KU LEUVEN

FACULTY OF SCIENCE

Making Machine Learning Fairer: How LLMs Can Empower Users

Yori Verbist

Supervisor: Prof. dr. K. Verbert
Mentor: *Ir. A. Bhattacharya*
Assessor: *Dr. Ir. L. Allein*

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Toegepaste Informatica,
option Artificial Intelligence

Academic year 2023-2024

Preface

This thesis is the culmination of my journey through the complex and fascinating world of artificial intelligence, specifically in the realm of bais and transparency in machine learning models. What began as a curiosity about the ethical implications of AI has evolved into a deep exploration of how technology can be designed to be more equitable and accessible to users.

The work presented here is the result of countless hours of research, experimentation, and reflection. It has been both challenging and rewarding, pushing me to expand my knowledge and think critically about the role of AI in society. Throughout this journey, I have had the privilege of working with and learning from many talented and dedicated individuals, whose support has been invaluable.

I would like to express my deepest gratitude to my supervisor, Professor Verbert, whose guidance, encouragement, and insightful feedback have been instrumental in shaping this research. Your expertise and patience have been a constant source of motivation, and I am incredibly grateful for your mentorship. Also, a special thanks to Aditya Bhattacharya for answering my countless questions and guiding my towards the right direction.

A special thanks goes to the participants who took part in the user study. Your willingness to engage with the FairML platform and provide candid feedback has been essential to the success of this research. Without your contributions, this work would not have been possible.

Finally, I would like to thank my family and friends for their unwavering support and understanding throughout this journey. Your belief in me has been a constant source of strength, and I am deeply grateful for your encouragement.

Yori Verbist

Contents

Abstract

As artificial intelligence (AI) becomes increasingly integrated into critical sectors such as healthcare, finance, and criminal justice, the need to ensure the fairness and transparency of these systems is growing. Bias in AI models, caused by biased training data or flawed algorithms, can lead to unfair and discriminatory outcomes, undermining public trust. This thesis explores the potential of large language models (LLMs) and conversational interfaces to support users, particularly non-experts, in identifying and reducing bias in machine learning (ML) models.

The research focuses on the FairML platform, a tool that combines both manual and conversational model steering. The platform is designed to enhance the interpretability of AI systems and enable users to interact with ML models through natural language queries. The study evaluates how conversational explanations impact the understandability of bias, how LLM chatbots facilitate model steering, and how these interfaces can provide causal reasoning for bias.

The findings show that conversational explanations significantly improve the understanding of complex bias-related concepts, making AI models more accessible to non-experts. While the LLM chatbot was effective in model steering and bias reduction, users noted there is room for improvement in the consistency and depth of responses. This research highlights the potential of LLM-based interfaces to promote fairness and transparency in AI systems.

The findings have important implications for the design of AI tools and emphasize the importance of user-centered design in creating fair and reliable AI technologies. Suggestions for future research include testing the platform with a larger and more diverse group of users and further developing the conversational capabilities of AI systems.

Samenvatting

Naarmate kunstmatige intelligentie (AI) steeds meer wordt toegepast in kritieke sectoren zoals de gezondheidszorg, financiën en het strafrecht, groeit de noodzaak om de eerlijkheid en transparantie van deze systemen te waarborgen. Vooringenomenheid in AI-modellen, veroorzaakt door bevooroordeelde trainingsdata of gebrekkige algoritmen, kan leiden tot oneerlijke en discriminerende resultaten, wat het publieke vertrouwen ondermijnt. Deze thesis onderzoekt het potentieel van grote taalmodellen (LLM's) en conversatie-interfaces om gebruikers, met name niet-experts, te ondersteunen bij het identificeren en verminderen van vooringenomenheid in machine learning (ML)-modellen.

Het onderzoek richt zich op het FairML-platform, een tool die zowel handmatige als conversatie-gebaseerde modelsturing combineert. Het platform is ontworpen om de interpreteerbaarheid van AI-systemen te verbeteren en gebruikers in staat te stellen via natuurlijke taalvragen met ML-modellen te communiceren. De studie beoordeelt hoe conversatieverklaringen de begrijpelijkheid van vooringenomenheid beïnvloeden, hoe LLM-chatbots modelsturing vergemakkelijken en hoe deze interfaces causale redeneerprocessen kunnen bieden.

De resultaten tonen aan dat conversatieverklaringen de begrijpelijkheid van complexe bias-gerelateerde concepten aanzienlijk verbeteren, waardoor AI-modellen toegankelijker worden voor niet-experts. Hoewel de LLM-chatbot effectief bleek in modelsturing en biasreductie, gaven gebruikers aan dat er ruimte is voor verbetering in consistentie en diepgang van de antwoorden. Dit onderzoek onderstreept het potentieel van LLM-gebaseerde interfaces om eerlijkheid en transparantie in AI-systemen te bevorderen.

De bevindingen hebben belangrijke implicaties voor het ontwerp van AI-hulpmiddelen en benadrukken het belang van gebruikersgericht ontwerp bij het creëren van eerlijke en betrouwbare AI-technologieën. Suggesties voor toekomstig onderzoek omvatten het testen van het platform met een grotere en meer diverse groep gebruikers en het verder ontwikkelen van de conversatiemogelijkheden van AI-systemen.

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
XAI	Explainable AI
UI	User Interface
UX	User Expierence
UCD	User-Centered Design

Chapter 1

Introduction

Artificial Intelligence (AI) has become an integral part of various sectors, ranging from healthcare and finance to transportation and criminal justice. The predictive power and efficiency of AI models have revolutionized decision-making processes, offering unprecedented accuracy and insights. However, as AI systems become more pervasive, concerns about their fairness and transparency have also intensified. Bias in AI models, stemming from biased training data or flawed algorithms, can lead to unfair and discriminatory outcomes, undermining the trust and credibility of these systems.

Ensuring fairness in AI is not merely a technical challenge but a societal imperative. The deployment of AI systems that are transparent, explainable, and free from bias is crucial for maintaining public trust and ensuring ethical outcomes. This calls for the development of methods and tools that help identify, understand, and mitigate bias in AI models, making them more accessible and trustworthy for a diverse range of users, including those without deep technical expertise.

Explainable AI (XAI) has emerged as a promising approach to address these concerns. By making AI models more interpretable, XAI techniques enable users to understand the decision-making processes of these models. This transparency is essential for identifying potential biases and for making informed adjustments to improve model fairness. Among the various XAI approaches, conversational explanations, facilitated by large language models (LLMs), offer an intuitive and user-friendly way to interact with AI systems, potentially democratizing access to AI tools and empowering users to take a more active role in bias mitigation.

This thesis explores the potential of LLM-based conversational explanations to empower users, in identifying and mitigating bias in machine learning (ML) models. The research focuses on the development and evaluation of the FairML platform, a tool that integrates both manual and conversational model steering components. The platform is designed to enhance the interpretability of AI systems and support users in making less biased decisions by allowing them to interact with and adjust ML models through natural language queries.

The primary objectives of this research are encapsulated in the following research questions:

How do conversational explanations impact the understandability of bias issues in ML models?

How do LLM chatbots facilitate users in model steering to reduce bias issues in ML models?

How do conversational explanations provide causal reasoning for model bias?

To address these questions, the research employs a mixed-methods approach, combining quantitative and qualitative data to provide a comprehensive evaluation of the FairML platform. The platform's development, guided by principles of user-centered design, involved iterative testing and feedback to ensure that it meets the needs of its intended users.

The structure of this thesis is as follows:

- Chapter 2: Related Work - Provides an overview of existing research on explainable AI, bias in AI, model steering approaches, and causal explainability.
- Chapter 3: Methodology - Describes the research design, including the mixed-methods approach, participant recruitment, and the structure of the FairML platform.
- Chapter 4: Development - Details the iterative development process of the FairML platform, including usability testing and the integration of user feedback.
- Chapter 5: Results - Presents the findings from the study, analyzing the effectiveness of conversational explanations and LLM chatbots in enhancing model interpretability and reducing bias.
- Chapter 6: Discussion - Discusses the implications of the findings, the limitations of the study, and potential areas for future research.
- Chapter 7: Conclusion - Summarizes the key contributions of the research and its significance for the field of AI fairness.

Through this research, the thesis aims to contribute to the development of more transparent, fair, and trustworthy AI systems. By demonstrating the potential of conversational interfaces in bias mitigation, this work seeks to make machine learning fairer and more accessible for everyone, thus advancing the ethical deployment of AI technologies across various domains.

Chapter 2

Related Work

This chapter gives the needed information to ground and position this thesis. It gives a brief overview of the relevant research related to explainable AI, bias in AI and causal explainability.

2.1 Explainable AI

Explainable AI (XAI) is a field of research that focuses on making AI models more interpretable for humans. Since the invention of deep neural networks, these models have become highly effective in many domains, but they are also very complex and hard to interpret. Especially in high-risk fields as healthcare, finance and criminal justice, it is important to explain why a model made a certain prediction. Since, wrong predictions in these fields can have severe consequences and there exists this big question about who is responsible when something goes wrong; the AI expert who made the model, or the domain expert who used and trusted the model? [12]

Even for AI experts themselves, it is often impossible to understand why a model makes a certain prediction. That is why XAI methods are being developed, to make these models more interpretable for users. When these models are more interpretable, it becomes easier to tell why a certain model made a particular prediction.

Methods for explaining models are classified into two distinct categories: model-specific and model-agnostic, depending on their level of specificity. Explanations that are tailored to particular model structures and algorithms fall under the category of model-specific explanations. Examples include Saliency Maps [18], and Grad-CAMs [16], which are designed for specific types of models. In contrast, model-agnostic explanations are versatile tools that can elucidate the workings of any model, regardless of the underlying algorithms. Widely recognized methods in the realm of XAI, such as LIME [14], SHAP [10], are representative of model-agnostic approaches.

Explainability is essential in machine learning because it provides insights into the sources of bias and the impact of bias mitigation techniques. By making the decision-making process of AI systems transparent, explainability tools enable stakeholders to detect and understand biases in model predictions. This understanding is crucial for developing effective strategies to mitigate bias and ensure that AI systems operate fairly and

ethically. For instance, SHAP can highlight which features contribute most to decisions, guiding developers in adjusting the model or data to reduce bias.

Methods are further classified into two categories: local and global explanations. Local explanations are concerned with providing insights into individual predictions by focusing on a specific instance in the data. Conversely, global explanations offer an overarching understanding of the model trained across the entire dataset. Studies have demonstrated that global explanations are more effective in instilling confidence about the model's workings than local explanations [6].

2.2 Bias in AI

Bias is not only pivotal but also multifaceted. It can originate from various sources, including but not limited to the data used to train AI systems, the design of the algorithms themselves, and the social context in which AI is applied.

There exist a lot of different kinds of biases [11]. Ranging from algorithmic biases, which introduce bias through the design of the algorithm used in the specific model. While data biases are introduced in the datasets that are used to train the models. The latter are the kind of biases this thesis is going to focus on.

2.2.1 Representation Bias

This thesis focuses on representation bias [17] specifically. Representation bias stems from the methods employed in sampling populations during the data-gathering phase. Samples that do not accurately reflect the population's diversity are deficient, omitting certain subgroups and exhibiting various irregularities.

There are two ways one can measure representation bias: (1) Representation rate and (2) Data coverage [17]. The representation rate for a subgroup is the ratio of its sample size to the largest sample size among all subgroups. For instance, if a study variable D encompasses subgroups s, t, and u, summing up to a total sample count S (where $s+t+u=S$), then the representation rate for subgroup s is determined by $r_s = \frac{s}{max(s,t,u)}$. Let's give an example with a medical dataset categorizing patients by disease type, such as diabetes, hypertension, and heart disease. If there are 900 patients in total, with 200 diabetic patients, 500 hypertensive patients, and 200 heart disease patients, then the representation rate for diabetic patients is $r_s = \frac{200}{max(200,500,200)} = 0.4$. Conversely, the representation rates for hypertensive and heart disease patients are 1.0 and 0.4, respectively, indicating a lower representation of diabetic and heart disease patients compared to those with hypertension.

