations **(RQ3)**.
- **Counterfactual** explanations are rated as significantly more explainable than **Nothing** ($p = 0.012$).

5.2. Perceptions of Social Competence and Trust in xAI

We applied Spearman's correlation with explainability as the independent variable and social competence as the dependent variable. We found that explainability was significantly correlated with impressions of the agent's social competence ($\rho = 0.43$, $p < 0.0001$) as measured by our xAI survey and the Godspeed survey (Bartneck et al., 2009). We did not find any statistically significant change in perceptions of social competence or intelligence of the virtual agent across our conditions.

Next, we applied Spearman's correlation with explainability as the independent variable and trust as the dependent variable. We found that explainability was significantly correlated with trust ($\rho = 0.56$, $p < 0.0001$), as measured by our xAI survey and the trust-in-automation survey (Jian et al., 2000). We further found that no individual condition in our study was rated as significantly more trustworthy than any other. Finally, we did not find statistically significant trends for compliance with the agent suggestions in our study nor for reliance on the agent suggestions (i.e.,

accepting correct advice or accepting incorrect advice). Neither explainability nor condition had any effect on the number of times that participants chose to accept the agent's advice, suggesting that trust was unrelated to participants' proclivity to accept advice from the agent (either correct advice or incorrect advice). We include results for participants' self-reported agreement with incorrect agent suggestions in the appendix, showing significant differences between the **Decision Tree**, **Feature Importance**, and **Templated Language** conditions. As our self-reported agreement and understanding results are drawn from a single Likert item rather than a full scale with multiple correlated items, we do not report those results in the main body of this work.

5.3. Objective Performance

By applying Spearman's correlation with explainability as the independent variable and performance as the dependent variable, we found that explainability was also correlated with human-machine team performance (i.e., decision-making accuracy) in our study ($\rho = 0.15$, $p = 0.01$), as measured by our xAI survey and the participant's final score on the primary task of our study. We additionally found that trust was correlated with accuracy ($\rho = 0.15$, $p = 0.012$) via Spearman's correlation with trust as the independent variable and accuracy as the dependent variable. While the agent offered more correct than incorrect suggestions, our results on compliance with the agent suggest that participants did not blindly rely on agent suggestions in any condition, regardless of their trust in the agent. We therefore hypothesize that the correlation between trust and accuracy is independent of the number of correct suggestions provided by the agent, though this hypothesis must be tested in future work. Finally, we did not find any statistically significant change in accuracy across our conditions. We found no effects for explainability nor condition (i.e., xAI method) on completion time.

5.4. Explainability by Condition

An ANCOVA showed that certain conditions in our experiment were rated as significantly more explainable than others ($F_{7,277} = 4.20$, $p < 0.001$). Our independent variable is the explainability method (condition) and our dependent variable is the participant's score on our xAI survey. We include, as a covariate, participants' baseline xAI survey scores after the priming task. A Tukey's HSD post-hoc analysis revealed that **Probability Scores** scored significantly lower on our xAI survey than all other explanation conditions, including **Counterfactual** (Cohen's $d = 0.918$, $SE = 0.25$, $p < 0.001$), **Case Based** ($d = 1.093$, $SE = 0.26$, $p < 0.001$), **Templated Language** ($d = 0.738$, $SE = 0.24$, $p < 0.001$), **Feature Importance** ($d = 0.608$, $SE = 0.24$, $p < 0.01$), **Crowd Sourced** ($d = 0.714$, $SE = 0.27$, $p = 0.025$), and **Decision Tree** ($d = 0.597$, $SE = 0.25$, $p = 0.039$) explanations. Similarly, **Counterfactual** explanations scored higher than the **Nothing** condition ($d = 0.701$, $SE = 0.23$, $p = 0.012$). Finally, we find no statistically significant differences in xAI survey scores from our priming task across the experimental conditions in our between-subjects design (Section 4.5) ($F_{7,277} = 1.338$, $p = 0.232$). We therefore attribute the differences in xAI ratings to the differences among the conditions rather than any differences between the subjects, who were randomly assigned to the experiment conditions..

Our results shed interesting insight into **RQ3** and the effects of different xAI

conditions on explainability. We found that the **Probability Scores** condition was statistically significantly worse than all other approaches to explainability and scored the lowest of all conditions on our xAI survey.

6. Discussion

6.1. Trust in xAI

Our results showed that trust and explainability (**RQ1**) were correlated measures. We found that an increase in explainability was correlated with an increase in participant-rated trust. Surprisingly, no condition was rated as significantly more trustworthy than another, despite the strong correlation between explainability and trust.

In finding a positive correlation between trust and explainability, we confirmed the intuition that an explainable agent is inherently more trustworthy. Regardless of the mechanism of explainability, an agent that is *perceived* to be more explainable is rated as more trustworthy. This finding also supports the validity of our xAI survey, as our explainability metric is correlated with the validated trust-in-automation survey (Jian et al., 2000). Importantly, there are not any overlapping questions between the trust-in-automation survey and our xAI survey, and each survey targets fundamentally different topics. While the trust-in-automation survey asks for ratings with respect to the robot, our xAI survey is entirely centered around the explanations and their utility. Regardless of these distinctions and differences, we find a significant correlation between the two measures. This significant correlation is therefore not a function of survey overlap or redundancy – instead we find that the concepts of explainability and trust are truly correlated.

Observing no significant difference in trustworthiness across our conditions (**Case Based**, **Counterfactual**, **Crowd Sourced**, **Decision Tree**, **Feature Importance**, **Nothing**, **Probability Scores**, **Templated Language**), we stumbled upon a surprising result. Despite our intuition that certain conditions would be distinguished by trustworthiness, we did not find a statistically significant difference in trustworthiness by method. While it is reasonable to expect that an agent that uses natural-language would be perceived as more relatable and trustworthy or that an explicit decision-tree would be more simulatable and verifiable, we found no condition was significantly more trustworthy than another. In conclusion, we found that none of our categories of xAI methods was definitively more trustworthy than any other.

Our findings suggest important avenues for future research. First, we corroborated the initial findings of prior work (Poursabzi-Sangdeh et al., 2021) that explainability alone will not reduce human over-reliance on automated decision-aids. Our work generalizes this result, showing that compliance is nearly constant for all xAI methods in our study. Therefore, future work must devise new approaches to human-agent interaction that specifically target compliance and reliance, as such effects will not simply be resolved by developing more explainable decision-aids. Second, future research into trust and explainability must consider domain details and investigate the utility of personalization. Future xAI systems will likely need to meet users halfway, conforming to their preferred mode of explanation in order to maximize trust and utility (Ehsan & Riedl, 2020). Our proposed xAI survey helps to guide such work, providing a quantitative benchmark for human-rated explainability of an agent partner.

