AFRL-RY-WP-TP-2023-0063

EVALUATING EXPLAINABLE AI (XAI) IN TERMS OF USER GENDER AND EDUCATIONAL BACKGROUND (Preprint)

Samuel Reeder, Joshua Jensen, and Robert Ball
Weber State University

MAY 2023
Final Report

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.
See additional restrictions described on inside pages

STINFO COPY

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE May 2023	2. REPORT TYPE Conference Paper Preprint	3. DATES COVERED	
		START DATE 24 April 2023	END DATE 24 April 2023

4. TITLE AND SUBTITLE
EVALUATING EXPLAINABLE AI (XAI) IN TERMS OF USER GENDER AND EDUCATIONAL BACKGROUND (Preprint)

5a. CONTRACT NUMBER FA8650-20-F-1956	5b. GRANT NUMBER N/A	5c. PROGRAM ELEMENT NUMBER 64294D/62204F/63680F
5d. PROJECT NUMBER 2002/5280	5e. TASK NUMBER N/A	5f. WORK UNIT NUMBER Y21D

6. AUTHOR(S)
Samuel Reeder, Joshua Jensen, and Robert Ball

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Weber State University 3848 Harrison Blvd, Ogden, UT 84408	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory, Sensors Directorate Wright-Patterson Air Force Base, OH 45433-7320 Air Force Materiel Command, United States Air Forces	10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RYDT	11. SPONSOR/MONITOR'SREPORT NUMBER(S) AFRL-RY-WP-TP-2023-0063

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
PAO case number AFRL-2023-1488, Clearance Date 24 April 23. This work was funded in whole or in part by Department of the Air Force contract FA8650-20-F-1956. The U.S. Government has for itself and others acting on its behalf an unlimited, paid-up, nonexclusive, irrevocable worldwide license to use, modify, reproduce, release, perform, display, or disclose the work by or on behalf of the U. S. Government. Report contains color.

14. ABSTRACT
For the purposes of this paper, unless specified, we will include both Machine Learn-ing (ML) and Artificial Intelligence into one single label of "AI."

15. SUBJECT TERMS
artificial intelligence, machine learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT Unclassified	b. ABSTRACT Unclassified	C. THIS PAGE Unclassified	SAR	27

19a. NAME OF RESPONSIBLE PERSON Todd James	19b. PHONE NUMBER *(Include area code)* N/A

PREVIOUS EDITION IS OBSOLETE.

Evaluating Explainable AI (XAI) In Terms of User Gender and Educational Background

Samuel Reeder, Joshua Jensen, Robert Ball

Weber State University, Ogden, UT 84408, USA

Abstract. The abstract should summarize the contents of the paper in short terms, i.e. 150-250 words.

Keywords: First Keyword, Second Keyword, Third Keyword.

1 Introduction

In 1950 Alan Turing published his landmark paper titled "Computing Machinery and Intelligence" where he proposed the question "Can machines think?" [1]. Since then artificial intelligence (AI) has now become a part of everyday life in the modern world with AI-driven systems being developed all around us. The use of AI powered systems has taken many forms ranging from machines that play chess better than humans [2] to complex networks that drive cars [3].

For the purposes of this paper, unless specified, we will include both Machine Learning (ML) and Artificial Intelligence into one single label of "AI."

Of particular concern to many people in the private, public, and military sectors is the use of AI in recommendation systems. Recommendation systems are the primary drivers for billions of dollars annually. From Amazom.com to Google.com, recommendation engines suggest and steer billions of people to particular products or search results.

If Alan Turing asked if machines can think then we ask the following question: How do machines reach their conclusions, particularly in recommendation systems? This question is answered by the subfield of Explainable AI (XAI). Explanations for recommendations from recommendation systems are important for a number of reasons, from legal compliance, de-bugging the system, curiosity about how the system works, etc.

In our related works section we show that there has been a large amount of work done to create many different types of explanations. However, little work has been done to take these explanations and evaluate how effective they are with actual people.

In this paper we share the results of an in-depth study that compares how people empirically react to different types of XAI, particularly different types of textual explanations and word clouds. Specifically, the purpose of this research is to investigate how explanations in recommendation engines affects a user's trust and comprehension in a recommendation system.

We found that gender and educational background had an interactive effect on both trust and under-standing and that different explanations were preferred by different groups. However, the word cloud XAI was overwhelmingly rejected by all groups.

In other words, we conclusively found that different types of XAI are preferred based on gender and educational background. In particular, we found that women's trust and understanding of the explanations did not vary significantly based on educational background, but that men's trust and understanding of explanations did statistically differ based on their backgrounds.

Although the study focused on recommendation engines, the results most likely generalize to all XAI.

Explaining why a decision was made by a person is a daunting task and not easily explained [4]. A study of history shows that even with all or most of the facts present, different motivations and accompanying explanations are posited for why individuals or groups made the choices that they did. Research into how people make decisions is an active topic (e.g., [5] and [6]).

In addition, distrust in technology and online technologies is a particular problem in today's world. Recent research shows that trust in online content can be based on simple things like the graphics and structure of websites [7]. Also, distrust in content, like in Facebook with terms like "misinformation," is a problem for many users [8].

Given many people's mistrust and non-understanding of how most technologies work, we address the following question: How effective are XAI solutions for actual people? We will show in the related works section that extensive amounts of work has been put into creating AI and ML algorithms and creating accompanying XAI, but little research in comparison has been done in evaluating how effective XAI is for actual people.

In this paper we answer that question by reporting on trust and understanding in a laboratory-controlled experiment. We exposed users to an explainable recommendation system and gauged the effects of the explanations through a series of questions. We utilized three approaches to demonstrating the explanations: simple textual, technical textual, and visual explanations.

We particularly focused on how people's back-grounds (STEM [Science, Technology, Engineering, and Mathematics] vs non-STEM [e.g., humanities, arts, etc.]) and gender (male vs female) affected their trust and understanding of the explanations given by the recommendation system.

2 Related Works

The importance of XAI cannot be overstated. For ex-ample, Dattner, et al explain in the Harvard Business Review that there are legal and ethical implications of using artificial intelligence-based systems to aid in the hiring process for businesses. The authors points out that since the underlying AI is not well understood it is unclear if these systems comply with nondiscrimination laws, the ADA (American with Disabilities Act), as well as others [9].

Another area where AI accountability is particularly important is when these systems are used for military applications. David Gunning and David W. Aha describe a system that can account for choices made as well as inform users of its shortcomings so that AI systems that could provide justification for the suggestions that were made. The purpose of this type of AI system is to provide a tool to the military that could help in making strategic choices while at the same time providing enough transparency to justify its use [10].

In the end, there are numerous examples that make the case that XAI is important and needed. Some ML algorithms, like neural networks, are not explainable and certain properties, such as the use of randomized neuron weights, as well as other factors contribute to this [11].

There are many other types of machine learning algorithms that suffer from this same issue in the con-text of comprehensibility. In general, the issue of comprehensibility in machine learning and AI is known as the "black box problem." In an exhaustive survey of this issue Guidotti, et al defines the black box problem as follows: "In recent years, many accurate decision support systems have been constructed as black boxes, that is as systems that hide their in-ternal logic to the user. This lack of explanation constitutes both a practical and an ethical issue" [12].