Moreover, data coverage is defined as the essential minimum number of samples each subgroup should possess. For example, if the data coverage requirement is 250 patients, then both the diabetes and heart disease subgroups do not meet the data coverage criteria, as they each have only 200 patients. It is crucial, regardless of the breadth of the dataset, to ensure ample coverage for all pivotal subgroups to affirm their adequate representation.

This thesis is going to use representation rate because getting good data coverage can be difficult in a lot of domains. Since there exist a lot of variables in a lot of domains that are more frequent than other variables. Take, for example, a dataset with different types of cancers. There will always be some kind of cancer that is more frequent than other kinds. This makes using data coverage instead of representation rate impractical.

2.3 Model Steering Approaches

In the field of interactive machine learning (IML), model steering refers to the process of incorporating user feedback directly into the model refinement process to enhance predictive algorithms. Researchers have predominantly explored various methods for embedding user input into training, optimizing, and troubleshooting predictive models [2, 20]. These methods are crucial for improving model accuracy and reliability by leveraging domain-specific knowledge [8].

While recent work has introduced innovative methods that enable domain specialists to guide predictive models [2, 15], significant gaps remain in utilizing conversational agents for this purpose. Current approaches often lack the ability to integrate user feedback in a seamless, real-time, and user-friendly manner. For example, many existing systems require users to have the substantial technical expertise to influence model behavior effectively, which can be a barrier for non-experts.

Our research aims to bridge this gap by developing methods that allow AI experts to engage in mitigating biases in predictive models using conversational agents. Specifically, we will explore the integration of Large Language Models (LLMs) with model steering processes. By enabling real-time, interactive feedback and adjustments, this approach seeks to make model steering more intuitive and accessible, thus enhancing the overall usability and effectiveness of predictive models.

2.4 Causal Explainability

Causal explanations are a type of explanations that are based on causal relationships between features. It gives an intuitive explanation of why the model made a certain prediction, by showing which features have the most influence on the prediction [13, 23].

2.4.1 Conversational Explainability

There has also been done some work on causal explainability with large language models (LLMs) [24]. The main difference here is that an LLM can not show any visual representaions of causal relationships. LLMs can only show textual causal explanations through conversations.

As explained in [24] we can ask an LLM three kinds of causal questions. First, we have to identify causal relationships using domain knowledge. For example, a patient

asks: Will my bad eating habits cause a higher chance of getting diabetes? These are the types of questions that most modern LLMs can answer already. The second type of question is discovering new knowledge from data. Here, we give the LLM access to new data so that it can use this extra data to reason with. At last, we have a quantitative estimate of the consequence of actions. An example is, "Medical doctor: This is the third time that this patient has returned with lumbago. The epidural steroid injections helped him before, but not for long. I injected 12mn betamethasone the last two times. What is the dose that I should use this time?" (Zhang, 2024, p.2).

This work focuses on the second type where the LLM gets more knowledge from a particular dataset. This kind of questions also help uncover unknown causal relationships between features. Since it can be hard for humans to find relationships in big datasets.

Chapter 3

Methodology

The last chapter defined the main goals of this research. Based on these goals, we define the following research questions.

Research Question 1

How do conversational explanations impact the understandability of bias issues in ML models?

Research Question 2

How do LLM chatbots facilitate users in model steering to reduce bias issues in ML models?

Research Question 3

How do conversational explanations provide causal reasoning of model bias?

This chapter outlines the necessary arrangements and the logic applied (where relevant) to tackle the specified research inquiries. The conducted research has been approved by *Sociaal- Maatschappelijke Ethische Commissie* (SMEC) (File: G-2023-7463-R2(MIN)).

3.1 Approach

In this study, our methodological approach is structured around addressing three primary research questions aimed at investigating the impact of conversational explanations and Large Language Model (LLM) chatbots on the interpretability and mitigation of bias issues in machine learning (ML) models.

For RQ1, the research question is: "How do conversational explanations impact the understandability of bias issues in ML models?" The hypothesis posits that users can more clearly interpret the model's bias through conversational explanations. This hypothesis guides the exploration into the effectiveness of conversational explanations in enhancing users' understanding of bias issues in ML models.

RQ2 focuses on understanding the effectiveness of LLM chatbots in facilitating users in model steering to reduce bias issues. The hypothesis states that "LLM chatbot-based

model steering is adequate for reducing bias issues." The study aims to evaluate the efficacy of LLM chatbots in assisting users with model steering to mitigate bias issues inherent in ML models.

Lastly, RQ3 investigates how conversational explanations can illuminate the causal reasoning of model bias. It is hypothesized that conversational explanations can effectively clarify the causal factors leading to model bias. This research question concentrates on the capacity of conversational explanations to improve users' comprehension of the origins and implications of bias in models.

To address these research questions, a mixed-methods approach is employed, utilizing both quantitative and qualitative data. This approach was chosen to capture both the measurable outcomes of user interactions with the FairML platform and the nuanced experiences that quantitative data alone cannot fully explain. By combining these methods, the study aims to provide a comprehensive evaluation of the platform's effectiveness in enhancing model interpretability and bias mitigation.

Quantitative data is collected to measure the effectiveness of conversational explanations and LLM chatbots in enhancing users' understanding and mitigating bias issues. This involves using structured questionnaires, such as the System Usability Scale (SUS) and NASA Task Load Index (NASA-TLX), as well as performance metrics to gather numerical data that can be statistically analyzed. The use of these instruments allows for objective measurement and comparison of different explanation modalities and model steering methods.

The study employs a within-subjects design, where each participant engages with both the manual model steering and the conversational model steering components of the FairML platform. This design allows for direct comparison of the two approaches while controlling for individual differences among participants. The quantitative approach provides a robust foundation for assessing the effectiveness of the different components of the FairML platform.

Qualitative data is gathered to provide deeper insights into users' perceptions and experiences with the explanation and steering approaches. This involves conducting semi-structured interviews and analyzing open-ended responses from questionnaires. The qualitative approach is essential for understanding the nuanced aspects of user interactions and experiences that cannot be captured through quantitative measures alone.

The interviews are designed to explore participants' thoughts on the clarity, usability, and effectiveness of the conversational explanations and the LLM chatbot. This qualitative data helps contextualize the quantitative findings and offers rich insights into how users interact with the FairML platform, providing a more comprehensive understanding of its strengths and areas for improvement.

This methodological approach is integral to achieving the thesis's broader objective of evaluating the effectiveness of conversational interfaces in mitigating bias in AI. By combining quantitative and qualitative data, the study provides a holistic assessment that informs the development of more transparent and accessible AI tools. The findings from this study will contribute to the broader field of explainable AI and bias mitigation,

offering insights that could be applied to improve the fairness and usability of AI systems.

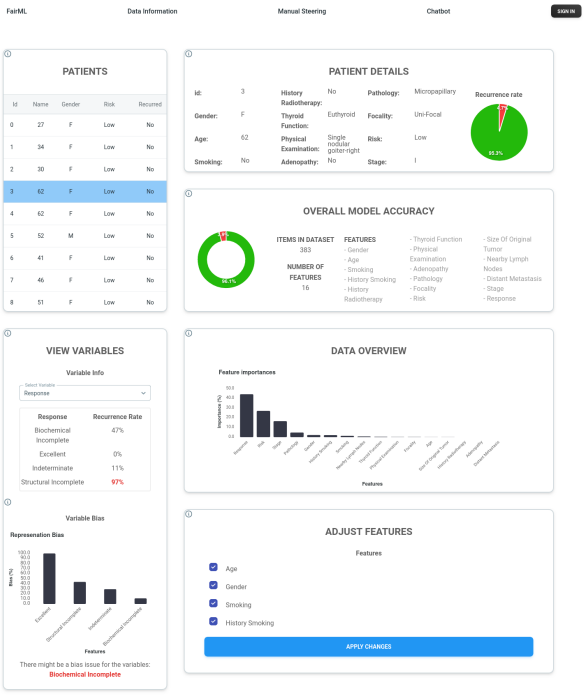
3.2 FairML

The FairML platform was developed as a comprehensive tool to facilitate the research and address the challenges of bias identification and mitigation in machine learning models. This platform plays a crucial role in the study, providing a practical environment where users can interact with models and directly engage in bias management. The platform consists of two primary components, each designed to cater to different aspects of model steering: a manual steering component and a conversational model steering component, as illustrated in Figures 3.1 and 3.2.

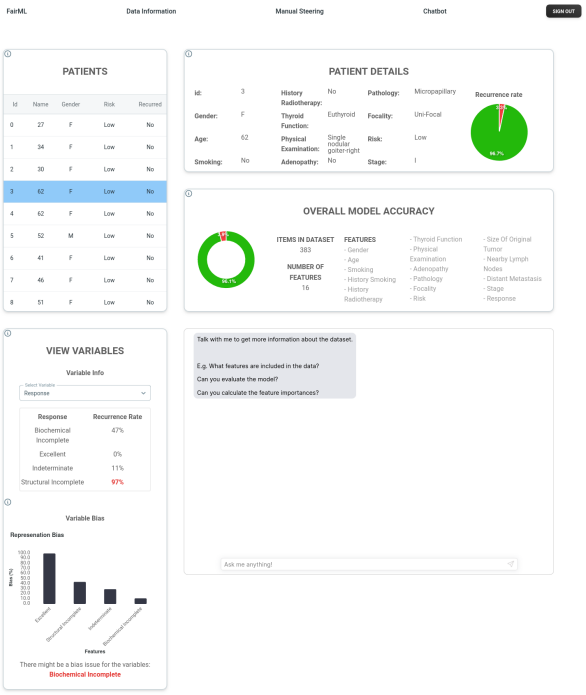
The manual steering component provides users with a hands-on approach to interacting with the machine learning model. Through a graphical user interface, users can manually select and adjust features, observing how these changes affect model predictions. This component is particularly valuable for users who prefer direct control over the model's behavior, allowing them to address potential biases by selecting and deselecting features in a tangible and precise manner. By enabling users to interact directly with the model, the manual steering component fosters a deeper understanding of the relationship between features and model outcomes, which is essential for effective bias mitigation.

In contrast, the conversational model steering component leverages the power of large language models (LLMs) to offer a more intuitive and accessible interaction method. This component employs a chatbot interface that allows users to engage in natural language conversations with the system. Through these interactions, users can ask questions about the model's behavior, seek explanations, and receive guidance on how to adjust the model to reduce bias. The conversational approach is designed to lower the barrier to entry, making it easier for users who may not have deep technical expertise to still effectively steer the model and mitigate biases. By integrating LLMs, the conversational component offers a user-friendly alternative that complements the precision of manual steering, ensuring that a wide range of users can engage with the platform effectively.

The integration of both manual and conversational steering components within FairML creates a versatile and robust platform for bias management. The manual component offers precision and control, ideal for users who prefer a more detailed and hands-on approach. Meanwhile, the conversational component democratizes access to bias mitigation tools by simplifying complex tasks through natural language interactions. Together, these components work synergistically to enhance the overall effectiveness of the platform, providing users with the necessary tools to understand and address biases in a comprehensive manner. This integration ensures that FairML is not only flexible but also highly effective in catering to different user needs and expertise levels, thereby supporting the broader goal of making machine learning models fairer and more transparent.