6.2. Objective Performance

Our results regarding performance and explainability (**RQ2**) were mixed. We found that participants performed slightly better when they perceived their virtual agent assistant to be more explainable ($p = 0.01$). Furthermore, we found that trust in the agent was a factor in this result and accounts for some portion of the participant's increase in performance. As our virtual agent provided the correct answer for nine of the fourteen scenarios, it is reasonable to expect that participants who trusted the agent's suggestions would have an above-average final score in our study. We speculate that, in a study with more questions or a virtual agent with higher accuracy, this effect would be more pronounced and there would be a stronger correlation between accuracy and explainability.

While we found that explainability was correlated with accuracy, we did not find any effect of condition or explainability on completion time. This result is surprising, as one might expect the **Nothing** condition to have the lowest completion time (because there is less information to review in each scenario). Instead, we found that *no* condition was significantly faster than any other. Again, this result suggests that efficiency may be domain-, or individual-specific, and that adapting to users may enable improved human-agent team fluency (Ehsan & Riedl, 2020).

Our findings suggest that explainability significantly improves performance for question-answering tasks and did not reveal an efficiency penalty incurred by adding explainability. This result is significant, as it suggests that there be a minimal efficiency *cost* associated with deploying xAI and there is a performance *benefit* to be gained by leveraging xAI.

6.3. Social Competence

Our findings support the notion that an explainable agent is perceived as more socially competent (**RQ1**). Participants rated their virtual agent assistants much higher on the Godspeed questionnaire (Bartneck et al., 2009) when they perceived those agents to be more explainable. This finding again validates our xAI survey, as our explainability metric is once again correlated to the previously-validated Godspeed questionnaire (Bartneck et al., 2009), and the correlation between the two is expected for a reasonable explainability metric. Notably, our survey asks fundamentally different questions from the Godspeed questionnaire, as we drive at the utility and explainability of an agent rather than its perceived intelligence and likeability.

Interestingly, we did not find significance across conditions for perceptions of social competence or intelligence in our xAI conditions. This finding is not as surprising, as all conditions used similar images of NAO agents, all conditions included some form of natural-language communication, and all agents had single-character names. These name, appearance, and communication modalities would likely have made a difference in social perceptions of the agent, as prior work has demonstrated that anthropomorphism plays a significant role on such metrics (Natarajan & Gombolay, 2020). This result indicates that the appearance and communication modalities of an agent may be greater factors in social perceptions of agents than xAI mechanisms. Despite not finding significant differences across conditions, our results show that any agent that is perceived to be more explainable will also be perceived as more socially competent.

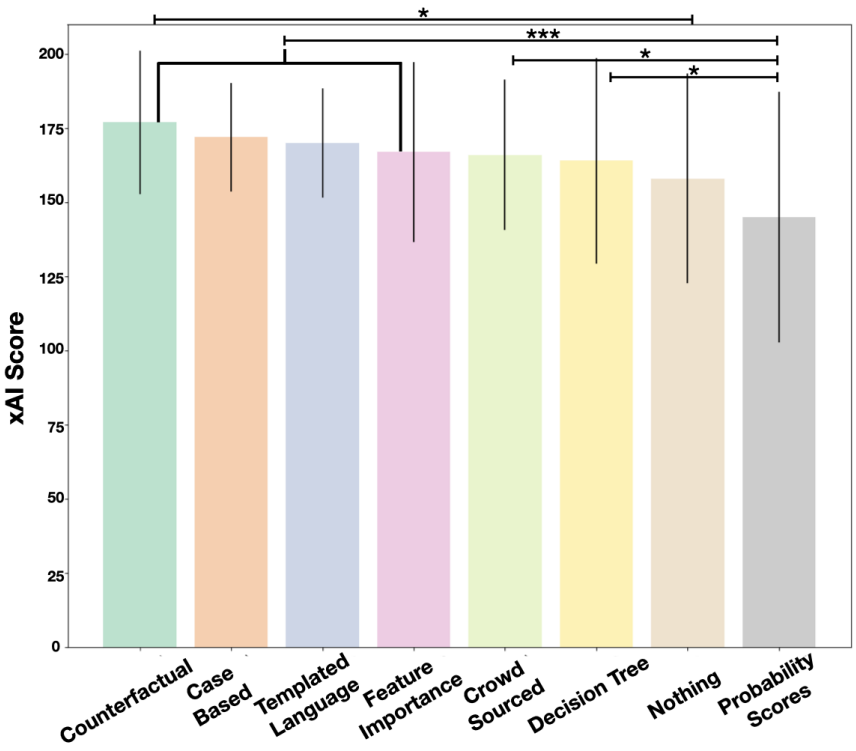


Figure 4.: xAI mean scores for all methods. Our results show that **Probability Scores** scored significantly lower on our xAI survey than all other explanation conditions, including **Counterfactual** ($p < 0.001$), **Case Based** ($p < 0.001$), **Templated Language** ($p < 0.001$), **Feature Importance** ($p < 0.01$), **Crowd Sourced** ($p = 0.025$), and **Decision Tree** ($p = 0.039$) explanations, and **Counterfactual** explanations scored higher than the **Nothing** condition ($p = 0.012$).

6.4. Explainability by Condition

Our results shed interesting insight into **RQ3** and the effects of different xAI conditions on explainability. We found that the **Probability Scores** condition was significantly worse than other approaches to xAI, including **Decision Trees**, **Crowd-Sourced**, **Feature Importance**, **Case Based**, **Templated Language**, and **Counterfactual**.

Our intuition regarding explainability by condition is that the simplest or clearest explanations are the explanations which receive the highest scores according to our xAI metric, as supported by prior research on simplicity in explanations (Lombrozo, 2007). Simple natural language explanations, such as in the **Templated Language** and **Counterfactual** conditions, were rated as significantly more explainable than an obscure explanation such as in the **Probability Scores** condition. Additionally, we found **Counterfactual** was the only condition to be rated as significantly more explainable than **Nothing**. We found further support for this observation in examining two pairs of very similar conditions in our study: **Templated Language** vs. **Feature Importance**, and **Crowd Sourced** vs. **Probability Scores**. Recall that **Templated Language** and **Crowd Sourced** presented the top feature/answer in the form of a sentence, while **Feature Importance** and **Probability Scores** presented a table of features/answers and probability scores for each. In both pairs of conditions, both **Templated Language** and **Crowd Sourced** *removed* information, yet were rated

higher for explainability than their probability-weighting counterparts.