There have been attempts to make black box ML algorithms more comprehensible. For example, by using decision trees, which are considered white box algorithms and are comprehensive, to mimic the behavior of neural networks output [13]. This can be difficult, but others have proposed solutions with pruning the nodes in decision trees to improve comprehensibility [14].

Others have proposed using the concept of a rule set. The idea is that a black box model could be explained by generating a set of rules that humans can understand that dictate a prediction for a given input [15]. Support Vector Machines (SVM) have also been used with the concept of a rule set (e.g., [16] and [17]). Another example of complex models that can be explained via a rule set is tree ensembles [18]. Barakat and Bradley provide a general review of rule set implementations with different ML algorithms [19].

The idea of local explanations is presented by Guidotti et al. where they describe a model that is locally explainable as a model the has that following property: "Is able to explain the prediction of the black box in understandable terms for humans for a specific instance or record" [12].

An interesting example of this type of explanation is given in the work of Xu et al. In this work a system that automatically generates captions describing an image is given. The system writes one word per pass over the picture. The algorithm provides an ex-planation for why each word in the caption is included by showing the portion of the picture that was used to generate the word [20].

Zhang, et al. present a survey of explainable recommendation systems. The paper breaks down XAI into two subgroups: the method of explanation and the models the generate them. The paper also breaks down the types of explanations into five categories: based on relevant users/items, feature based explanations, textual based explanations, visual explanations, and social explanations [21].

Within the context of recommendation engines, text-based explanations often involve an attempt to leverage text from reviews about the items being recommended. Some form of natural language processing (NLP) is employed with the goal of attempting to determine a set of features or sentiment that could be used to explain why the recommendation was given. NLP often utilizes sentiment analysis as a means for providing explanations to a user. Zhang et al. provide an example of sentiment analysis with XAI. They describe a framework called Sentires. The Sentires framework takes reviews of a certain type of product and generates triplets in the forms of aspect-opinion-sentiment and are used as an explanation for an individual recommendation [22].

The key idea with sentiment analysis is to deter-mine if a user feels negatively or positively about a product. One of the challenges with doing this is a lack of labeled training data for recommendation systems to learn from. This problem is highlighted in by Guan, et al. They introduce a method for deriving a larger labeled set of training data for algorithms to train on. They use a semi-supervised model to add sentiment labels to unlabeled data [23].

There are many types of recommendation engines. A good overview of these systems can be found in [24]. At a high level, the idea is that a user will likely be interested in items that are similar to ones that they have indicated a preference for. Recommendation systems generally rely on prior knowledge about users and provide suggestions to the user based on that history.

Y. Zhang, et al. have noted that latent factor models have become popular in recommendation systems because of their high prediction accuracy. However, latent factors are black boxes and make explanations of recommendations difficult. To solve this issue, they propose a model referred to as explicit factor model. The idea is that reviews on a product or type of product can be used to identify features of a product as well as the sentiment users have about them. An example sentence follows, "You might be interested in [feature], on which this product performs well" [25].

Explanations like the one above are useful, but the templated sentence structure approach can feel mechanical. Other research has suggested more personalized text [26].

Explanations are sometimes visual. The concept of a visual explanation is when a product is recommended and the justification is provided in a visual format. This can take the form of a picture or a picture accompanying a textual explanation.

Wu and Ester provide an example of this principle in action. They describe an algorithm called Factorized Latent ModEL (FLAME) that attempts to solve the problem of personalized latent aspect rating analysis. The FLAME model presents the explanation for the recommendation as a word cloud. The word clouds generated aims to give the user a sense of what features are most prominent for the recommended product [27].

Visual aids may also be used to help augment text-based explanations as shown in the work of Lin, et al. The authors provide a system that pairs images of products with a sentence explaining the recommendation [28].

Another suggested presentation idea for XAI is by using graph-based models. He, et al. propose a system that utilizes a tripartite graph to rank aspect opinion data about items based on users' reviews. This graph is used to create the explanations that accompany recommendations. This approach was de-vised to address the shortcomings of matrix factorization models that are often used in collaborative filtering recommendation systems [29].

There are numerous other papers that suggest different ways to produce XAI for various ML algorithms. For example, the use of gradient boosted decision trees combined with matrix factorization they produce recommendations as well as generating hu-man comprehensible explanations for them [30] to various deep learning neural networks (e.g., [31] and [32]).

It is clear from the above works that there is a lot of interest in explainable recommendations. The overarching assumption is that that explainable recommendations are valuable because they increase a user's trust in the recommendation system. It is understandable why many people make that particular assumption. However it shouldn't simply be assumed. In the rest of this paper we will present re-search that attempts to provide some verification for this claim.

Our experiment was heavily impacted by the de-scription text-based explainable recommendation systems as was shown in [22], [23] and [25]. The concept of using the sentiment found in reviews to generate explanations is a good basis for comparing to other types of explanations. We also used the concept of using a word cloud as a means of explanation a recommendation as a form to incorporate some visual explanations. We used the technique from [37].

The recommendation system strategies that have been discussed so far represent the foundation of recommendation systems. In practice, these strategies are mixed and combined with many forms of advanced machine learning to create reliable recommendations. One of the more famous examples of this can be found in the "Netflix Prize" competition. In 2006 the video streaming company Netflix announced a coding challenge. The premise was that Netflix wanted to see if a team could develop an algorithm better than their Cinematch recommendation system. Any team who could produce a system that made a 10% or more improvement over the Cine-match system would be awarded one million dollars. The winning team that to beat the native Netflix recommendation engine had many different statistical and machines learning algorithms were blended to get the improved performance [33].

There have been many papers published on creating XAI and creating frameworks that provide direction for explanations. Gilpin, Leilani H., et al. provide a general survey of 87 XAI papers for numerous types of ML algorithms. They conclude that "for ma-chine learning systems to achieve wider acceptance among a skeptical populace, it is crucial that such systems be able to provide or permit satisfactory explanations of their decisions" [34].

One fair question to ask about these systems is can these complicated systems be decomposed such that individual recommendations be explained in a meaningful way? Furthermore, if they can be explained, does it matter to the users of the recommendation?

The goal of this paper is to provide some insight into the question of whether XAI matters or not to the user.

3 Experiment and Methodology

As stated above, the purpose of this research is to investigate how explanations in recommendation engines affects a user's trust and comprehension in a recommendation system based on their gender and educational and professional background. To do this we designed an experiment that exposes participants to a series of recommendations where each recommendation is paired with a collection of explanations. After viewing each recommendation, a participant was asked to respond to a series of questions.

The purpose of the questions was to gather information about how trust and comprehension develop over time as the participant progresses through the experiment. After viewing all the recommendations, the participants were asked to respond to a final set of questions.

3.1 Experimental Format

The experiment was delivered via website to each participant. There were three distinct parts: The tutorial stage, the recommendations stage, and the final survey.

The purpose of the tutorial section was twofold. First, to make sure that each participant was familiar with the mechanics of the experiment. The tutorial was designed with the goal of answering any questions a participant might have about how the experiment was to be taken. Each participant was given a phone number to call in the event they encountered any issues. The tutorial was delivered in the form of a video that each participant was invited to watch. The video presented a live walk-through of the pages in the tutorial section as well as commentary on the purpose of the experiment. The provided commentary and instructions were carefully designed not to introduce any bias towards our research questions.