(a) Manual Model Steering



(b) Conversational Model Steering

Figure 3.1: The developed FairML platform to conduct the research with.

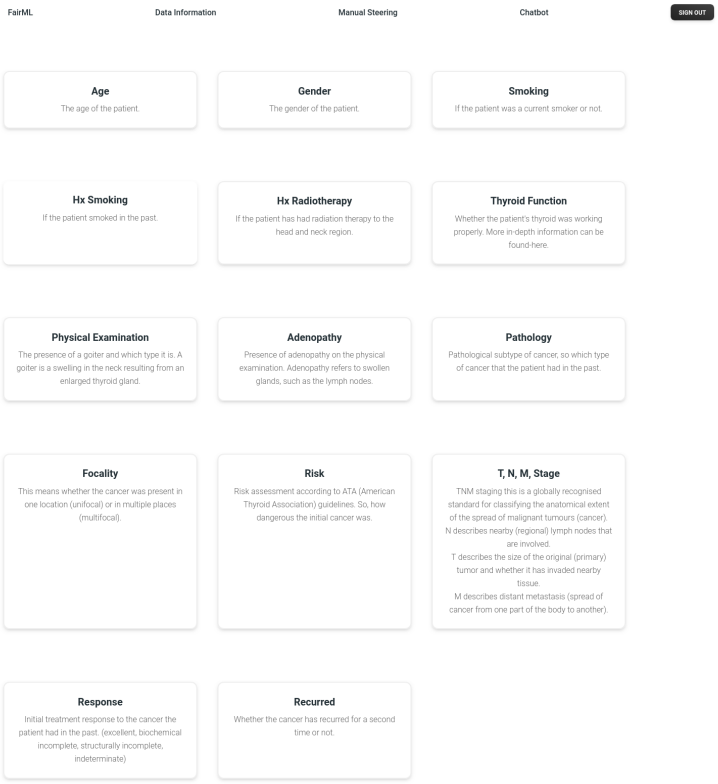


Figure 3.2: Information page about the different features in the dataset.

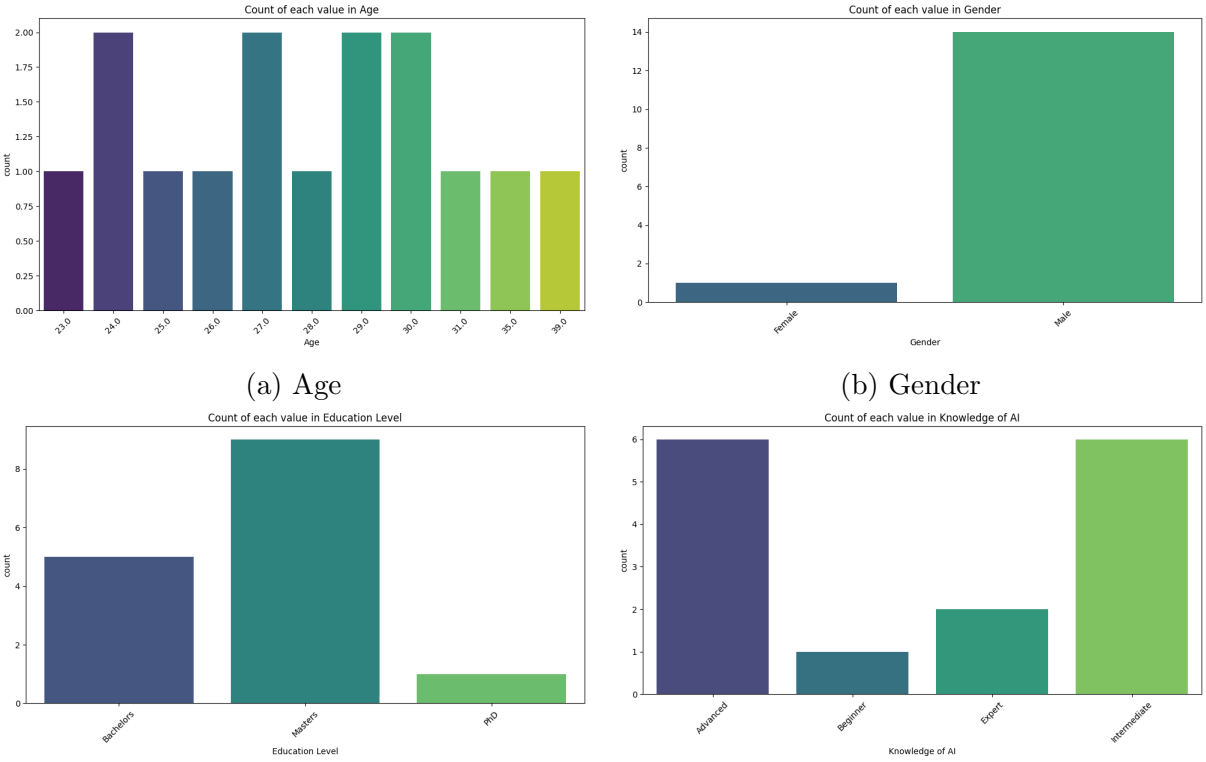
3.3 Participants

In this section, we discuss the recruitment process and the demographics of the participants involved in the study. The target audience for this research was AI experts, given their familiarity with machine learning concepts and their ability to provide informed feedback on the system's capabilities and limitations.

3.3.1 Recruitment Process

Participants were recruited through social media platforms, where AI experts were invited to participate in the study. The eligibility criteria included being over 18 years of age and having experience with AI. There were no additional restrictions to ensure a diverse pool of participants within the target demographic.

Those who expressed interest in participating were provided with an informational brochure detailing the research objectives, and methodology, as well as additional information regarding the handling and storage of data. These participants were required to sign an informed consent form. All the documents relating to the recruitment process (information brochure and informed consent form) can be found in appendix A.1.



(a) Age

(b) Gender

(c) Education Level

(d) Experience with AI

Figure 3.3: Demographics of the final participants

3.3.2 Final Participants

A total of 15 AI experts were successfully recruited for the study. The demographic information of the final participants is shown in Figure 3.3. All participants were male except one and worked in Belgium, with ages ranging from 23 to 39 years. The participants' educational backgrounds varied, with different levels of expertise in AI, as illustrated in the demographic breakdown.

Although the aim was to recruit a larger number of participants, several potential participants did not respond when it was time to participate in the study. Consequently, the study was conducted with fewer participants than initially planned. Given the limited number of participants, a within-subjects user study design was employed. This design choice ensured that each participant experienced both conditions (manual and conversational model steering). The within-subjects design was chosen to maximize the data obtained from each participant and to allow direct comparison between the two steering methods, reducing the variability associated with individual differences.

3.4 Flow of Study

The final user study was designed to rigorously evaluate the FairML platform by involving participants in a structured sequence of interactions and assessments. To ensure the reliability and comparability of the results, all participants followed a consistent procedure throughout the study.

Pre-Study Questionnaires

Initially, participants were required to complete a pre-study questionnaire aimed at gathering general demographic information and details about their experience with AI. This step was crucial for understanding the background of the participants and segmenting the data analysis based on varying levels of AI expertise. The questionnaire, detailed in Appendix B, included questions on age, gender, educational level, and AI experience. This data was essential for contextualizing the participants' interactions with the platform and assessing whether previous experience influenced their ability to identify and mitigate bias.

Interaction Phases: Manual and Conversational Steering

Following the pre-study questionnaire, participants interacted with two distinct components of the FairML platform: the manual model steering component and the conversational model steering component. The study design intentionally had each participant first engage with the manual model steering interface. During this phase, participants used sliders and other manual controls to address potential bias issues within the dataset. This phase was critical for establishing a baseline of participant capability, allowing for a direct comparison of how users performed bias mitigation tasks manually.

After completing the manual steering tasks, participants proceeded to the chatbot model steering interface. In this phase, they interacted with an AI-driven chatbot designed to assist in identifying and mitigating bias in the ML models. The structured order of these interactions was deliberately chosen to minimize learning effects that might arise if participants were allowed to switch freely between the two methods. By structuring the interaction sequence in this way, the study aimed to isolate the impact of the conversational interface on participants' understanding and effectiveness in bias mitigation, thus directly addressing Research Questions 1 and 2.

The tasks in each phase were designed to be comparable, focusing on similar bias issues within the same dataset. For instance, participants were asked to adjust the influence of features such as gender and age on the model's predictions in both the manual and chatbot interfaces. This consistency ensured that the differences observed between the two methods could be attributed to the interface used rather than the complexity or nature of the tasks.

Post-Study Questionnaires

Upon completing their interaction with both components, participants were asked to fill out a series of post-study questionnaires. These questionnaires were designed to assess various aspects of their experience, including perceived understandability, usability, and workload associated with the system. The questionnaires included standardized tools such as the System Usability Scale (SUS) and the NASA Task Load Index (NASA-TLX), alongside custom questions tailored to the specific features of FairML. Detailed information about these questionnaires and their structure can be found in Section 3.5 and Appendix B.

The flow of the study, as described, is illustrated in Figure 3.4. This consistent flow

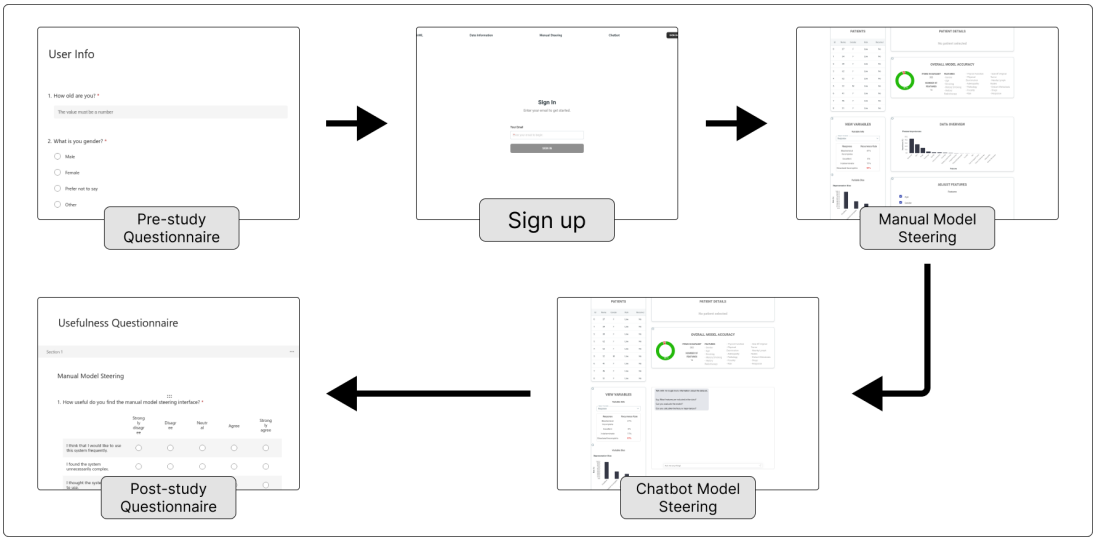


Figure 3.4: Flow the participants could follow when using FairML

is crucial for ensuring that the data collected is comparable across all participants, thus enhancing the validity of the study's findings.

3.5 Questionnaires

This section provides a detailed overview of the questionnaires used in the study, highlighting their design, purpose, and how they contributed to addressing the research objectives.

3.5.1 Pre-study Questionnaire

Before starting the study, participants were required to complete a pre-study questionnaire designed to gather general demographic information and assess their prior experience with AI. This questionnaire included questions about age, gender, educational level, and self-reported AI knowledge (Appendix B, table: B.1). The data collected from this questionnaire was critical for contextualizing the participants' interactions with the FairML platform. For example, understanding the participants' level of AI expertise allowed for a more nuanced analysis of how different backgrounds might influence their ability to identify and mitigate bias. These insights were particularly relevant when interpreting the differences in participants' performance and feedback during the study.