Our results suggest interesting avenues for future work. We observe that one condition that did not rely on natural language was still rated very highly: **Case Based** explanations. This observation suggests that case-based reasoning may be a fruitful avenue for explainable decision-aids across many other domains, particularly those for which it is well suited (Caruana et al., 2015). Similarly, we find that **Templated Language** receives above-average xAI ratings on our survey, despite the lack of clear "features" to be used in the language for many scenarios.

Finally, future work should consider ways to maximize the faithfulness of counterfactual explanations. We find strong support for **Counterfactual** explanations as an avenue for explainability with human participants, being rated as the most explainable of all of our conditions and supported by research on human factors (Miller, 2019). However, recent research (White & Garcez, 2021) suggests that counterfactual explanations are not often actionable or understandable explanations and may be poor approximations of black-box model logic. Additional research on methodologies for faithful construction of counterfactuals may help to yield powerful and readily usable xAI technology.

7. Future Work

Our work introduces a survey designed to measure human-rated explainability of different xAI mechanisms. In future work, we plan to validate this survey through additional studies with new participant populations. We will also seek to replicate our study in additional domains and with different participant populations (e.g., domain experts). Finally, we aim to create a concise survey to be used by domain experts when evaluating suggestions by xAI agents. In this regard we created a shortened survey using factor analysis methods to remove redundancy. The validation of this shortened survey is also left to future work. In the remainder of this section, we present the design of this reduced xAI survey.

The participants in our study completed our 30-question xAI survey and we used their responses to create a reduced version of the full survey that measures the same primary components. We first conducted a factor analysis to analyze the different questions in our survey (Spearman, 1904; Watkins, 2018). After we removed all items with low factor loadings, the factor analysis reported that three factors were sufficient ($p = .165$). We also ensured that each factor had at least four items, as concepts such as usability, transparency, and simulatability are abstract and complex constructs that fewer items may not adequately capture (Schrum, Johnson, Ghuy, & Gombolay, 2020). We then tested the reliability of each subscale using Cronbach's alpha with $\alpha_1 = 0.83$, $\alpha_2 = 0.82$, and $\alpha_3 = 0.81$. This process resulted in a 14-question survey that approximately captures explainability along these three axes– transparency, usability, and simulatability. Our 14-question xAI survey is given in the appendix. We further analyzed the reliability of the survey by sampling a third of the data and recalculating Cronbach's alpha for the subscales. After taking fifteen samples, we calculated the mean, standard deviation, and 99% confidence interval for the Cronbach's alpha for each subscale: Factor 1 ($M = .824$, $SD = .031$, 99% $CI = (.804, .845)$), Factor 2 ($M = .823$, $SD = .037$, 99% $CI = (.798, .848)$), Factor 3 ($M = .809$, $SD = .016$, 99% $CI = (.798, .820)$). These results show that the subscales for transparency, usability, and simulatability consistently have internal reliability ($\alpha > .7$). The final "explainability score" for our xAI survey is computed as the sum of all items (with negative items

inverted), as in the full 30-question survey. We present analysis of our results using the reduced 14-question xAI survey in the appendix, and we leave validation of this reduced survey to future work.

8. Limitations

The primary limitation of our work is that our task was limited in scope, being confined to a set of multiple-choice questions involving common sense inference. Our study included a set of scenarios that did not significantly test any pre-existing knowledge or reading comprehension. Our study drew on common sense and inference about everyday life, revealing statistically significant differences across our population of untrained users. Additional deployments of our study targeted at populations of experts (such as medical professions, pilots, engineers, etc.) may yield additional domain-specific and nuanced results around compliance, trust, and performance.

Our study could also be extended by applying real xAI techniques to produce explanations for participants as explanations in our study were generated via WoZ. However, a deployment of our study with state-of-the-art language generation systems or feature-importance mechanisms may yield additional insights into the current failings of xAI research.

Finally, we have produced a reduced version of our xAI survey that correlates with the full version, but has not been empirically validated or verified through independent study. Future work will investigate the legitimacy of the reduced, 14-Q xAI survey as a tool for measuring participant-rated explainability.

In an effort to overcome these limitations in future work and to facilitate deployment of studies similar to our own, we provide study resources (e.g., survey files, questions, and tests) to the community. By leveraging our resources, other researchers will be able to quickly deploy their own versions of our xAI study to different domains or populations. By deploying more xAI studies to a wider variety of problems and demographics, we can begin to draw broader conclusions about xAI applied to a broader variety of challenges.

9. Societal Impact

Our work offers insight into the benefits of explainability when deployed to a question-answering task with a population of non-expert users. In finding support for explainability improving trust, social competence, and performance, we hope that our work will encourage the wider adoption and deployment of xAI to the real world.

Generalizing findings of prior work (Poursabzi-Sangdeh et al., 2021), we found several classes of xAI techniques may yield increased reliance on an agent decision-aid, even when such a decision-aid is incorrect. Therefore, it is critical that future developments and deployments of xAI take this finding into account. Without further research on how to mitigate such over-reliance when deploying explainable agents, xAI may inadvertently lead experts to make more mistakes, while simultaneously reinforcing such mistakes with inaccurate explanations.

10. Conclusion

In this work, we have described the design and results of a study to provide the first quantitative insights into the effects of explainability on trust, performance, and perceptions of social competence of virtual agents. We found that explainability was significantly correlated with trust, accuracy, and social competence, and that such findings were not dependent upon the method of explainability. We further found that simple language-based explanations and case-based explanations were all perceived as significantly more explainable than class-wise probability scores. Finally, we have proposed an xAI survey to measure human ratings of explainable AI, supported via correlations to trust, performance, and social competence. Our survey will be verified in future work, and will help xAI researchers more rigorously evaluate their work with human participants with a standardized measurement scale that can be applied to any xAI deployed to human users.

11. Acknowledgments

This work was sponsored by MIT Lincoln Laboratory grant 7000437192, NASA Early Career Fellowship grant 80HQTR19NOA01-19ECF-B1, a gift to the Georgia Tech Foundation from Konica Minolta, Inc, and the National Science Foundation (20-604).