During the tutorial, the participants were told that they were helping to validate the performance of a recommendation engine designed to recommend cookie recipes. To help with this task, participants were introduced to a current user of the recommendation software named "Steve," who was introduced via his user profile. The profile shows a list of Steve's favorite types of cookies, a list of his favorite ingredients, and a final list showing the names of Steve's favorite recipes that he had found through the recommendation software. Participants were told that they were supposed to imagine that they were Steve while they were viewing the recommendations and explanations. The purpose of this was to remove the need for each participant to have to create their own profile with enough data to overcome the cold-start problem.

The main feature of the tutorial was showing the participant how they were supposed to view recommendations as well as how to take the survey that accompanies each recommendation. During this part of the tutorial a demonstration recommendation was shown, each button on the screen explained, with emphasis given to the buttons that show the explanations. The survey was explained question by question with an explanation given for what individual questions mean as well as how the responses are to be input. This also included the necessary IRB disclaimers, and explanations required by our institution's IRB.

The last stage of the experiment involved collecting demographic data about each participant. Each participant was asked to provide their gender, age, and status as a technical or non-technical person. The definition of a technical person was given as any person who meets one of more of the criteria listed below:

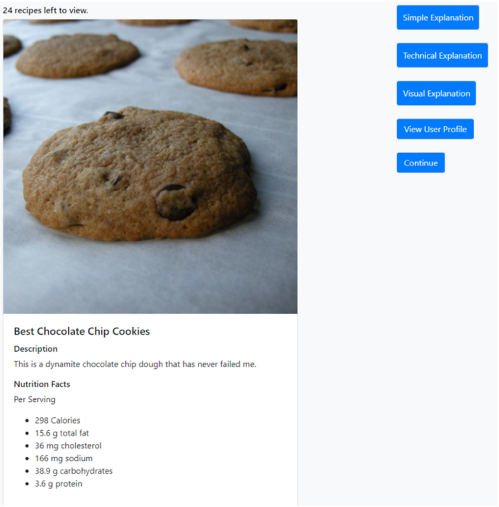
- Earned a degree in a STEM field,
- Currently enrolled in a university level computer science program,
- Or currently employed in a STEM related job.

Each step of the tutorial was shown in the recorded video, but participants were required to visit each page of the tutorial section to make sure that they understood the content of the video as well.

3.2 The Recommendation State

During the actual experiment, users viewed 25 different recommendations. Each recommendation had three explanations given as justifications for the recommendation. An example of how recommendations appear is given in Figure 1 below.

Participants were encouraged to view all the explanations for each recommendation, but the testing software did not require it. After each recommendation, participants were required to answer six questions.



24 recipes left to view.

Best Chocolate Chip Cookies

Description

This is a dynamite chocolate chip dough that has never failed me.

Nutrition Facts

Per Serving

- 298 Calories
- 15.6 g total fat
- 36 mg cholesterol
- 166 mg sodium
- 38.9 g carbohydrates
- 3.6 g protein

Fig. 1. An Example of an experimental recommendation. This is the layout of an experimental recommendation. Each recommendation displays a picture of the recommended cookie. Meta data about the cookie given. Users can click on buttons bellow the nutrition facts section to see directions for the recipe and ingredients (not shown in the figure). Explanations are viewed by clicking the buttons on the right-hand side of the screen.

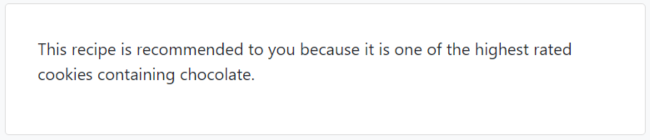
The last step of the experiment involved asking participants to give some feedback on the recommendation software. This feedback was collected in the form of four open-ended questions. The purpose of this last section was to gather information about each participant's perspective on key elements of the experiment. The hope was that this feedback could provide some insight and context to the results of the experiment.

3.3 The Recommendation System

The recommendation system used in this experiment provided recommendations about cookie recipes to its users. The recipe data was extracted from an online recipe website. For each recommendation, three types of explanations were provided to the user. These explanations are categorized as simple, technical, and visual.

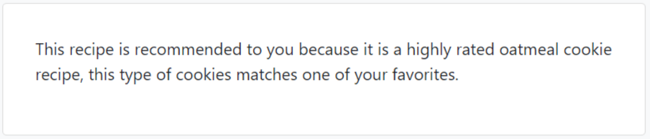
A simple explanation represents a basic type of explanation for a recommendation that a participant should be familiar with from using real recommendation engines. A simple explanation is one to two sentences long. The goal of the explanation is to show the user how the recommendation ties back to Steve's profile. Figure 2 and Figure 3 give examples of simple explanations that were used in the experiment.

The intent of this recommendation type is to provide an explanation that is more understandable and substantive than something like a star rating that you might encounter when looking at suggested products like on Amazon.com. The other goal of the simple explanation is to be basic enough that anyone would be able to understand it and connect it to Steve's profile. This type of explanation was inspired by some of the papers related works section, specifically references [23] and [25].



This recipe is recommended to you because it is one of the highest rated cookies containing chocolate.

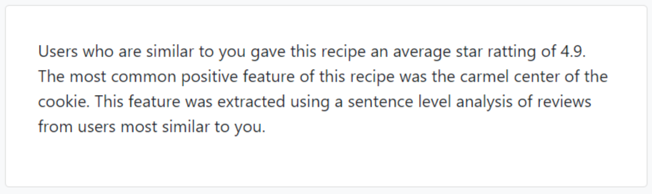
Fig. 2. An example of a simple explanation for "Carmel Filled Chocolate Cookies".



This recipe is recommended to you because it is a highly rated oatmeal cookie recipe, this type of cookies matches one of your favorites.

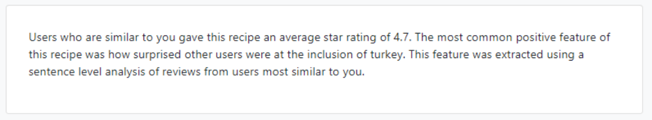
Fig. 3. Another example of a simple explanation given for a recipe called "Delicious Raspberry Oatmeal Cookie Bars."

The technical explanations build off the simple explanations by attempting to give some insight into how the recommendation engine may have generated the explanation by exposing the internal mechanism of the software. These explanations are still text based but provide an additional technical vantage point. Figure 4 and Figure 5 provide examples of these explanations.



Users who are similar to you gave this recipe an average star ratting of 4.9. The most common positive feature of this recipe was the carmel center of the cookie. This feature was extracted using a sentence level analysis of reviews from users most similar to you.

Fig. 4. An example of a technical explanation. This is the technical explanation given for the recipe “Carmel Filled Chocolate cookies.”



Users who are similar to you gave this recipe an average star rating of 4.7. The most common positive feature of this recipe was how surprised other users were at the inclusion of turkey. This feature was extracted using a sentence level analysis of reviews from users most similar to you.

Fig. 5. Another example of a technical explanation.