3.5.2 Post-study Questionnaire

At the end of the study, participants completed a series of post-study questionnaires designed to evaluate their experiences with the FairML platform. These questionnaires aimed to capture detailed feedback on the perceived understandability, usability, and cognitive workload associated with the system. The following standardized tools and custom questions were used:

Perceived Understandability: The perceived understandability questionnaire assessed how well participants understood the system's behavior and its assistance in decision-

making. Questions focused on the clarity, predictability, and ease of following the system's actions. This was crucial for determining whether users felt confident in their ability to use the system effectively without requiring extensive prior knowledge. High ratings in this area would suggest that the conversational explanations were successful in making complex concepts more accessible, which directly addresses Research Question 1.

Usability Assessment: The System Usability Scale (SUS) [4] was used to evaluate the practical aspects and ease of use of the FairML platform. Participants rated their experiences on various aspects, including system complexity, ease of use, and overall user confidence. The SUS scores provided a reliable measure of the system's usability, offering insights into how intuitive and user-friendly the platform was, which is vital for determining its practicality for non-expert users. These results were particularly relevant to Research Question 2, which focused on the effectiveness of the LLM chatbot in assisting with model steering.

Workload Evaluation: The NASA Task Load Index (NASA-TLX) [5] was employed to assess the cognitive and physical demands placed on participants during the tasks. This tool measured mental and perceptual activity, physical effort, time pressure, and emotional response. The NASA-TLX scores were essential for identifying any areas where the system might have been overly demanding, providing a comprehensive measure of workload that could inform future system improvements.

Open-ended Feedback: In addition to the structured questions, open-ended questions were included to capture more detailed qualitative feedback. These questions allowed participants to provide specific insights into their preferences, potential areas for improvement, and comparative evaluations between the manual and chatbot interaction methods. The qualitative data gathered from these responses was invaluable for understanding the nuanced experiences of users and for guiding the future development of the system. This data also helped address Research Question 3, which explored how well the conversational explanations provided causal reasoning for model bias.

Rationale for Question Selection: The post-study questionnaire was carefully designed to provide a comprehensive evaluation of the FairML platform from multiple perspectives. By combining standardized tools like the SUS and NASA-TLX with open-ended responses, the questionnaire aimed to capture both quantitative and qualitative data. This approach ensured that the evaluation was thorough and nuanced, capable of informing iterative improvements to the platform based on user feedback.

3.6 Analysis of Results

In this section, the analysis methods used to interpret the results of the user study are outlined. Given the nature of the data collected, a combination of quantitative and qualitative analyses was employed to provide a comprehensive understanding of the effectiveness and usability of the FairML platform.

To analyze the quantitative data, we first considered the type of data collected and

the assumptions underlying different statistical tests. Since the majority of our data is derived from Likert scale questionnaires, which produce ordinal data, non-parametric statistical tests were deemed most appropriate. While some argue that parametric tests can be applied to Likert scale data [7], the non-parametric approach is generally preferred due to fewer assumptions about the data distribution.

For comparing the effectiveness of the conversational model steering, the Mann-Whitney U Test was utilized. This test is particularly suitable for within-subject designs where participants experience both conditions, as it compares the differences in their responses without assuming a normal distribution of those differences. The test was applied to the data from the post-study questionnaires to assess whether there was a statistically significant difference in the perceived usefulness and workload between the two model steering approaches.

The analysis was conducted using Python, with the pandas library [19] for data manipulation, SciPy [21] for statistical testing, and seaborn [22] and Matplotlib [9] for data visualization. These tools provided the necessary functionality to perform the analyses and generate visual representations of the results.

By employing the Mann-Whitney U test, we ensured that our analysis was both statistically rigorous and appropriately tailored to the nature of the data. This approach provides a solid foundation for drawing meaningful conclusions about the effectiveness of the conversational model steering component in the FairML platform.

Chapter 4

Development

In this chapter we outline the iterative development process and the development of the FairML application.

4.1 Iterative Development Process

The development of FairML was guided by the principles of User-Centered Design (UCD), ensuring that the application was molded around the specific requirements of its users, primarily AI professionals. Given the wide-ranging characteristics of this demographic, maintaining a connection with AI professionals throughout the development phase was essential to address potential divergence in values and existing knowledge.

To bridge this gap, usability testing was employed, providing developers with the opportunity to identify and rectify errors promptly while integrating valuable user feedback. The development process followed an iterative approach, progressing through two main stages: the initial high-fidelity model and the final proof of concept. Each stage is depicted in Figure 4.1 and involves continuous refinement based on user feedback.

Usability testing was a critical component at both stages, with think-aloud studies chosen for their simplicity and efficacy in gathering user experience insights. In the think-aloud approach, participants engaged with the prototype while verbalizing their thought process as they completed a set of tasks. The observing developer closely monitored the participants' rationale and interactions. The insights gained from these observations served as critical feedback for refining the prototype.

The first stage involved the development of a high-fidelity prototype using the Figma platform. This stage included evaluation sessions with two UX experts to ensure the prototype adhered to established usability standards. Participants' feedback from these

Figure 4.1: Development stages of FairML

sessions highlighted issues such as the complexity of the causal graph and the clarity of variable bias representation.

Based on this feedback, significant changes were made in the second stage. For example, the causal graph was replaced with a bar chart to better illustrate feature importance, as the bar chart provided a clearer and faster way for users to understand which features were most important. Additionally, the application explicitly indicated the use of representation bias and included tooltips for users unfamiliar with this concept.

These iterative improvements were validated through further think-aloud studies with AI and UI experts, ensuring that the final proof of concept was both user-friendly and effective in addressing bias issues in machine learning models.

By systematically incorporating user feedback into each development stage, the iterative development process ensured that FairML met the needs of its users, ultimately enhancing its usability and functionality.

4.2 Prototype

The primary goal of developing the high-fidelity prototype was to ensure that the interface was both user-friendly and aligned with the research objectives of facilitating model steering for AI experts. The prototype was developed using Figma[1], a collaborative design platform, allowing for detailed visual representation and interactive mockups of the FairML platform's interface.

The prototype was designed with a focus on creating a clear and intuitive user experience. Initial designs featured a manual model steering interface where users could adjust model parameters via sliders, as illustrated in Figure 4.2. This design aimed to simplify the process of model interaction, enabling users to easily manipulate features and observe the resulting changes in model outputs.

To ensure the prototype adhered to usability standards, evaluation sessions were conducted with UX experts. These sessions focused on several key usability criteria, including navigation efficiency, clarity of information presentation, and user engagement. The UX experts provided valuable feedback, highlighting areas that needed refinement, such as the complexity of the causal graph and the overall layout of the interface.

Based on this feedback, significant changes were made to the prototype. For instance, participants in the think-aloud study, two AI experts and one UI expert, found the causal graph too complex to interpret quickly. As a result, it was replaced with a bar chart, which provided a more straightforward visualization of feature importance and allowed users to identify key features more easily. Significantly improved the users' ability to interact with the model and understand its outputs.

[1] https://www.figma.com/

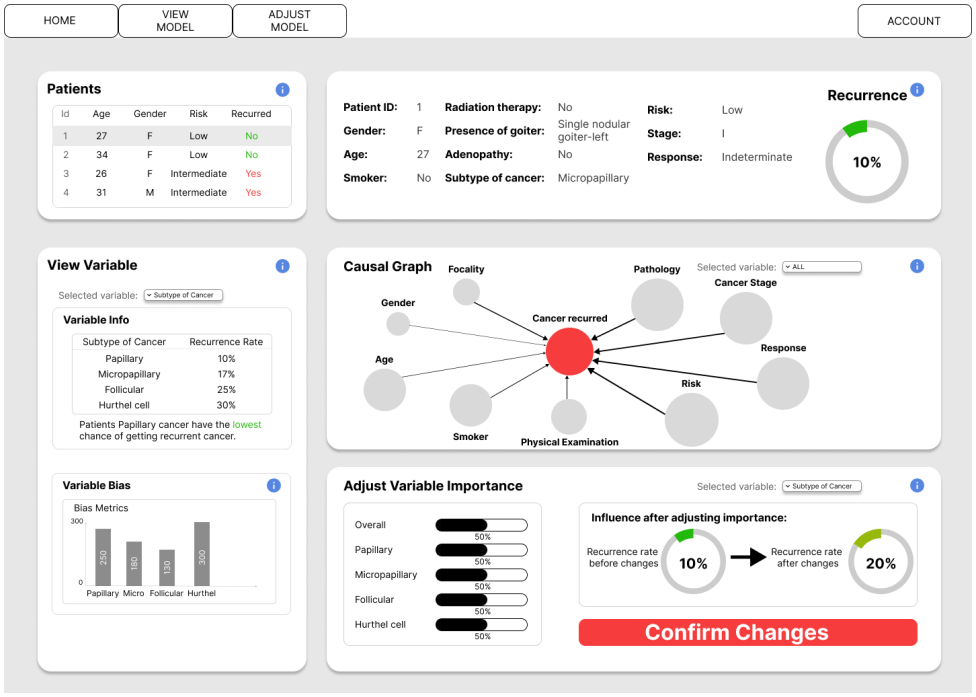


Figure 4.2: Prototype of the screen with manual model steering through sliders.

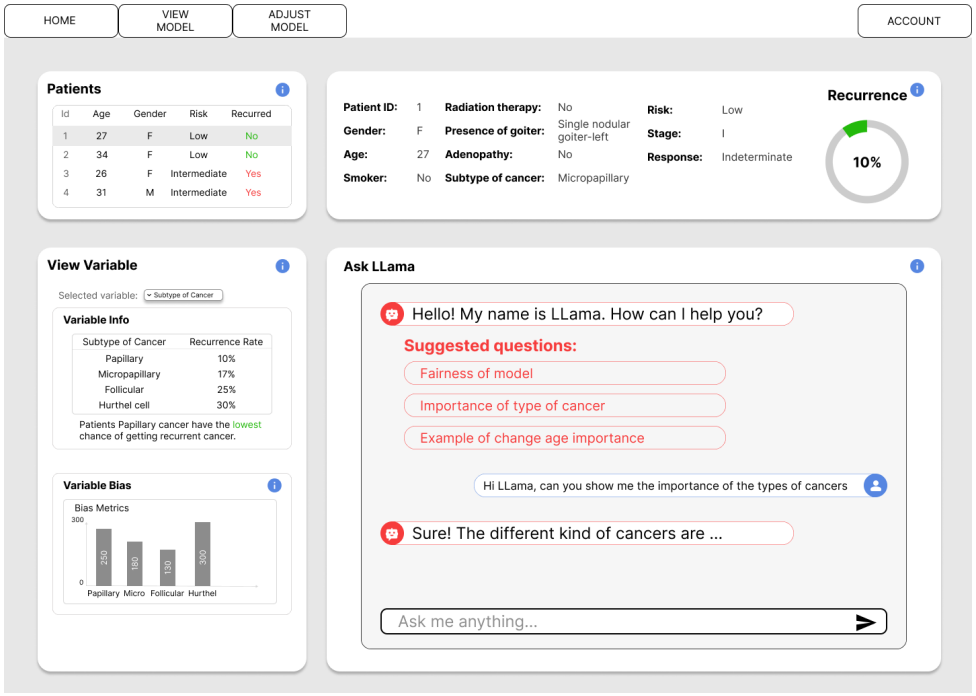


Figure 4.3: Prototype of the screen with conversational model steering through a chatbot.