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160.

Agarwal, N., & Das, S. (2020). Interpretable machine learning tools: A survey. In *2020 ieee symposium series on computational intelligence (ssci)* (pp. 1528–1534).

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, *18*(1), 8753–8830.

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., & Rudin, C. (2021). Iaia-bl: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *arXiv preprint arXiv:2103.12308*.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, *1*(1), 71–81.

Basak, J. (2004). Online adaptive decision trees. *Neural computation*, *16*(9), 1959–1981.

Bastani, O., Pu, Y., & Solar-Lezama, A. (2018). Verifiable reinforcement learning via policy extraction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc.

Bedny, G., & Karwowski, W. (2003). A systemic-structural activity approach to the design of human-computer interaction tasks. *International Journal of Human-Computer Interaction*, *16*(2), 235–260.

Boies, K., Fiset, J., & Gill, H. (2015). Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. *The Leadership Quarterly*, *26*(6), 1080-1094. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1048984315000934`

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees*. CRC press.

Brooke, J. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4–7.

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior research methods*, *50*(6), 2586–2596.

Caruana, R., Kangarloo, H., Dionisio, J. D., Sinha, U., & Johnson, D. (1999). Case-based explanation of non-case-based learning methods. In *Proceedings of the amia symposium* (p. 212).

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (p. 1721–1730). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2783258.2788613`

Chen, C., & Rudin, C. (2017). An optimization approach to learning falling rule lists. *arXiv preprint arXiv:1710.02572*.

Chen, H., Chen, X., Shi, S., & Zhang, Y. (2021). Generate natural language explanations for recommendation. *arXiv preprint arXiv:2101.03392*.

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

Craven, M. W., & Shavlik, J. W. (1995). Extracting tree-structured representations of trained networks. In *Proceedings of the 8th international conference on neural information processing systems* (p. 24–30). Cambridge, MA, USA: MIT Press.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2019). Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International conference on human-computer interaction* (pp. 449–466).

Elaad, E., et al. (2015). The distrusted truth: Examination of challenged perceptions and expectations. *Psychology*, *6*(05), 560.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research*, *42*(3), 237–288.

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

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International journal of human-computer interaction*, *13*(4), 373–410.

Hase, P., & Bansal, M. (2020, July). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5540–5552). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.acl-main.491`

Hedlund, E., Johnson, M., & Gombolay, M. (2021). The Effects of a Robot's Performance on Human Teachers for Learning from Demonstration Tasks. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 207–215). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3434073.3444664`

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hogan, T., & Kailkhura, B. (2018). *Universal hard-label black-box perturbations: Breaking security-through-obscurity defenses* (Tech. Rep.). Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations:

the system causability scale (scs). *KI-Künstliche Intelligenz*, 1–6.
Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2018). A benchmark for interpretability methods in deep neural networks. *arXiv preprint arXiv:1806.10758*.
Hutton, A., Liu, A., & Martin, C. (2012). Crowdsourcing evaluations of classifier interpretability. In *2012 aaai spring symposium series.*
Jain, S., & Wallace, B. C. (2019, June). Attention is not Explanation. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3543–3556). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N19-1357`
Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, *4*(1), 53–71.
Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, *89*(6), 899.
Kailkhura, B., Gallagher, B., Kim, S., Hiszpanski, A., & Han, T. Y.-J. (2019). Reliable and explainable machine-learning methods for accelerated material discovery. *npj Computational Materials*, *5*(1), 1–9.
Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.
Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 353–362).
Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, *2*, 26-41.
Klein, G. A. (1993). A recognition-primed decision (rpd) model of rapid decision making. *Decision making in action: Models and methods*, *5*(4), 138–147.
Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885–1894).
Lage, I., Chen, E., He, J., Narayanan, M., Gershman, S., Kim, B., & Doshi-Velez, F. (2018). *An evaluation of the human-interpretability of explanation.*
Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350–1371.
Liao, Q. V., Gruen, D., & 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* (p. 1–15). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3313831.3376590`
Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.
Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.
Lloyd, G. E. R., & Lloyd, G. E. R. (1996). *Adversaries and authorities: Investigations into ancient greek and chinese science* (Vol. 42). Cambridge University Press.
Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.
Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232–257.
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0004370218305988`
Mishra, S., & Rzeszotarski, J. M. (2021, April). Crowdsourcing and evaluating concept-driven explanations of machine learning models. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1).

Retrieved from https://doi.org/10.1145/3449213
Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery*, *2*(4), 345–389.
Natarajan, M., & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (p. 33–42). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3319502.3374839
Nguyen, D. (2018, June). Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1069–1078). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1097
Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, *7*(3), 437–454.
Olaru, C., & Wehenkel, L. (2003). A complete fuzzy decision tree technique. *Fuzzy sets and systems*, *138*(2), 221–254.
O'Mara, E. M., Kunz, B. R., Receveur, A., & Corbin, S. (2019). Is self-promotion evaluated more positively if it is accurate? reexamining the role of accuracy and modesty on the perception of self-promotion. *Self and Identity*, *18*(4), 405-424. Retrieved from https://doi.org/10.1080/15298868.2018.1465846
Paleja, R., Ghuy, M., Arachchige, N. R., & Gombolay, M. (2021). The utility of explainable ai in ad hoc human-machine teaming. In *Proceedings of the conference on neural information processing systems (neurips)*.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.
Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, 114–133.
Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–52).
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 1135–1144).
Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? a proposal of user-centered explainable ai. In *Iui workshops* (Vol. 2327, p. 38).
Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature communications*, *11*(1), 1–11.
Rosenfeld, A. (2021). Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th international conference on autonomous agents and multiagent systems* (pp. 45–50).
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.
Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*.
Saragih, M., & Morrison, B. W. (2021). The effect of past algorithmic performance and decision significance on algorithmic advice acceptance. *International Journal of Human–Computer Interaction*, 1–10.
Schlenker, B. R., & Leary, M. R. (1982). Audiences' reactions to self-enhancing, self-denigrating, and accurate self-presentations. *Journal of Experimental Social Psychology*, *18*(1), 89-104. Retrieved from https://www.sciencedirect.com/science/article/pii/002210318290083X
Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., & van den Bosch, K. (2021).

Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, *154*, 102684. Retrieved from https://www.sciencedirect.com/science/article/pii/S1071581921001026

Schroeder, E., Tremblay, C. H., & Tremblay, V. J. (2021). Confidence bias and advertising in imperfectly competitive markets. *Managerial and Decision Economics*, *42*(4), 885–897.