The motivation for this explanation type stems from research around feature sentiment-based explanations such as those described in [22], [23], and [25]. The other motivating factor is to provide the participants some signals about different types of data and analysis strategies that could have been used to generate the explanation. For instance, the mention of similar users hints at the use of some type of collaborative filtering method. The last sentence of each explanation gives a clue that some type of natural language processing might have been used as well.

The last type of explanation were visual explanations. For this explanation type we took inspiration from [27] and used word clouds as visual explanation. Figure 6 gives an example of a word cloud used in the experiment.

There are some draw backs to this approach. For instance, the word clouds do not do a great job of associating a sentiment with the features. For example, the word cloud shown in Figure 6 gives the most emphasis to the words chewy, cookies, butter, and peanut when read from the top to bottom of the image. Each of the words on its own could hold positive or negative sentiment for a participant and it is not clear what the sentiment of reviews that generated the world cloud are for the emphasized features.

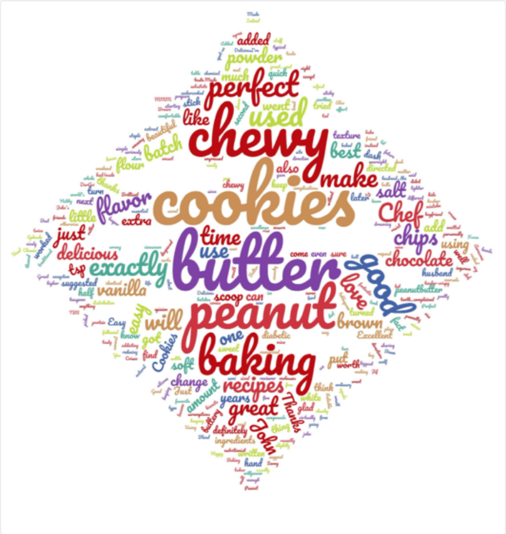


Fig. 6. An Example of a visual explanation. The word cloud shown here is meant to be an explanation for the recipe called "Chef John's Peanut Butter Cookies". to the terms used In the word cloud come from the most common words that appeared in reviews by other users.

3.4 Experimental Mechanics

For the purposes of experimentation and focusing on evaluating the explanations, each recommendation and the accompanying explanations were hard coded. This was so that the system used in the experiment would not introduce any additional variables to the experiment.

The hard-coded set of recommendations allowed the experiment to focus on the research questions without concern over recommendation variance or the accuracy of programmatically generated recommendation. If the experimental recommendation system generated a new set of recommendations for each participant there is no way to guarantee that each participant would have the same experience.

As mentioned above, the premise of the experiment given to the participant is that they were being asked to help validate a recommendation engine that gives recommendations about cookie recipes.

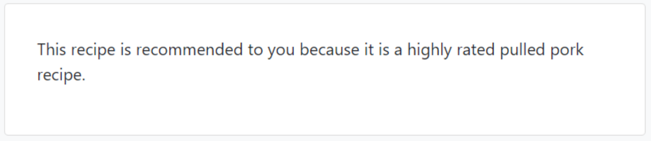
The goal of presenting this scenario to participants was to encourage them to pay attention while at the same time mask the primary purpose of the experiment. Since the experiment is aimed at gauging trust, we presented the experiment in a way that provided as little bias as possible where the software is concerned. The hope was that by asking the participants to validate the system it will give them some feeling of obligation to view at least some of the explanations.

The open nature of the experiment was meant to give each participant the opportunity for their trust and understand of the system to develop as organically as possible.

3.5 False Explanations

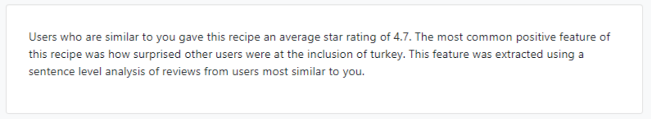
One of the key features of the experimental recommendation system was that five of the twenty-five recommendations provided false explanations. At a high level the purpose of these false explanations was to test participants to see if being presented with false explanations for an otherwise correct recommendation had an impact on their trust and comprehension.

For a recommendation that has false explanations each of the three types of explanations are explanations that are for a different recipe. Figures 7-9 show a set of false recommendations.



This recipe is recommended to you because it is a highly rated pulled pork recipe.

Fig. 7. An example of a false simple explanation. This simple explanation was given for a recipe called Beth’s Spicy Oatmeal Raisin Cookies.



Users who are similar to you gave this recipe an average star rating of 4.7. The most common positive feature of this recipe was how surprised other users were at the inclusion of turkey. This feature was extracted using a sentence level analysis of reviews from users most similar to you.

Fig. 8. An example of a false technical explanation. This technical explanation was given for a recipe called Beth’s Spicy Oatmeal Raisin Cookies.

The false explanations shown in figures 7-9 were inspired from a pulled pork recipe. The goal of these false explanations is to give the impression that the software was experiencing a bug.

One of the assumptions that the experiment had is that each participant has some level of trust in, and understanding of, recommendation engines. False explanations were given to provide a way to better understand how explanations were changing participants trust and understanding.

The other purpose of the false explanations was to provide a check to verify if the participants were paying attention to the explanations. The false explanations were obviously wrong so when analyzing the results if participants did not indicate that they had seen some questionable data then we could conclude that they had not read the explanations.

The false explanations were also important in measuring trust. Specifically, we tested this is by attempting to establish trust in the recommendations and then challenge that trust with a false explanation. To do this the experimental system presented the first five recipe recommendations with true explanations. The hope is that the participant would develop some level of trust while viewing these recommendations. Then on the sixth recipe recommendation the participant was given a set of false explanations. After that the remaining 4 false explanations were given at different intervals to the participants so as to appear to be random bugs in the system.

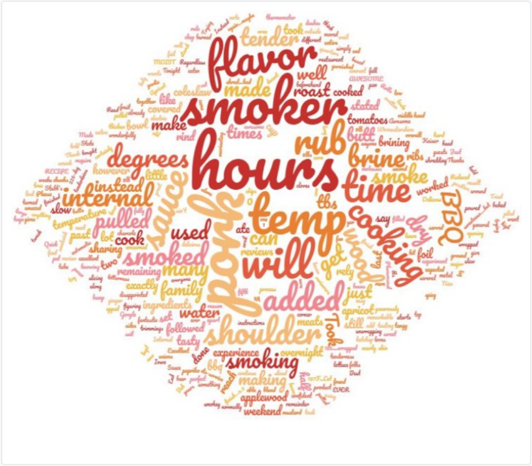
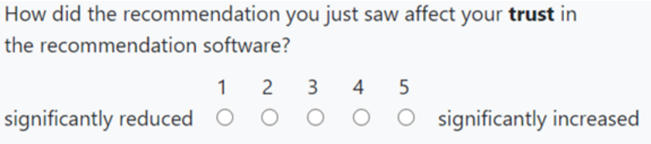


Fig. 9. An example of a false visual explanation. This visual explanation was given for a recipe called Beth's Spicy Oatmeal Raisin Cookies.