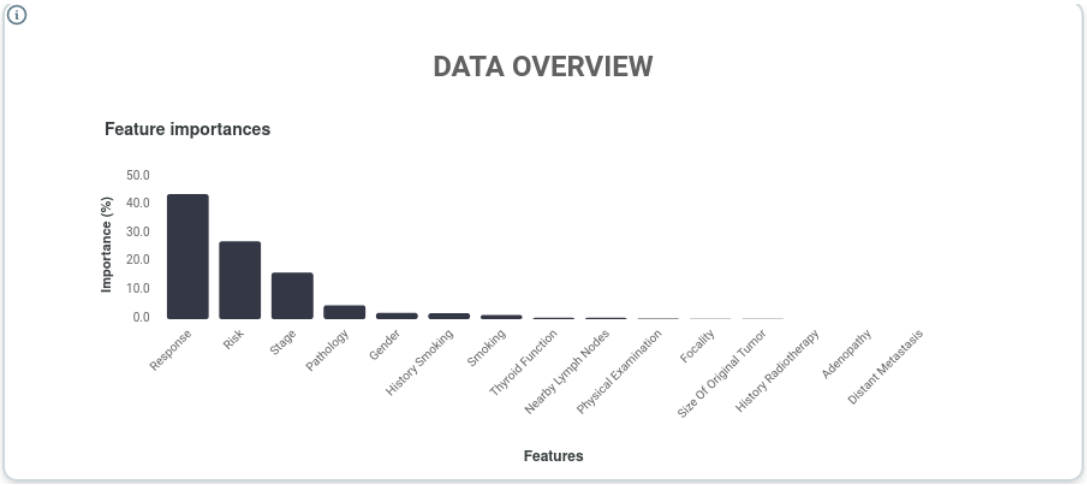


Figure 4.4: Bar chart of the feature importances

In addition to replacing the causal graph, other modifications were made to enhance the prototype based on both the think-aloud study and UX expert evaluations. Task-centric inquiries during the evaluation sessions revealed that users struggled with certain navigation paths, leading to revisions in the interface layout to streamline these processes. Feedback-centric inquiries, which gathered participants' overall impressions, indicated a need for more detailed information about each feature, prompting the inclusion of tool tips and additional explanatory text.

The iterative design process, informed by continuous feedback, ensured that the final prototype was well-aligned with the needs of its target users. Figures 4.2 and 4.3 show the progression from the initial design to the final interface, illustrating how user input directly influenced the evolution of the FairML platform.

The most significant changes from this think-aloud study are discussed in section 4.3.

4.3 Significant Changes

Upon examining the figures from the initial high-fidelity prototype and the ultimately deployed website, several notable modifications and exclusions are evident. The purpose of this section is to elucidate the rationale behind these alterations.

Feature Importances. In the prototype, a causal graph was used to show the importance of each feature for the predictions. However, feedback from the think-aloud studies revealed that participants, especially those unfamiliar with causal graphs, found it challenging to interpret. As a result, the causal graph was replaced with a bar chart in the final version of the website. The bar chart provides the same information in a more accessible format, allowing users to quickly identify the most important features by sorting the chart from most to least important (see Figure 4.4). This change was crucial for enhancing the usability of the platform by simplifying the visualization of feature importance.

Variable Bias. During the think-aloud studies, participants expressed uncertainty

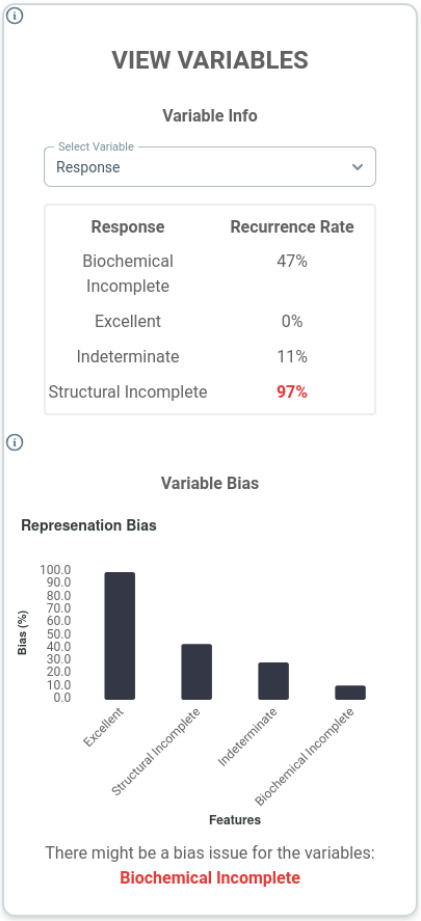


Figure 4.5: View Variable Component

about the type of bias being addressed in the View Variable view. To address this, we made it clear in the final application that representation bias was being used. Additionally, a tool tip was added to explain what representation bias is and how users should interpret it. This change was driven by the need to ensure that all users, regardless of their familiarity with different types of biases, could fully understand the system's functionality. At the bottom, we also indicated which variable values of the selected feature might have a bias issue. The selection criteria of these possible biased variables are when their representation bias [17] had a value smaller than 25%. These changes are shown in figure 4.5.

Information Page. The initial prototype lacked an information page to provide additional context about the features and variables in the dataset. This gap was highlighted during user testing, particularly for participants without a medical background, who struggled to understand the dataset. In response, we added an information page in the final version, where each feature is explained in detail. Furthermore, we incorporated hover-over tooltips across the application to offer quick, contextual information about each feature, as shown in Figure 4.6. This addition made the platform more user-friendly and accessible to a broader audience, ensuring that all users can interact with the data effectively.

These modifications significantly improved the usability of the FairML platform, align-

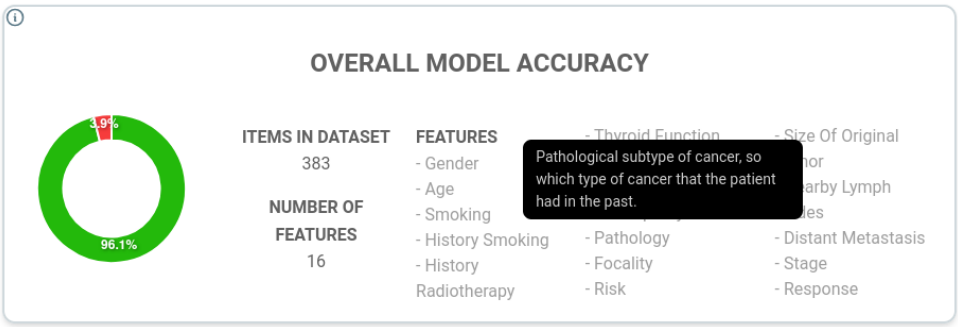


Figure 4.6: More details when a user hovers over a feature

ing it more closely with the needs and expectations of its users. By responding directly to user feedback, the platform was made more intuitive, accessible, and effective in supporting users in their efforts to identify and mitigate bias in machine learning models.

4.4 Model Steering Interface

The model steering interface is vital in this thesis. In this seciton, we will explain the thinking behind the concluding model steering interface. There are two versions of model steering in this thesis, manual model steering and conversational model steering.

4.4.1 Manual Model Steering

Figure 4.7 illustrates the manual model steering component. For the manual model steering part the user uses checkboxes to select the features they want for the model to use. Originally, all the features are selected. All of this can be seen in section B in the previously mentioned figure. The user also sees a bar chart with the feature importance for every feature that is selected (section A in the figure). Through this bar chart, the user can decide if a particular feature needs to be excluded from the model. When the user deselects a feature, it will also be excluded from the feature importance graph.

4.4.2 Conversational Model Steering

The conversational model steering component can be seen in figure 4.8. As explained in section 2.3 there still exist gaps for model steering through conversational agents. This is partly to the limitation of the existing conversational agents that can not interact with specific models/datasets. It is possible to fine-tune an existing model like ChatGPT on a custom dataset, but this would be really cumbersome when you are frequently changing datasets.

To make it possible to communicate about a specific dataset and do model steering on a specific model, this thesis uses LangChain[2]. LangChain makes it possible to add custom functions to an LLM which it can then can call to interact with a specific dataset and model. A more in-depth description of how LangChain is used is outlined in section 4.6.

[2] https://python.langchain.com/v0.2/docs/introduction/

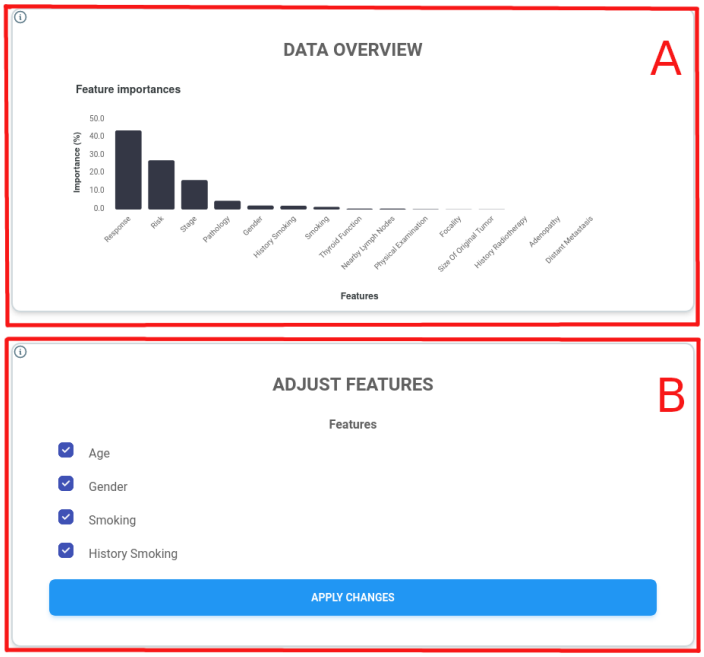


Figure 4.7: Marked version of the manual model steering interface

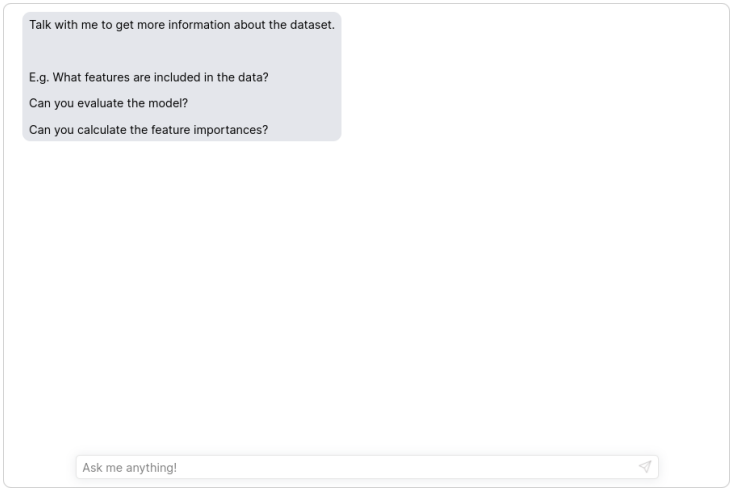


Figure 4.8: Conversational model steering interface

4.5 Takeaways

We provide a brief summary of the key insights gained from the development process. These observations stem from the think-aloud studies conducted as part of this research and may not be universally applicable. However, they offer valuable considerations for designing a similar application for a comparable target audience.

Varied Familiarity with Large Language Models (LLMs): Participants demonstrated varying levels of familiarity with LLMs, which impacted their ability to steer the model using conversational methods. The performance of the LLM was notably influenced by how participants formulated their questions, underscoring the importance of clear and effective communication with the model.

Future designs should consider providing users with guidelines or examples on how to interact with LLMs effectively. This could help bridge the gap between users with different levels of experience and ensure more consistent outcomes across diverse user groups.