Schrum, M. L., Johnson, M., Ghuy, M., & Gombolay, M. C. (2020). Four years in review: Statistical practices of likert scales in human-robot interaction studies. In *Companion of the 2020 acm/ieee international conference on human-robot interaction* (p. 43–52). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3371382.3380739

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, *146*, 102551. Retrieved from https://www.sciencedirect.com/science/article/pii/S1071581920301531

Silva, A., Gombolay, M., Killian, T., Jimenez, I., & Son, S.-H. (2020, 8). Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the twenty third international conference on artificial intelligence and statistics* (Vol. 108, pp. 1855–1865). PMLR. Retrieved from https://proceedings.mlr.press/v108/silva20a.html

Sokol, K., & Flach, P. (2020). Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 56–67).

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. Retrieved from http://www.jstor.org/stable/1412107

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*(3), 647–665.

Suau, X., Zappella, L., & Apostoloff, N. (2020). Finding experts in transformer models. *arXiv preprint arXiv:2005.07647*.

Swann, W. B., & Ely, R. J. (1984). A battle of wills: self-verification versus behavioral confirmation. *Journal of personality and social psychology*, *46*(6), 1287.

Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, *16*(1), 97-113. Retrieved from https://www.sciencedirect.com/science/article/pii/0167487094000326

Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 399–439.

van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, *144*, 102493.

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, *31*, 841.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 chi conference on human factors in computing systems* (p. 1–15). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3290605.3300831

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219-246. Retrieved from https://doi.org/10.1177/0095798418771807

Weiss, S. M., & Indurkhya, N. (1995). Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research*, *3*, 383–403.

White, A., & Garcez, A. d. (2021). Counterfactual instances explain little. *arXiv preprint*

arXiv:2109.09809.

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

Wiegreffe, S., & Pinter, Y. (2019, November). Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 11–20). Hong Kong, China: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D19-1002`

Wu, M., Hughes, M. C., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-second aaai conference on artificial intelligence.*

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th acm/ieee international conference on human-robot interaction (hri* (pp. 408–416).

Zhang, Y., Liao, Q. V., & 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* (pp. 295–305).

Appendix A. Reduced 14-Question xAI Survey

Echoing the results of our primary investigation with the full survey, here we present results according to our reduced xAI survey. An ANCOVA showed that certain conditions in our experiment were rated as significantly more explainable than others ($F(7, 277) = 3.14$, $p = 0.003$). Our independent variable is the explainability method and our dependent variable is the explainability score. We include as a covariate the participant's baseline explainability score. A Shapiro-Wilk test revealed that our data were not normally distributed, but we proceed with an ANCOVA due to a lack of non-parametric alternative and the robustness of the F-test (Cochran, 1947; Glass, Peckham, & Sanders, 1972; Hack, 1958; Pearson, 1931). A Tukey's HSD post-hoc analysis reveals that **Counterfactual** was rated as more explainable than **Probability Scores** ($p = 0.002$), as shown in Figure A1

The reduced questionnaire, after a factor analysis and verification is given in Table A1.

Appendix B. Understanding and Agreement

Our scenarios included two Likert items for every question in the study: "I understand the reasoning behind the agent's suggestion" and "I agree with the agent's suggestion." Taking results from all participants on all questions, we have 3,672 responses for each item. As each Likert item is not part of a full scale with additional context, prior work (Schrum et al., 2020) suggests that analysis on such data may lead to premature conclusions. However, owing to the added workload of a full Likert scale after each of the 20 scenarios in our study, we decided to reduce our data collection to only two items. Reducing the scale to two items drastically reduces the time and workload of our study, yet still presents interesting data for analysis. We acknowledge the limitations of statistical testing with single Likert items. As such, we present the following analyses as interesting case studies of single-item responses and as possible avenues for future work to explore further.

Table A1.: The reduced xAI Survey

Factor	**Question**
1	I had trouble using the explanations to answer the question.
1	I believe that the explanations would not help most people to answer the question.
1	Most people would not be able to apply the agent's explanations to the questions.
1	I would not understand how to apply the explanations to new questions.
1	The explanations were not relevant for the questions I was given.
2	The explanations were detailed enough for me to understand.
2	I understood the explanations within the context of the question.
2	The explanations provided enough information for me to understand.
2	The explanations were useful.
3	I am able to follow the agent's decision-making process step-by-step.
3	I would be able to repeat the steps that the agent took to reach its conclusion.
3	I understand why the agent used specific information in its explanation.
3	I could have applied the agent's reasoning to new problems, even if the agent didn't give me suggestions.
3	I believe that I could provide an explanation similar to the agent's explanation.