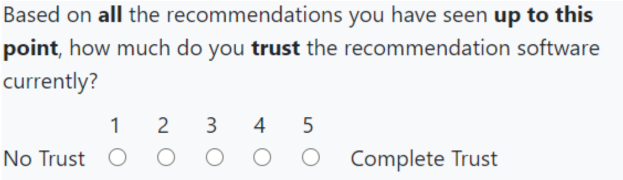
3.6 Measuring Trust and Understanding

In order to measure trust and comprehension, participants were presented with a series of questions. After each recommendation a participant was required to give an answer to six questions. The first four questions were based on a Likert scale. The questions were broken up into three categories with the first two questions shown in Figure 10 and Figure 11.


How did the recommendation you just saw affect your **trust** in the recommendation software?

	1	2	3	4	5	
significantly reduced	○	○	○	○	○	significantly increased

Fig. 10. Question 1 as it appeared to participants. How did the recommendation you just saw affect your trust in the recommendation software?


Based on **all** the recommendations you have seen **up to this point**, how much do you **trust** the recommendation software currently?

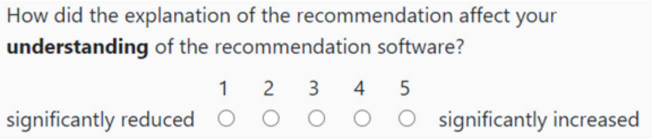
	1	2	3	4	5	
No Trust	○	○	○	○	○	Complete Trust

Fig. 11. Question 2 as it appeared to participants. Based on all of the recommendations you have seen up to this point, how much do you trust the recommendation software currently?

The general idea with these questions is that over time the different values reported can be interpreted as a trend for how the participants' trust changes over time. This change in time is a critical measure of participants use of the system coupled with the false explanations.

The first question, Figure 10, places emphasis on the most recent recommendation the participant saw. With this question we tried to capture a reaction from the participant about how each recipe recommendation and its explanation modified their trust in the context of a single recommendation. The second question, Figure 11, was designed to capture the overall sense of trust that the participant had in the recommendation system as the experiment progressed.

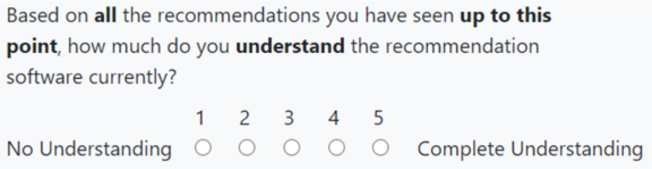
The next set of questions dealt with understanding. These questions are demonstrated in Figure 12 and Figure 13.



How did the explanation of the recommendation affect your **understanding** of the recommendation software?

	1	2	3	4	5	
significantly reduced	○	○	○	○	○	significantly increased

Fig. 12. Question 3 as it appeared to participants. How did the explanation of the recommendation affect your understanding of the recommendation software?



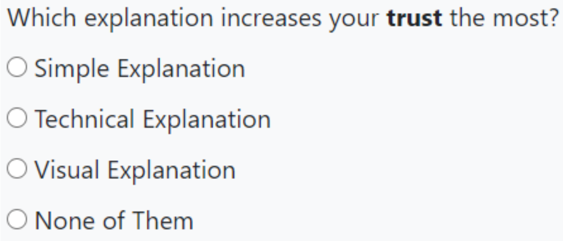
Based on **all** the recommendations you have seen **up to this point**, how much do you **understand** the recommendation software currently?

	1	2	3	4	5	
No Understanding	○	○	○	○	○	Complete Understanding

Fig. 13. Question 4 as it appeared to participants. Based on all the recommendations you have seen up to this point, how much do you understand the recommendation software currently?

These questions have the same goals as the first two questions but instead are meant to track how participants' understanding changes over time.

The last two questions, which were multiple choice, revisit trust and understanding by asking the participant about which explanation method had the most impact on their understanding and trust. These questions are shown in Figure 14 and Figure 15.



Which explanation increases your **trust** the most?

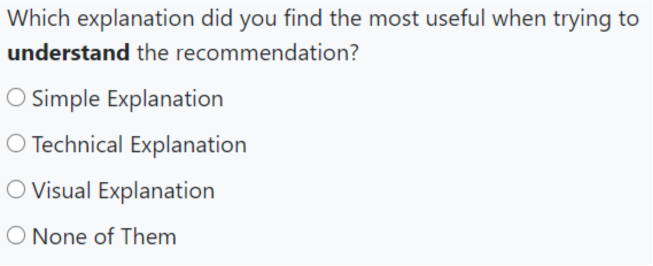
○ Simple Explanation

○ Technical Explanation

○ Visual Explanation

○ None of Them

Fig. 14. Question 5 as it appeared to participants. Which explanation increases your trust the most?



Which explanation did you find the most useful when trying to **understand** the recommendation?

○ Simple Explanation

○ Technical Explanation

○ Visual Explanation

○ None of Them

Fig. 15. Question 6 as it appeared to participants. Which explanation did you find the most useful when trying to understand the recommendation?

The purpose of these two questions is to get a sense of how participants were responding to the different types of explanations.

In addition to the multiple-choice questions, we collected quantitative data about the behavior of the participants as they interacted with the recipe recommendations. We recorded how many times the users clicked on each of the buttons on each page.

At the end of the experiment each participant was asked to provide qualitative data by giving feedback about the experiment by answering four open-ended questions about the recommendation engine in general. The questions asked about their final levels of trust, understanding, inconsistencies that they found, and asked if they understood how the recommendation engine worked internally.

The experiment had sixty participants after erroneous submissions were removed. As previously mentioned, the goal was to get an even distribution of male and female and well as between technical (STEM) and non-technical people.

Table 1 summarizes the population of the participants.

Table 1. Participants' demographics.

Gender	Total	Technical (STEM)	Non-Technical
Male	30	15	15
Female	30	15	15

4 Results

The results of this experiment show that trust and understanding are clearly affected by the presence of explanations. To show this we show the analysis of each set of questions presented to the participants.

For all analysis where a P value is used we chose to consider all results with a P value less than 0.05 as significant. This value was selected previous to data analysis.

4.1 Trust

The first survey question states, "How did the recommendation you just saw affect your trust in the recommendation software?" The participants were asked to answer this question using a Likert scale of 1-5 where 1 represented a significant reduction in trust and 5 indicated a significant increase. The first thing to note is the general trend of how the participants answered this question. This trend is given below in Figure 16.

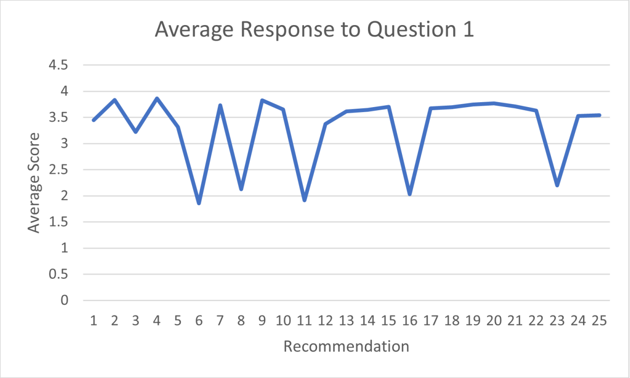


Fig. 16. Average response to question 1. The graph shows the response to survey question 1 for each recommendation averaged over all the participants.