Differences in Identifying Bias Issues: Despite all participants being AI professionals, there was a noticeable variation in their ability to quickly and accurately identify bias issues within the AI model. This discrepancy highlights the differing levels of expertise and experience among users, even within a specialized field.

The system could be enhanced by incorporating adaptive features that provide additional support or guidance for users who may struggle with identifying bias. For example, implementing a tiered assistance system that offers varying levels of detail based on the user's interaction history or performance could be beneficial.

These takeaways, while specific to the context of this study, offer practical insights for the development of similar systems. By addressing the challenges and leveraging the strengths identified during the think-aloud studies, developers can create more user-friendly and effective AI model steering interfaces.

4.6 Technical Implementation

This section provides an overview of the technical implementation of FairML, focusing on the architectural decisions and technologies used to build the platform. The FairML platform utilizes a combination of modern web development frameworks and machine learning tools to create an interactive and user-friendly environment for model steering.

4.6.1 Backend Architecture

The backend architecture of FairML is designed to be modular and scalable, with a focus on flexibility and performance. The platform's backend is built using FastAPI[3], a modern web framework known for its high performance and ease of use. FastAPI was chosen for its ability to handle asynchronous operations efficiently, which is critical for the real-time interactions required in FairML.

[3] https://fastapi.tiangolo.com/

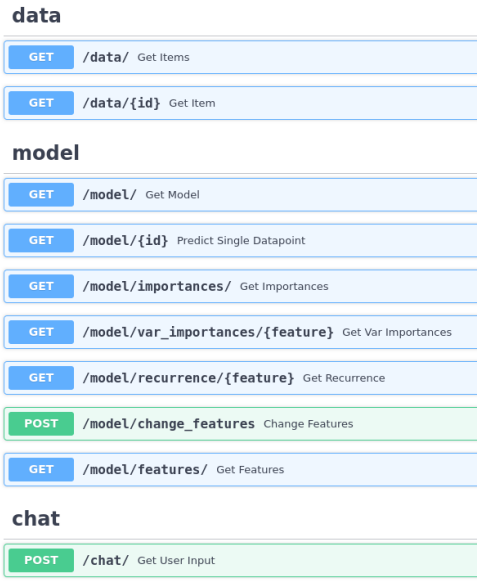


Figure 4.9: Possible endpoints of the API

The backend is divided into three main components each with specific responsibilities, a visualization of its endpoints can be seen in figure 4.9:

Data Endpoint: The Data endpoint is responsible for managing and retrieving the dataset used in the platform. Users can query the entire dataset or retrieve specific entries as needed. This component ensures that data handling is efficient and that users can access the necessary information quickly.

Model Endpoint: This endpoint handles all interactions with the machine learning model. The model used is a Support Vector Machine (SVM), chosen based on its superior performance on the dataset, as previously demonstrated in relevant studies [3]. The Model endpoint allows users to perform tasks such as retrieving overall model accuracy, predicting outcomes for specific data points, and calculating feature importances using SHAP (SHapley Additive exPlanations) [10].

Chatbot Endpoint: The Chatbot endpoint facilitates communication between the user and the LLM-powered chatbot. LangChain, a framework that extends the capabilities of LLMs, is used to integrate custom functions into the chatbot. This allows the chatbot to perform tasks such as feature importance calculations and dataset queries dynamically.

This works as follows:

1. An agent is created and initialized with a set of tools that it can use. These tools could be anything from search engines, APIs, databases, or custom functions that provide specific capabilities.
2. The agent uses the language model as a reasoning engine to decide which tool to use and in what sequence. This decision-making process is dynamic and based on the context of the task at hand. If no tool is needed, it answers the question as it would otherwise.

3. Once the agent decides on an action, it executes the tool associated with that action. The tool performs its function and returns the result to the agent.
4. The agent processes the results of the action. It may use the outcome to make further decisions, execute additional tools, or conclude the task.
5. The results of these actions are fed back into the agent, which then determines if more actions are needed or if the task is complete.

For the chatbot, OpenAI's GPT-3.5-turbo was selected as the underlying LLM due to its cost-effectiveness and compatibility with LangChain. This model supports a wide range of functionalities necessary for the conversational aspects of the platform.

The frontend of FairML is built using React[4] and tailwindcss[5]. React was chosen for its component-based architecture, which allows for the creation of a dynamic and responsive user interface. Tailwindcss provides utility-first CSS, enabling rapid and consistent styling across the application.

[4] https://react.dev/
[5] https://tailwindcss.com/

Chapter 5

Results

This chapter presents the findings from the user study conducted to evaluate the effectiveness of the FairML platform, with a focus on its ability to help users identify and mitigate bias in machine learning models. The study aimed to explore how conversational explanations and LLM chatbots impact users' understanding and ability to steer models toward fairer outcomes. The analysis is divided into several sections, each addressing different aspects of the study's results, including both quantitative and qualitative data.

The first section delves into the quantitative data derived from the post-study questionnaires, including the System Usability Scale (SUS) and NASA Task Load Index (NASA-TLX). These standardized instruments provide insights into the usability and perceived workload associated with both the manual and conversational model steering components of FairML. The quantitative analysis seeks to determine whether the different steering methods influenced participants' effectiveness in mitigating bias and their overall user experience.

Following the quantitative analysis, the chapter explores the qualitative data obtained from open-ended questions and participant feedback. This section provides a richer understanding of the user experience, highlighting participants' preferences, challenges, and suggestions for improvement. The qualitative insights are crucial for identifying specific areas where the platform excels and where it may require further development.

Finally, the chapter integrates both the quantitative and qualitative findings to offer a comprehensive assessment of the FairML platform. By combining these data sources, the analysis addresses the primary research questions regarding the impact of conversational explanations on the understandability of bias, the effectiveness of LLM chatbots in model steering, and the ability of these tools to provide causal reasoning for model bias.

Through this multi-faceted analysis, the chapter aims to provide robust evidence of the capabilities of the FairML platform and its potential to contribute to more transparent and fair AI systems. The findings discussed here will also inform the subsequent discussion and conclusions of this thesis, offering insights into the broader implications for the design and implementation of AI tools that prioritize fairness and accessibility.

Table 5.1: Results of the Mann-Whitney U test for the different questionnaires.

Questionnaire	**p-value**	**u-value**
Usefulness	0.459	14.0
Bias	0.721	12.5
Workload	0.493	7.0
Understandability	0.823	9.0

5.1 Responses

The responses from the post-study questionnaires (Table B) are visualized in Figures 5.1 and 5.2. These figures illustrate the participants' answers regarding bias, perceived understandability, manual model steering usefulness, and conversational model steering usefulness. Box plots of the answers can be seen in figure 5.3.

5.2 Conversational Model Steering

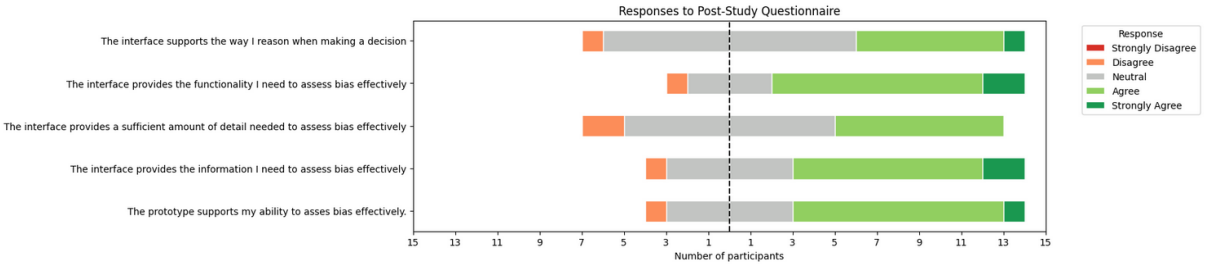
We employed the Mann-Whitney U test to assess whether there was a significant difference in responses between participants who performed well and those who did not perform as well in mitigating bias through conversational model steering. This test was chosen because it is well-suited for comparing two independent groups, particularly when the data does not necessarily follow a normal distribution. The null hypothesis and the alternative hypothesis for the Mann-Whitney U test are as follows [1]:

- **Hypothesis H_0:** There is no difference between the participants who performed badly on mitigating the bias compared to those who performed well.
- **Hypothesis H_a:** There is a difference between the participants who performed badly on mitigating the bias compared to those who performed well.

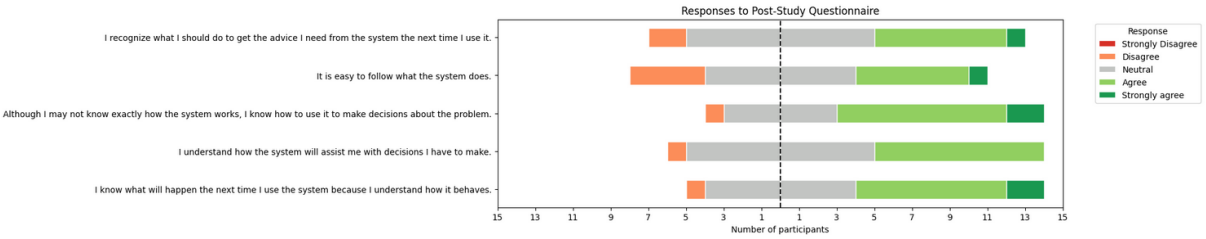
The results, summarized in Table 5.1, indicate that no significant differences were found between the two groups across the measured variables, as all p-values exceed the commonly accepted significance level of 0.05. Specifically, the smallest p-values observed were 0.459 for perceived usefulness and 0.493 for workload. These results suggest that the distributions of the two groups are quite similar, indicating that participants' ability to mitigate bias was not strongly influenced by the conversational model steering component.

Figures 5.4 to 5.7 provide a visual comparison of the responses from participants who performed well versus those who did not. For instance, Figure 5.4 illustrates the distribution of responses regarding the usefulness of the system, where the lack of significant difference between groups is visually apparent. Similarly, Figure 5.5 shows responses related to bias, with both groups reporting similar experiences.