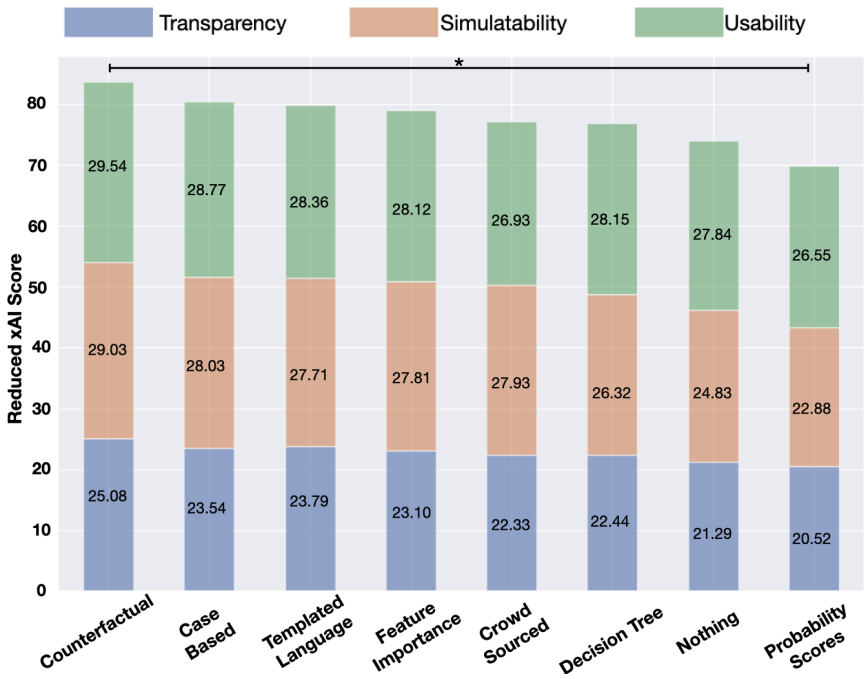
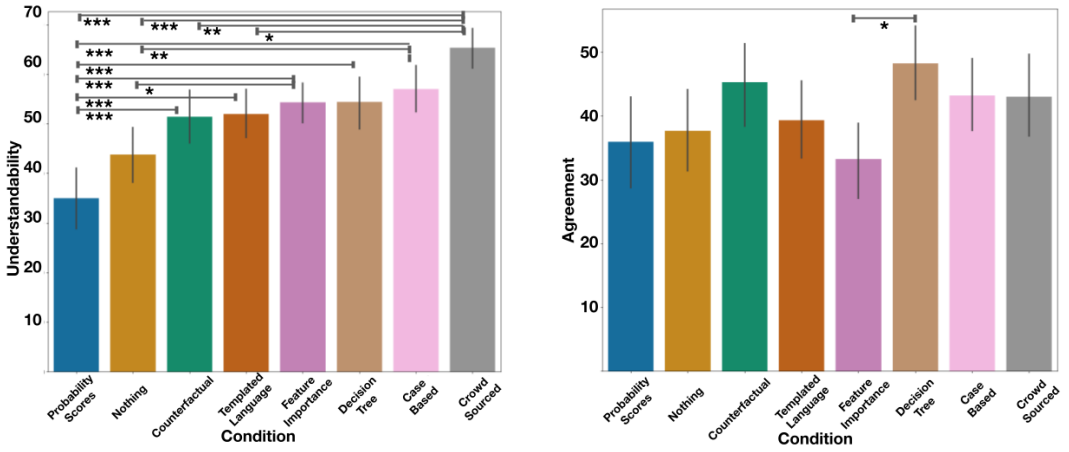


Figure A1.: Reduced xAI mean scores and sub-scale mean scores for all methods. The trends for the reduced scale match the full xAI scores, with the same ordering of conditions. While sub-scale scores for usability do not present much variance across conditions, the sub-scales of transparency and simulatability offer more variation across conditions. Statistical analysis of the aggregated reduced xAI scores reveal that counterfactual explanations score higher than probability scores ($p < 0.05$).

B.1. Understandability

An ANOVA showed that certain conditions in our experiment were rated as significantly more understandable for every question than others ($F(7, 3664) = 10.29$, $p < 0.001$). A Tukey's HSD post-hoc analysis revealed that the **Case Based** ($M = 56.98$, $SD = 50.85$), **Counterfactual** ($M = 51.41$, $SD = 59.79$), **Crowd Sourced** ($M = 65.37$, $SD = 38.43$), **Decision Tree** ($M = 54.45$, $SD = 58.39$), **Feature Importance** ($M = 54.35$, $SD = 48.24$), and **Templated Language** ($M = 52.02$, $SD = 58.12$) conditions were all rated as more understandable than the **Probability Scores** ($M = 34.98$, $SD = 60.82$) condition ($p < 0.001$). Similarly, the **Nothing** ($M = 43.84$, $SD = 61.54$) condition was rated as less understandable than the **Case Based** ($p = 0.006$), **Crowd Sourced** ($p < 0.001$), and **Feature Importance** ($p = 0.0497$) conditions. Finally, **Crowd Sourced** was rated as significantly more understandable than both the **Counterfactual** ($p = 0.007$) and **Templated Language** ($p = 0.011$) conditions. A comparison of all understandability ratings is shown in Figure B1a.

These results are very surprising. The **Crowd Sourced** condition presents the same information as the **Probability Scores** condition, the only difference is that the top confidence score is placed into a natural-language sequence and the three unused confidence scores are removed. For example, instead of showing a table with 85%, 10%, 5%, and 0%, as in the **Probability Scores** condition, the **Crowd Sourced** condition presents the sentence "85% of experts agreed on this answer." Despite presenting the same probability for the suggested answer in slightly different ways, we observe a



(a) (b)

Figure B1.: (a) Condition has a significant effect on participant understanding of agent suggestions, revealing that all xAI techniques are superior to softmax confidence scores, and three techniques (feature importance, case-based reasoning, and crowd-sourced scores) are superior to the "no explanation" condition. (b) Condition has a significant effect on participant agreement with an agent, with decision tree explanations prompting significantly more agreement.

significantly higher tendency for users to rate the **Crowd Sourced** condition as more understandable.

One possible reason for this disparity is in the wording of the prompt: "I understand the reasoning behind the agent's suggestion." While a set of confidence scores do not offer insight into *why* the agent arrived at an answer, saying that "85% of experts agreed on this answer" provides participants with enough information to infer the agent's reasoning. It is reasonable to assume that the agent chose the answer because the largest portion of experts agreed upon the answer. Despite the quantitative information being identical, users have more to infer with the **Crowd Sourced** condition. This line of reasoning may also explain the relative superiority of the **Case Based** condition, as users may infer that the agent's decisions arise from past experience. It is possible that our participants interpreted the prompt to be "I understand how this agent was trained," and imagined possible training data involving expert opinions or prior cases.

Finally, and amusingly, we note that even the **Nothing** condition achieves a higher mean-understandability score than the **Probability Scores** condition. Our hypothesis is that confidence scores are not useful signals to untrained human users, offering little insight into the decision-making process or the imagined training process of an agent assistant. Even having no information at all may be less confusing to human users.

B.2. Agreement

An ANOVA showed that participant-rated agreement was also significantly affected by condition, albeit to a lesser degree ($F(7, 3664) = 2.63$, $p = 0.011$). A Tukey's HSD post-hoc analysis revealed that participants were more likely to agree with an agent in the **Decision Tree** ($M = 48.23$, $SD = 63.13$) condition than in the **Feature Importance** ($M = 33.22$, $SD = 68.75$) condition ($p = 0.0136$). A comparison of

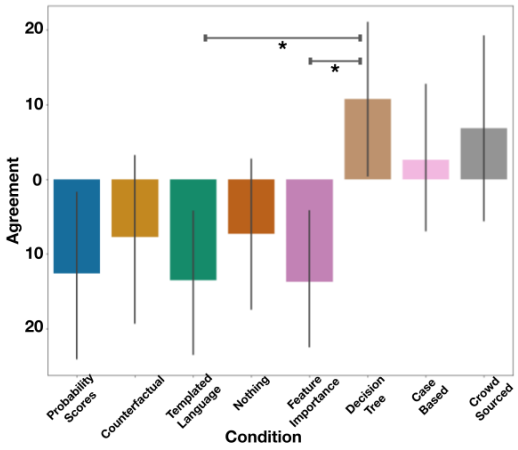


Figure B2.: Condition has a significant effect on participant agreement with an agent when the agent is offering incorrect suggestions, with decision tree explanations prompting significantly more agreement than feature importance scores or templated-language explanations.

all agreement ratings is shown in Figure B1b. We did not observe many statistically significant relationships between condition and participant-rated agreement with the virtual agent, which again corroborates our findings on xAI and compliance – namely, that compliance is unaffected by xAI condition.