The main feature of this graph is that is shows that false recommendations did affect trust. This can be seen by the dips at recommendations 6, 8, 11, 16, and 23 which used false explanations. Each time a false recommendation was encountered, on average participants reported that they had less trust in the recommendation software. This leads to the following question: Does the presence of false explanations damage trust over time? Figure 17, which summarizes question 2 measuring trust over time gives some insight into this.

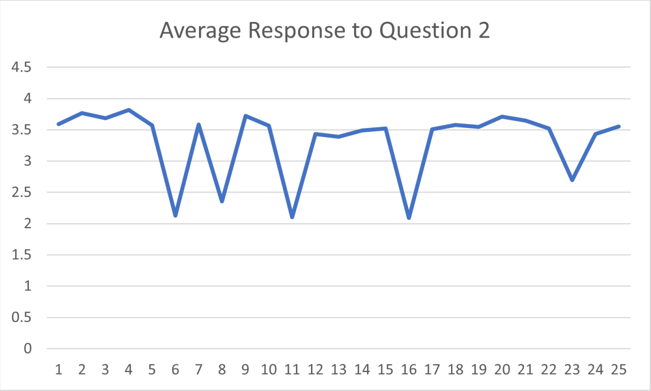


Fig. 17. Average response to question 2. The graph shows the response to survey question 2 for each recommendation averaged over all the participants.

Survey question two states, “Based on all the recommendations you have seen up to this point, how much do you trust the recommendation software?” We can see from Figure 17 that the trends in question two follow nearly identically the trend in question one (Figure 16). This leads to the conclusion that exposure to faulty recommendations does not impact trust negatively over the long term.

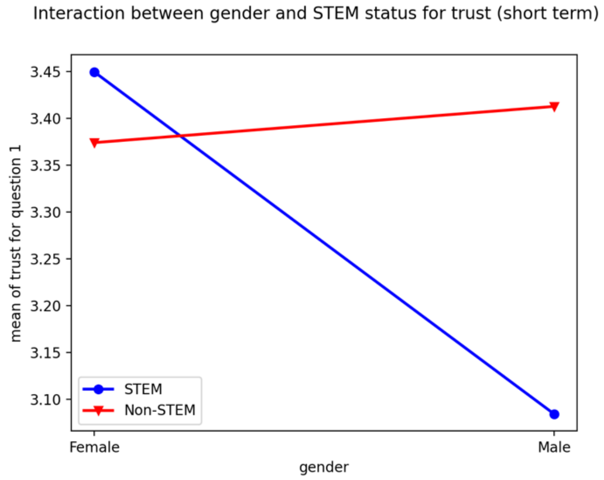


Fig. 18. Statistically significant interaction between gender and STEM status of participants for question one.

Performing a two-way ANOVA on gender and STEM status for question one results in a statistically significant interaction with an F score of F(11.1979, 1) = 0.0144. The interaction is visualized in Figure 18. The interaction in the graph shows insight into the difference in the genders relative to trust. The graph shows that, on average, men have a larger variance in their trust than women do, but both genders are affected by their educational and professional background. Specifically, men’s trust factor is negatively affected and women’s trust is positively affected by their STEM background, but non-STEM participants had approximately the same amount of trust regardless of gender.

Performing a two-way ANOVA on gender and STEM status for question 2 also results in a statistically significant interaction with an F score of F(11.1979, 1) = 0.0144. The interaction is visualized in Figure 19.

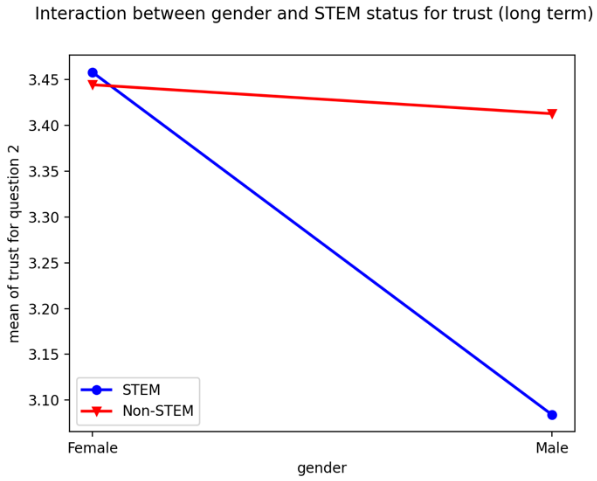


Fig. 19. Statistically significant interaction between gender and STEM status of participants for question two.

It is important to also note that there is an interaction between participants' STEM status and their trust as affected by false explanations. Performing a two-way ANOVA on STEM status and trust found results in an interaction with an F-score of F(9.1811, 1) = 0.0025. The interaction is shown in Figure 20. Specifically, participants' trust not in the STEM field had a greater variance based on bad or false explanations.

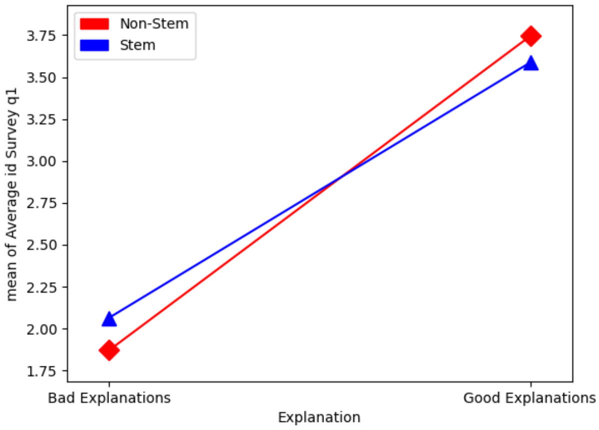


Fig. 20. Interaction between STEM status and reaction to false explanations for question 1.

4.2 Understanding

Questions three and four relate to understanding. These trends are shown in Figure 20 and Figure 21.

Question three states, “How did the explanation of the recommendation affect your understanding of the software?” The main result from this trend is that after participants see a false recommendation that their average understanding decreases. The highest average understanding occurs during the first five recommendations. Interestingly, after each recommendation that has a false explanation the understanding of future recommendations is impacted. This is especially true after recommendation 11. It appears that the false explanations create lingering confusion about how correct explanations were being generated.

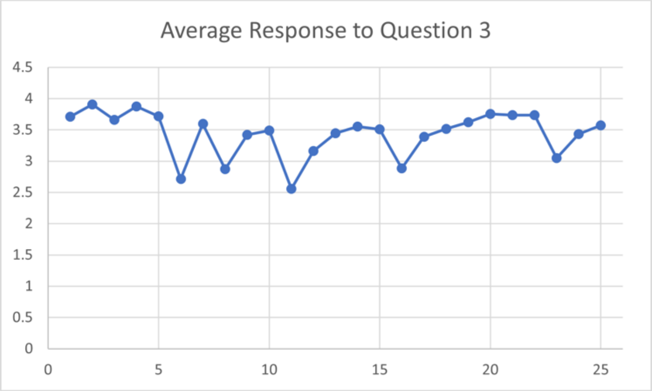


Fig. 21. Average response to question 3. The graph shows the response to survey question 3 for each recommendation averaged over all the participants.

The trend for question four (Figure 22) is similar to that of question three shown in Figure 21. Question four was “Based on all the recommendations you have seen up to this point, how much do you understand the recommendation software?”