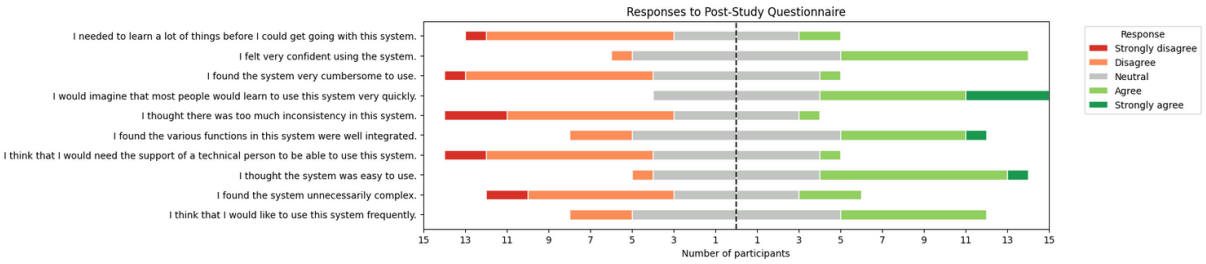
The analysis of these figures supports the statistical findings that there were no notable differences in how participants perceived the usefulness, workload, and bias in the system, regardless of their performance level. This consistency across groups may imply



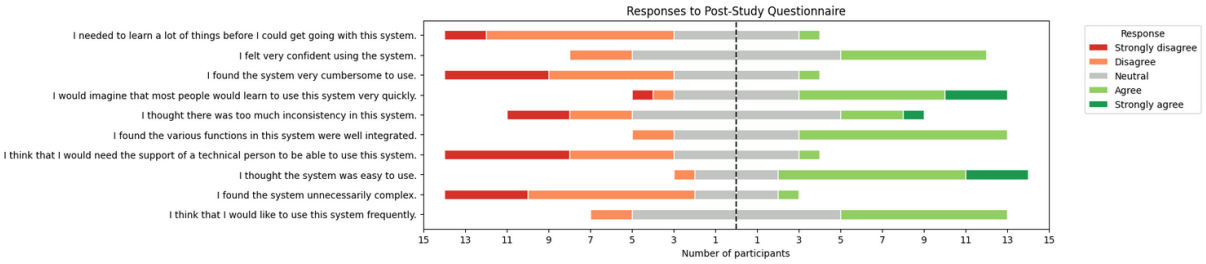
(a) Bias



(b) Perceived understandability



(c) Manual model steering usefulness



(d) Conversational model steering usefulness

Figure 5.1: Answers of the post-study questionnaires

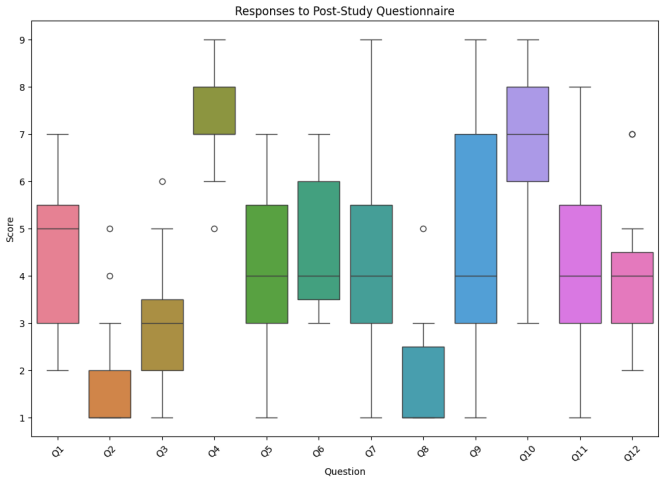
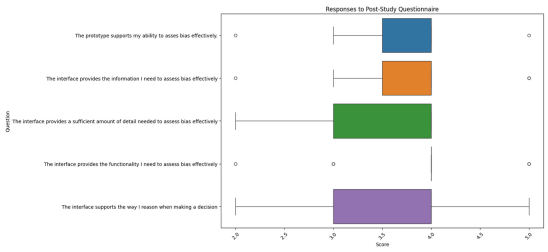
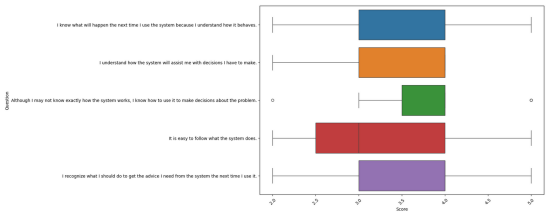


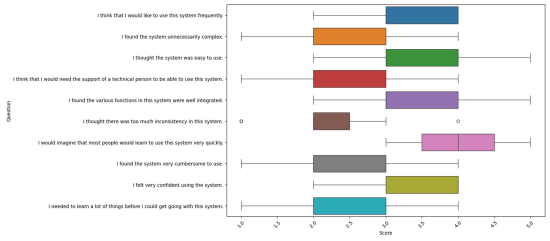
Figure 5.2: Answers of the NASA-TLX workload questionnaire



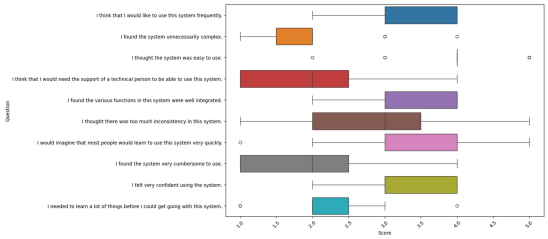
(a) Bias



(b) Perceived understandability



(c) Manual model steering usefulness



(d) Conversational model steering usefulness

Figure 5.3: Boxplots of the post-study questionnaires

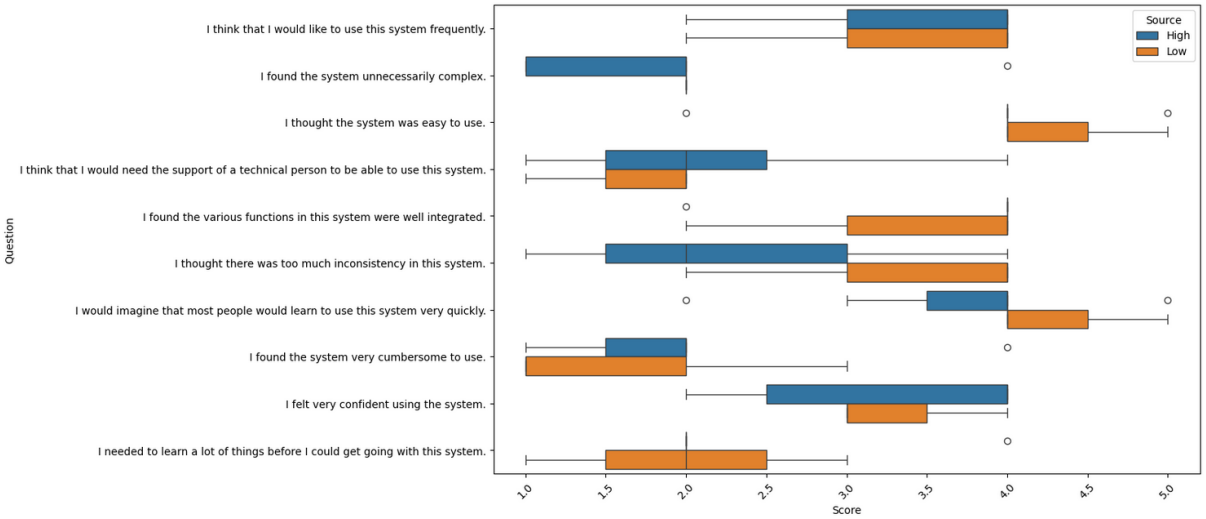


Figure 5.4: Comparison of the participants that performed well and poorly on the usefulness questionnaire.

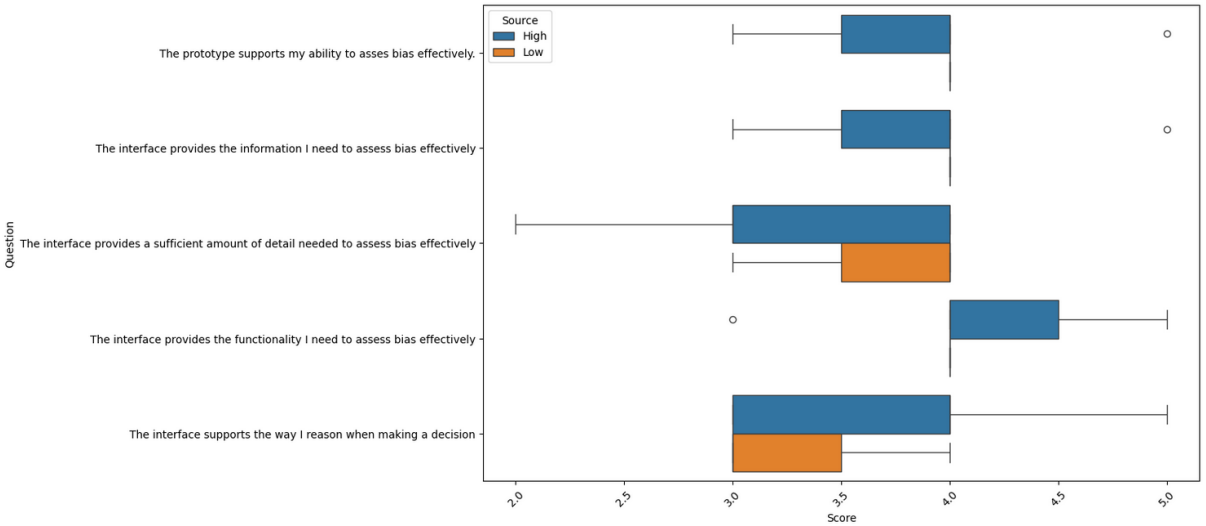


Figure 5.5: Comparison of the participants that performed well and poorly on the bias questionnaire.

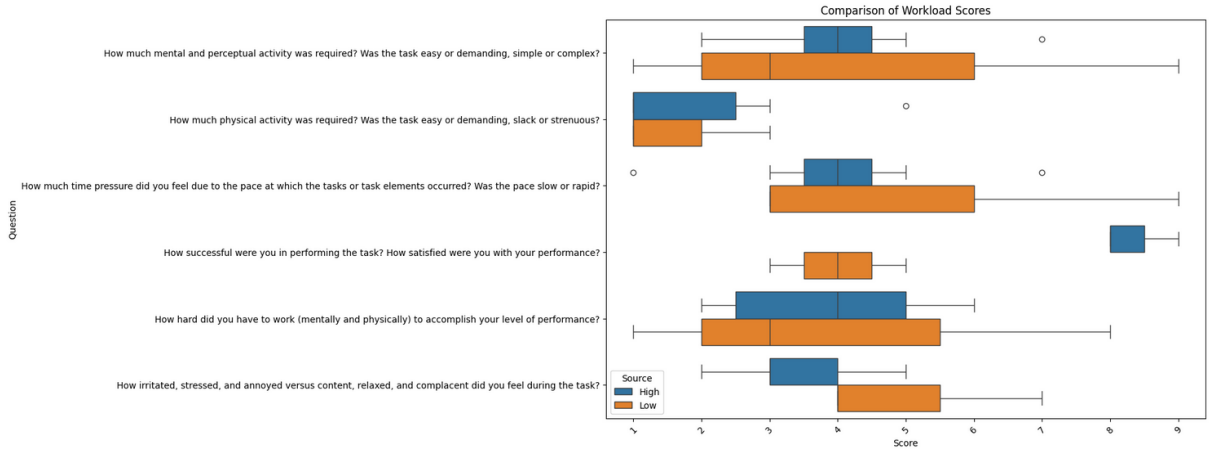


Figure 5.6: Comparison of the participants that performed well and poorly on the workload questionnaire.

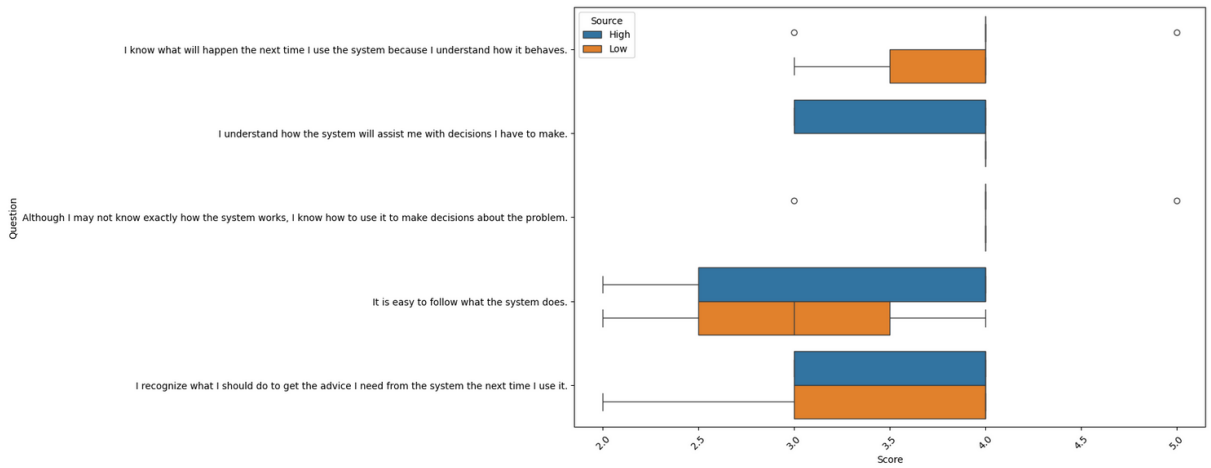


Figure 5.7: Comparison of the participants that performed well and poorly on the understandability questionnaire.

that while the conversational model steering was generally well-received, it did not necessarily enhance participants' ability to mitigate bias in a manner that differed significantly between those who performed well and those who did not.