When we specifically investigated agreement with *incorrect* suggestions, an interesting trend appeared. An ANOVA showed that participant-rated agreement was significantly affected by condition ($F(7, 1399) = 3.26$, $p = 0.0019$). A Tukey's HSD post-hoc revealed that, again, participants were more likely to agree with an agent in the **Decision Tree** ($M = 10.76$, $SD = 68.58$) condition than in the **Feature Importance** ($M = -13.71$, $SD = 66.20$) or **Templated Language** ($M = -13.51$, $SD = 71.35$) conditions at significance levels $p = 0.018$ and $p = 0.0203$, respectively. We also observe that, of all of our conditions, *only* **Case Based**, **Crowd Sourced**, and **Decision Trees** have *positive* average agreement scores. Results are shown in Figure B2.

Taking into context our results between xAI condition and the participants' compliance with the agent's suggestions, this result is surprising. Despite five of our eight conditions exhibiting *negative* average agreement with the virtual agent for incorrect suggestions, we do not observe significant differences between conditions for inappropriate compliance. In other words, our study suggests that untrained human users may accept an agent's suggestion *even if they disagree with the agent and the agent is wrong!* It is possible that, despite disagreeing with the agent, users were fooled by the agent's confidence in its suggestions (e.g., the agent never says "I'm not sure" or "Maybe the answer is..."), as there is abundant psychology research to suggest that humans tend to over-trust confidence (Elaad et al., 2015; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005; O'Mara, Kunz, Receveur, & Corbin, 2019; Rollwage et al., 2020; Schlenker & Leary, 1982; Schroeder, Tremblay, & Tremblay, 2021; Swann & Ely, 1984; Thomas & McFadyen, 1995). If someone got a previous question wrong, they might lose confidence in themselves and want to take the agent's suggestions (Hedlund, Johnson, & Gombolay, 2021). This result signals a need for xAI research to empower human users to actively challenge or interrogate their agent assistants, or

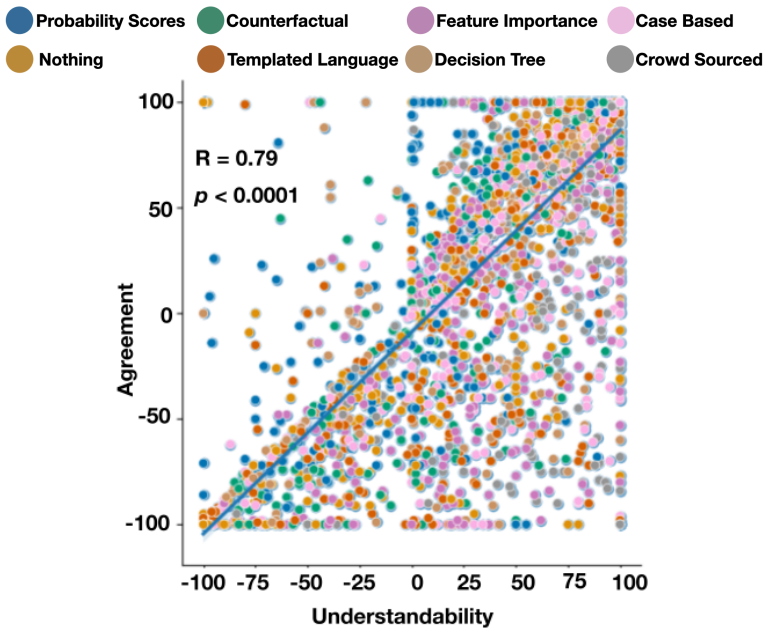


Figure B3.: Participant subjective agreement with agent suggestions is strongly correlated with participant understanding of agent suggestions ($R = 0.79$, $p < 0.0001$)

for xAI agents to regularly remind users of their fallibility (Natarajan & Gombolay, 2020). At present, our results suggest that users may be feeling pressure to accept agent suggestions even if they do not agree with such suggestions.

Finally, Pearson's correlations revealed that agreement, understandability, accuracy, and compliance were all statistically significantly correlated ($p < 0.0001$). Of these correlations, understandability and agreement were the strongest ($R = 0.79$), followed by agreement and compliance ($R = 0.41$), understandability and compliance ($R = 0.32$), agreement and accuracy ($R = 0.16$), and understandability and accuracy ($R = 0.13$). A comparison of understandability and agreement is shown in Figure B3.

Appendix C. Scenarios

In this section we present all scenarios used in the study. Scenario 1 was presented alongside instructions with how to use the interface and work with the virtual robot, while the remaining scenarios did not include additional instructions or content (apart from associated explanations). Scenarios 2-6 were used as the priming task, and scenarios 7-20 were used as the main body of the study. For each question, the correct answer is highlighted in bold, and the robots incorrect suggestion (where applicable) is highlighted in bold and red.

C.0.0.1. 1. A soccer player arrives to the training facility early every day. After several months of rigorous training and practice, the player still hasn't managed to make it into the starting team, with too much competition for their preferred position. However, the player has significantly improved coordination, acceleration, and top-speed. Which position is the player most likely to play?

(1) **Attacker**
(2) Defender
(3) Midfielder
(4) Goalkeeper

C.0.0.2. 2. Mark has just started running, and is trying to train for a local marathon. The marathon is set to take place in a month, so Mark has been training very hard. Unfortunately, a week before the marathon, Mark suffered an injury. Where was Mark injured?

(1) Elbow
(2) Neck
(3) **Knee**
(4) Back

C.0.0.3. 3. John is preparing a garden behind his building. He dug up an old tree stump and cleared out weeds to prepare a vegetable box, and must now prepare the ground for seeds. How should John fill in the vegetable box before planting seeds?