Performing a two-way ANOVA for STEM status and gender for question three and four resulted in a statistically significant interactions with a F score of F(5.3545, 1) = 0.0003 for question three and an F score of F(4.1216, 1) = 0.0425 for question four. The interactions are visualized in Figure 23 and Figure 24 respectively.

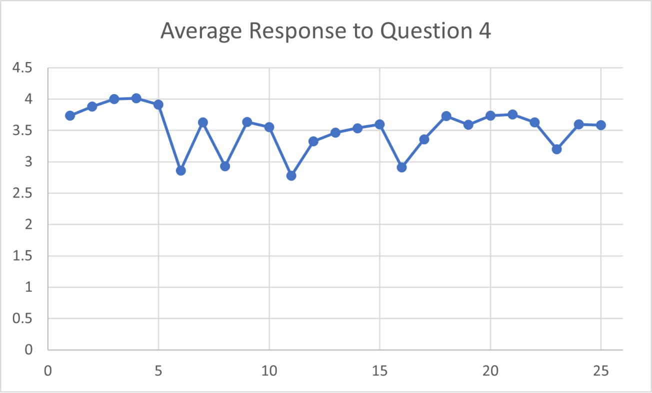


Fig. 22. Average response to question 4. The graph shows the response to survey question 4 for each recommendation averaged over all the participants.

These results are consistent with the two-way ANOVA results from questions one and two. This leads to the conclusion that for both trust and understanding an important factor was the interaction of STEM status and gender. In other words, men and women react differently in regard to trust and understanding to explanations based on their educational and professional background.

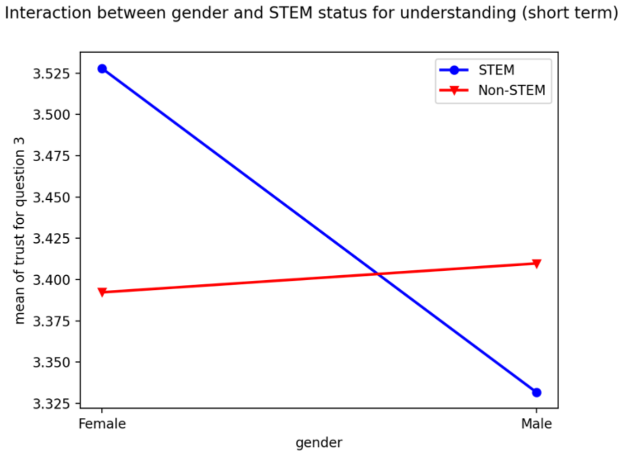


Fig. 23. Statistically significant interaction between gender and STEM status of participants for question 3.

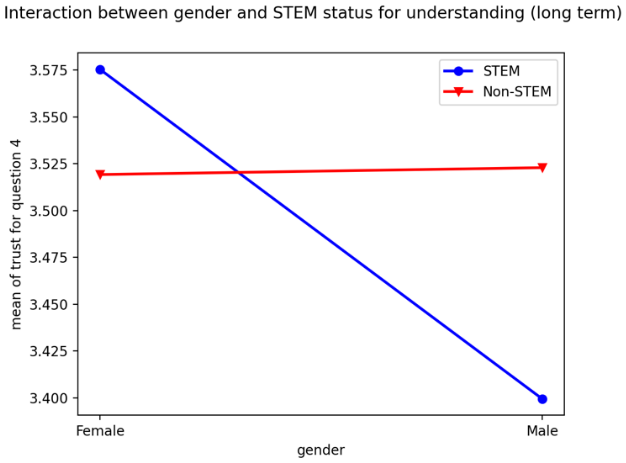


Fig. 24. Statistically significant interaction between gender and STEM status of participants for question 4.

4.3 Explanation Types

We now analyze the explanation types. There were three types of explanations: simple text, technical text, and visual word clouds.

Question five asks, “Which Explanation increases your trust the most?” Participants are asked to select which of the three types of explanations helped the most or if none of them were helpful.

Table 2 shows the results of a statistically significant Chi-squared analysis for question 5.

Table 2. Question 5 (explanation trust) chi-squared analysis results.

Gender	**Female**		**Male**	
Stem	No	Yes	No	Yes
Simple	194	203	180	191
Technical	196	104	116	149
Visual	9	14	7	17
None	67	62	74	71
P Value	0.00034			

Table 2 shows a summary of how each participant answered question 5, broken down by gender and STEM status.

The visual word cloud explanation was viewed by participants as far less effective at increasing trust when compared to the simple and technical explanations. Despite this, it is interesting to note that both men and women who were considered STEM preferred the visual explanations when compared to their not STEM counterparts.

For the technical text explanations, we see that non-technical women preferred the technical explanation over STEM women. In men it was the opposite, STEM men preferred the technical explanations more than non-STEM men.

Simple text explanations are preferred roughly the same between men and women as well as their STEM and non-STEM groups. These findings continue the theme that the main factors affecting participant trust and understanding is the interaction of gender and STEM status that is present in the findings from the first 4 questions.

These patterns also appear, with some variation in the results of the Chi-squared analysis of question six. Survey question six states, "which explanation did you find the most useful when trying to understand the recommendation?" Table 3 shows the results of the analysis.

Again, we see that the visual word cloud explanations were the least preferred type when participants were trying to understand the recommendation software. We again see that STEM women preferred the visual explanations vs non-STEM women. However, for the men, we see that non-STEM men prefer the visual explanations more than STEM men did. In terms of the technical explanations, the results show that STEM men preferred technical explanations far more than non-STEM men. This preference is reversed for the STEM/non-STEM women. Interestingly, non-STEM women found the simple explanation much more useful than STEM women did, meanwhile, the men were roughly even as they were in question five.

Table 3. Question 6 (explanation understanding) Chi-squared analysis results.

Gender	**Female**		**Male**	
Stem	No	Yes	No	Yes
Simple	181	106	140	148
Technical	213	189	115	208
Visual	15	36	62	18
None	57	52	60	54
P Value	5.20e-15			

4.4 Qualitative Results

The concepts of trust and understanding can be hard to quantify. Because of this, the comments made by participants provide useful insight into the results in the previous section.

The final section of the survey asked open-ended questions. The first question asked participants about which explanation did the best job of increasing their trust and why. Of the STEM females who replied, the consensus was that the simple textual explanations were best. A few examples responses included conciseness, clarity, and simplicity as to their reasons.

There were a few responses where STEM females note that it was a combination of the simple and technical explanations but only stated that the technical explanations were best at increasing their trust.

Non-STEM females who responded were evenly split between the simple and technical. In general, there was concern that the simple textual explanation was too simple or condescending, while the technical textual explanation was sometimes more detailed than needed.

This feedback is consistent with the outcome of the Chi-squared analysis of question 5. It also gives valuable context to what participants may have been thinking as they went through the experiment.

For STEM males the responses were more split. However, the participants who responded seemed to favor the simple textual explanations as well.

Non-STEM males were also somewhat divided between technical and simple explanations, but they leaned more towards the simple explanations.

The second open-ended question asked participants to indicate which explanation method was most helpful in understanding the recommendations and asked the participants to give their best guess as to how the software was generating recommendations.