(1) **Mixing soil and fertilizer**
(2) Mixing soil and clay
(3) Mixing clay and fertilizer
(4) Mixing clay and gravel

C.0.0.4. 4. Jane needs to attend a meeting on the other side of the country tomorrow. Her company will pay for her expenses, the top priority is for her to physically attend the meeting. What is the best way for Jane to get to the meeting on time?

(1) Take a train
(2) **Fly**
(3) Drive
(4) Take a bus

C.0.0.5. 5. Monica has been working from home for the past several months, and is constantly suffering from eye-strain and headaches from staring at her computer monitor all day. Which of the following is the **least likely** to help reduce Monica's headaches and eye-strain?

(1) Use a blue light filter on her computer
(2) **Do more work in the dark with the lights turned off**
(3) **Break up the day with walks outside**
(4) Regularly take breaks to stare at distant objects

C.0.0.6. 6. James stops at a lookout while driving across the country to rest. While there, he looks out across a herd grazing on a plain, composed of animals native to North America. Which animals is James looking at?

(1) Cattle
(2) Domestic Sheep
(3) Bobcats
(4) **Bison**

C.0.0.7. 7. Everyday at 8:00 AM and 6:00 PM, a person's pet needs to be fed a scoop of food. The pet's space in the house needs to be cleaned weekly and typically takes under an hour to clean. The pet needs to go to the vet every 6 months. What type of animal is the pet?

(1) Dog
(2) **Cat**
(3) Hamster
(4) Fish

C.0.0.8. 8. Jamie is an avid hiker. She loves to explore outdoors in cool weather with just a light jacket, without worrying about bugs or heavy rainstorms, and particularly enjoys venturing up into the mountains to walk along small streams. Unfortunately for Jamie, her allergies always flare up as flowers bloom. Which season is best for Jamie to go hiking

(1) Spring
(2) Summer
(3) **Fall**
(4) Winter

C.0.0.9. 9. Patrick has struggled to recreate a recipe he found online. He hasn't ever tried to cook this particular dish before, and he is finding it difficult to replicate precisely. Because of the ways that altitude can affect cook times in ovens, Patrick is finding that his finished product doesn't look like the example online. Which food is Patrick preparing?

(1) **Bread**
(2) Chicken
(3) Veggie Platter
(4) Homemade Chocolate

C.0.0.10. 10. Shelby loves to read. In the past year, she's finished several books by Dostoevsky and Tolstoy, and others set in a violent coup or in a gulag. Which genre does Shelby seem to prefer?

(1) Biographies
(2) Historical Fiction
(3) **Russian Literature**
(4) Gritty Fantasy

C.0.0.11. 11. Jean is busy training for the upcoming finals in her favorite sport. To train appropriately, Jean is dedicating an hour each day to stretching and warming up, and then alternating between 5 to 8 miles of distance training or an hour of speed training. Jean's teammates are also pushing themselves very hard, as they'll all be competing for first place. Which sport is Jean training for?

(1) Hurdles
(2) **Cross Country**
(3) **Swimming**
(4) Soccer

C.0.0.12. 12. Arlo spends his days on his feet. He is often talking to other people, though other people will only seldom have the opportunity to respond or to interject. Arlo's audiences often pay very close attention for an hour at a time, and then rotate out for a new audience. Which profession best matches Arlo?

(1) Doctor
(2) Stage Performer
(3) Lawyer
(4) **Teacher**

C.0.0.13. 13. Carl enjoys the same drink every day. After he arrives to work, stressed from the chaos of his commute, he always goes straight to the break room to catch up with co-workers. Carl usually takes this time to calm himself down and try to relax, not needing any more stimulation after his commute. Which drink does Carl prefer in the break room before beginning work?

(1) **Tea**
(2) **Coffee**
(3) Whisky
(4) Soda

C.0.0.14. 14. Charon's favorite music helps her get through tough workouts. When she is exhausted and worn-out, the predictable and energetic rhythms of her favorite songs will always help motivate her to finish. What type of music does Charon enjoy for her exercise?

(1) Cinematic Soundtracks
(2) **Jazz**
(3) **Rock**

(4) Classical

C.0.0.15. 15. Charlie lives 4 miles from his workplace in a city with heavy traffic. His workplace is near a subway station, but Charlie's house is 2 miles from a station and he doesn't like physical activity. Fortunately, his workplace offers bike racks in the parking deck. Which mode of transportation best fits Charlie's commute?

(1) Bike
(2) **Electric Scooter**
(3) Car
(4) Subway

C.0.0.16. 16. Jay is suffering from chronic headaches. He has been feeling bad for a few months, ever since a knee injury forced him to stop running every afternoon. With the added time, Jay has been much more active on social media, and he is excitedly considering a career as an influencer. Which of the following is likely the cause of Jay's headaches?

(1) **Increased screen time**
(2) Less running every afternoon
(3) A change in diet
(4) **Increased stress over his career choice**

C.0.0.17. 17. Taylor tried to bake bread for the first time last week. Unfortunately, he forgot to account for the mess created by kneading dough, causing him to coat his hands in sticky dough and his clothes in flour. He also misread the instructions, as his eyes were burning from chopping onions that he used in his dinner. Before he tries baking again tonight, which change to his outfit should Taylor make?

(1) Wear white clothes
(2) Put on gloves
(3) Wear goggles
(4) **Put on an apron**

C.0.0.18. 18. Persephone is trying to cut down a tree to make more space for her cars behind her house. After a few hours of exhausting work on a rainy summer day, she managed to get the tree down and out of her yard. However, she was left with a stump about four inches high and 10 inches across in the middle of her yard. How should Persephone deal with the stump?

(1) Use a sledgehammer to hit it down into the ground
(2) Use an axe to cut more of the trunk away
(3) **Use a shovel to dig it out**
(4) **Use a controlled fire to burn it away**

C.0.0.19. 19. George is trying to complete a tour of European capitals before he graduates. Next month, he will begin a study-abroad in Lyon, France where he will be

able to visit new cities every month. Having never visited Europe before, where will George go first?

(1) **Paris**
(2) Madrid
(3) London
(4) Berlin

C.0.0.20. 20. When the total solar eclipse crossed the country on a Wednesday afternoon, thousands of tourists flocked to a narrow band of space where they would be able to see the total eclipse. There were not many restaurants or shops to visit in the path of the eclipse. How was traffic on the roads after the eclipse passed?

(1) No traffic
(2) **Heavy traffic**
(3) Light traffic
(4) Moderate traffic