STEM females who responded preferred the technical explanations generally. Almost all the participants who responded were able to give a high-level description of how hybrid recommendation systems work.

Non-STEM females also tended to prefer the technical explanations.

STEM males had a response like the females in that they generally also preferred technical explanations. This group was also able to provide well educated guesses for how the software might be working.

Non-STEM males who responded, were more divided between technical and simple explanations being the best for increasing understanding. Most of the non-STEM males who responded were able to make a good guess, but some of them were less specific and in one case, a participant admits that they did not know.

4.5 Summary of Results

We have shown that the trust and understanding users have in recommendation engines is influenced by the presence of explanations. Specifically, we found that there is an interaction between a participant’s education and professional background (STEM status) and their gender.

We also found that the presence of explanations was impactful to a user’s trust of the system overall. The consistent appearance of the STEM status and gender interaction across the first four questions provides an explanation of what factors are important in influencing users trust and explanation in a recommendation system.

We also found that different explanation methods have different impacts on participants based on their STEM status and gender. In general, we found that simple explanations were best at increasing trust while technical explanations were best at aiding in understanding the recommendation system. It was also shown that visual explanations, as they were expressed in the experiment, were by far the least useful for increasing trust and understanding.

5 Conclusion

In this paper we report on an experiment to understand how trust and understanding in a recommendation engine are affected by explainable AI (XAI). We specifically tested 60 people divided into different categories, namely STEM status (educational and professional STEM background) and gender, which resulted in 15 people per category. Based on related works, we also tested two text-based explanations and one visual explanation, specifically a word cloud. We also introduced false explanations, explanations that were obviously wrong, for the purpose of more deeply seeing if explanations mattered to participants.

We found the following summarized results:

Gender and STEM status had an interactive effect on both trust and understanding.

Different explanations were preferred by different groups; however, the visual word cloud was overwhelmingly rejected.

The results section overwhelmingly show that the gender and STEM status of the participant were important factors in determining trust and understanding. In other words, there is not one particular way to show explanations that is best for all audiences because both gender and the person's education and professional background play a part in determining their trust and understanding in the system.

In addition, although a number of publications have promoted visual word clouds as a way to show XAI, our participants rejected that type of explanation and preferred textual explanations.

6 Acknowledgments

This work was supported in part by a grant from AFRL contract #FA 8650-20-F-1956.

References

[1] Turing, A. "Computing Machinery and Intelligence," *Mind*, vol. 59, pp. 433-460, 1950.

[2] Silver, D., et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, no. 6419, pp. 1140-1144, 2007.

[3] S. Grigorescu, et al. "A survey of deep learning techniques for autonomous driving." *Journal of Field Robotics*, 2019.

[4] Klein, G. *Sources of power: How people make decisions*. MIT press, 2017.

[5] Reyna, V. "How people make decisions that involve risk: A dual-processes approach." *Current directions in psychological science*, vol. 13, no .2, pp. 60-66, 2004.

[6] Glöckner, A. and Betsch, T. "Do people make decisions under risk based on ignorance? An empirical test of the priority heuristic against cumulative prospect theory." *Organizational Behavior and Human Decision Processes*, vol. 107, no. 1, pp. 75-95, 2008.

[7] Seckler, M., et al. "Trust and distrust on the web: User experiences and website characteristics." *Computers in human behavior*, vol. 45, pp. 39-50, 2015.

[8] Cheng, Y., and Zifei, F. "Encountering misinformation online: antecedents of trust and distrust and their impact on the intensity of Facebook use." *Online Information Review*, 2020.

[9] Dattner, B., et al. "The legal and ethical implications of using AI in hiring." *Harvard Business Review*, vol. 25, 2019.

[10] Gunning, D., and Aha, D. "DARPA's explainable artificial intelligence (XAI) program." *AI Magazine*, vol. 40, no. 2, pp. 44-58, 2019.

[11] Cao, W., et al. "A review on neural networks with random weights," *Neurocomputing*, vol. 275, pp. 278-287, 2018.

[12] Guidotti, R., et al. "A Survey of Methods for Explaining Black Box Models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1-42, 2019.

[13] Craven, M. and Shavlik, J. "Extracting tree-structured representations of trained networks," In *Proceedings of the Conference on Advances in Neural Information Processing System*, pp. 24–30, 1996.

[14] Boz, O. "Extracting decision trees from trained neural networks." In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 456–461, 2002.

[15] Augasta, M and Kathirvalavakumar, T. "Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems," *Neural Process Letters*, vol. 35, no. 2, pp. 131–150, 2012.

[16] Barakat, Nahla, and Andrew P. Bradley. "Rule extraction from support vector machines: a review." *Neurocomputing*, vol. 74, no .1, pp. 178-190, 2010.

[17] D. Martens, et al. "Comprehensible credit scoring models using rule extraction from support vector machines," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1466-1476, 2007.

[18] Deng, Houtao. "Interpreting tree ensembles with intrees." *International Journal of Data Science and Analytics* vol. 7, no .4, pp. 277-287, 2019.

[19] N. Barakat and A. P. Bradley. "Rule extraction from support vector machines: A review," *Neurocomputing*, vol. 74, pp. 178-190, 2010.

[20] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention," In *Proceedings of the International Conference on Machine Learning*, pp. 2048–2057, 2015.

[21] Zhang, Y. and Xu, C. "Explainable recommendation: A survey and new perspectives." arXiv preprint arXiv:1804.11192, 2018.

[22] Zhang, Y., et al. "Do users rate or review? Boost phrase-level sentiment labeling with review-level sentiment classification." Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014.

[23] Guan, X., et al. "Attentive Aspect Modeling for Review-Aware Recommendation." *ACM Transactions on Information Systems (TOIS)*, vol. 37, no. 3, pp. 1-27, 2019.

[24] Aggarwal, C. *Recommender Systems: The Textbook*. Springer, 2016.

[25] Zhang, Y., et al. “Explicit factor models for explainable recommendation based on phrase-level sentiment analysis,” In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014.

[26] Chang. S., et al. “Crowd-based personalized natural language explanations for recommendations,” In *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

[27] Wu, Y. and Ester, M. “FLAME: A probabilistic model combining aspect based opinion mining and collaborative filtering,” In *Proceedings of the eighth ACM international conference on web search and data mining*, 2015.

[28] Lin, Y., et al. “Explainable Outfit Recommendation with Joint Outfit Matching and Comment Generation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1502-1516, 2020.

[29] He, X., et al. "Trirank: Review-aware explainable recommendation by modeling aspects." In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.

[30] Wang, X, et al. “TEM: Tree-enhanced Embedding Model for Explainable Recommendation.” In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, 2018.

[31] Wang, N., et al. “Explainable Recommendation via Multi-Task Learning in Opinionated Text Data,” In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*, pp. 165–174, 2018.

[32] Seo, S., et al. “Interpretable convolutional neural networks with dual local and global attention for review rating prediction.” *Proceedings of the eleventh ACM conference on recommender systems*, 2017.

[33] Koren, Y., et al. "Matrix factorization techniques for recommender systems." *Computer*, vol. 42, no.8, pp. 30-37, 2009.

[34] Leilani, G., et al. "Explaining explanations: An overview of interpretability of machine learning." *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE*, 2018.