In summary, while the Mann-Whitney U test did not reveal any significant differences between the two groups, the data provides valuable insights into how participants interacted with the conversational model steering component. These findings are crucial for understanding the broader implications of using conversational agents in bias mitigation and highlight areas for potential improvement in future iterations of the FairML platform.

5.3 Qualitative Data

Participants were asked to provide feedback on their preferences for using conversational model steering, as well as suggestions for improvements. This feedback is crucial in understanding the user experience and identifying areas for enhancement, especially given the novel nature of this research on conversational model steering.

Overall, participants highlighted several key themes in their responses. Those who found it challenging to address bias issues often mentioned the need for more guidance and clearer feedback from the system. Conversely, participants who were more successful in addressing bias issues appreciated the system's ability to handle complex queries and provide in-depth responses.

Here, we provide some of the interesting textual responses given by the participants. Moreover, we add the users' response to the question "*How successful were you in performing the task? How satisfied were you with your performance?*" to see if the given response was from a user who was able to successfully address the bias issues through the conversational model steering or not (a score from 1 through 10).

In which cases would you prefer to use the chatbot model steering?

Participants expressed varied preferences for using the chatbot model steering. Notably, those who found it challenging to address bias issues provided the following insights:

- "*When trying things out or when you don't really know yet what to do.*" (3)
- "*Using the chatbot it's very fast to get the wanted information*" (4)

Conversely, participants who were more successful in addressing bias issues indicated:

- "*Contextual / comparison questions. Natural language interfacing is easy to use in most cases* " (9)
- "*When I want to ask more in-depth questions.*" (8)
- "*Using the chatbot it's very fast to get the wanted information*" (8)

Regardless of their success in addressing bias, both groups highlighted the chatbot's utility for in-depth inquiries, which is not feasible with manual steering interfaces.

Do you have suggestions on how we could improve the chatbot model interaction?

Participants also provided valuable suggestions for enhancing the conversational model steering:

- "*Feedback on when the chatbot actually changed something. It will tell you it did but I did not always believe the chatbot.*" (3)
- "*The response accuracy need to be improved, also if [it] could [be] possible and in the scope to make it more smart using AI will be wonderful.*" (4)

Participants who were able to address the bias issues suggested the following:

- "*The possibility to change between different models*" (9)
- "*Different languages, voice activated*" (8)
- "*Better overview of chatbot functionalities*" (8)

Participants who struggled with bias issues focused on improving consistency and reliability, while those who succeeded suggested adding new functionalities. This indicates that while the basic chatbot functions need refinement, expanding its capabilities could significantly enhance user experience for more advanced users.

Some of these expanding capabilities depend on which LLM the application uses (e.g. GPT-4, Llama, Mistral, etc.). Since, some models have more built-in functionalities like; multiple languages that are supported, voice input, etc.

In addition to the general feedback provided, it is essential to contextualize the qualitative insights with the corresponding quantitative performance data (Section 5.2) to better understand the patterns in user experiences. The participants' success in mitigating bias using the conversational model steering component was varied, and this variance is reflected in the qualitative feedback.

For instance, participants who rated their performance as low (scoring between 3 and 4) often indicated a need for clearer guidance and more explicit feedback from the system. Their comments, such as needing more clarity on when the chatbot actually implemented a change, suggest that these users struggled with trust and transparency issues. This feedback aligns with the quantitative findings where no significant difference was observed between different user groups, indicating that the system's current design may not adequately support less confident users.

Conversely, those participants who reported higher satisfaction with their performance (scoring 8 or 9) tended to provide more positive feedback and suggested advanced functionalities, like switching between models or using voice-activated commands. These suggestions point to a more sophisticated interaction with the system, where the users are not

only engaging with the provided features but are also thinking about potential enhancements. This advanced engagement could be correlated with their better performance in mitigating bias, as these users are likely more comfortable navigating and utilizing the platform's capabilities.

Furthermore, the qualitative responses reveal a nuanced understanding of the system's capabilities and limitations. For example, successful users identified specific scenarios where the chatbot's natural language processing was particularly beneficial, such as in contextual or comparative inquiries. These insights suggest that while the chatbot is effective in handling complex queries, there may be opportunities to improve its performance in more straightforward tasks to assist users who may find these challenging.

The qualitative data, when combined with the quantitative analysis, underscores the importance of tailoring the conversational model steering to different user experience levels. This could involve developing adaptive features that adjust the level of guidance or feedback based on the user's interactions and performance. Such an approach could enhance the system's overall usability and ensure that it supports users across a broader spectrum of expertise.

In summary, the qualitative feedback highlights the diverse experiences of participants and provides valuable direction for refining the FairML platform. By addressing the concerns of less successful users while also incorporating the advanced suggestions from more adept users, future iterations of the system can better meet the needs of all users and improve their ability to identify and mitigate bias in machine learning models.

5.4 Logged Data

The FairML platform also records detailed logs of user interactions to provide insights into user behavior and engagement with the system. These logs capture every interaction a user has with the platform, including the timestamps of actions taken. By analyzing these logs, we can infer the duration of a user's engagement with the platform by calculating the time difference between the first and last recorded interaction. This, however, is an approximate measure since users might pause or multitask during their session.

The logged data also records the number of interactions each user has with the system, which offers a glimpse into the user's level of engagement and interaction style. Figure 5.8 visualizes the number of interactions for users categorized into two groups: those who were successful in mitigating bias and those who were not. The data reveals a clear distinction between the two groups, with participants who were confident in their ability to mitigate bias demonstrating a higher number of interactions compared to those who were less successful.

These insights highlight the importance of user interaction frequency as a potential indicator of success in bias mitigation. The higher engagement level among successful users suggests that more frequent interaction with the platform could be associated with better outcomes. However, it is important to consider that high interaction alone does not

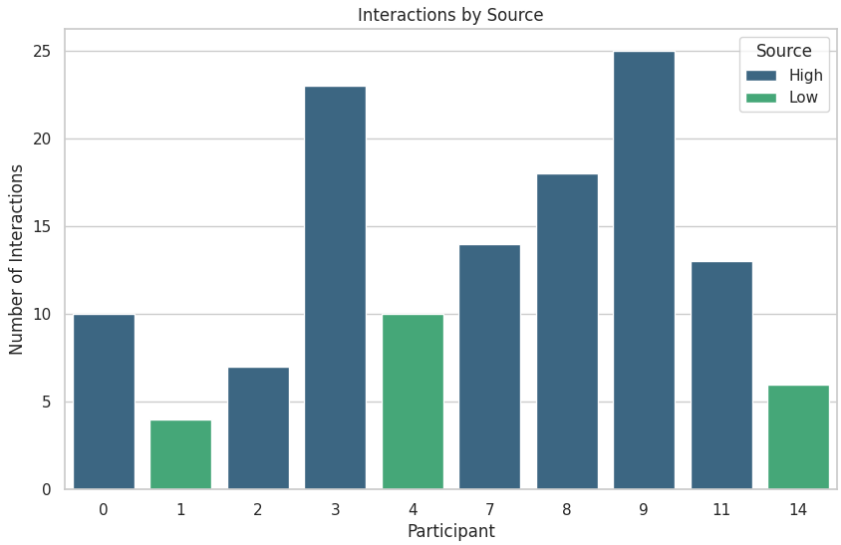


Figure 5.8: Amount of interactions for the users of the two groups, high indicates the users that were able to perform the bias mitigation, while low indicates the users that were not able to mitigate the bias.

guarantee success; the quality and purpose of these interactions also play a significant role.

The logged data, therefore, provides valuable information not only about how users engage with the platform but also about the potential relationship between user behavior and the effectiveness of bias mitigation strategies. This information can be instrumental in refining the platform to better support users, particularly those who may need more guidance or encouragement to interact more frequently and effectively with the system.

In future iterations of FairML, it may be beneficial to explore how the system can be designed to encourage more meaningful interactions, especially for users who may be struggling to achieve successful bias mitigation. This could involve implementing adaptive feedback mechanisms that guide users based on their interaction patterns, ensuring that the platform is both responsive and supportive of different user needs.

Chapter 6

Discussion

This chapter provides an in-depth interpretation of the results presented in the previous chapter, discussing their implications in the context of the research questions and the broader field of AI bias and model steering. The discussion integrates both the quantitative and qualitative findings, offering a comprehensive analysis of the effectiveness and usability of the FairML platform.

The chapter begins by revisiting the primary research questions:

1. How do conversational explanations impact the understandability of bias issues in ML models?

2. How do LLM chatbots facilitate users in model steering to reduce bias issues in ML models?

3. How do conversational explanations provide causal reasoning for model bias?

Each research question is addressed in turn, drawing on the data to evaluate how well the FairML platform meets these objectives. The discussion highlights the strengths of the platform, such as its ability to improve users' understanding of bias and its user-friendly interface, as well as any limitations identified during the study.

Additionally, this chapter explores the practical implications of the findings for the design and implementation of AI systems. It considers how the insights gained from this research can inform the development of more effective tools for bias detection and mitigation, emphasizing the importance of user-centered design and the potential of conversational AI to enhance model interpretability.

The discussion also addresses the study's limitations, such as the small sample size and the homogeneity of the participant group, and suggests directions for future research. These suggestions aim to build on the current study's findings, proposing further investigation into the scalability of the FairML platform and its applicability to a broader range of users and contexts.

6.1 Conversational Model Steering

Research Question 1

How do conversational explanations impact the understandability of bias issues in ML models?

The study's findings highlight the significant impact that conversational explanations have on improving users' understanding of bias issues in machine learning (ML) models. Specifically, the FairML platform's conversational interface, which allows users to interact with the model in natural language, was shown to enhance comprehension of complex bias-related concepts.

Chapter 5 provides detailed insights into this impact. Figures 5.1b and 5.3b illustrate that participants rated the perceived understandability of the system highly, especially when using the conversational model steering component. Users found that the ability to query the system in natural language and receive detailed, context-specific explanations helped them to better understand how certain biases were manifesting in the model's predictions. For example, participants noted that the chatbot's explanations of how particular features contributed to biased outcomes were much clearer than traditional, more technical explanations. This finding was further supported by qualitative feedback, where users expressed that the conversational approach made it easier for them to engage with the system without needing extensive prior knowledge in ML or bias detection.

Moreover, the study revealed that the conversational explanations were particularly effective in breaking down the decision-making processes of the model into understandable segments. Users who may have struggled with interpreting complex statistical data or algorithmic outputs found that the conversational interface provided a more intuitive and accessible means of understanding these processes. This enhancement in understandability is crucial, as it allows a broader range of users—including those without technical expertise—to critically evaluate and potentially mitigate bias in ML models.

Research Question 2

How do LLM chatbots facilitate users in model steering to reduce bias issues in ML models?

The integration of large language model (LLM) chatbots within the FairML platform played a pivotal role in facilitating users' efforts to steer models towards bias mitigation. The effectiveness of the LLM chatbots in assisting users was evident in both the quantitative and qualitative data collected during the study.

The Mann-Whitney U test results, as detailed in Table 5.1, showed no statistically significant difference between participants who performed well and those who did not in bias mitigation. However, this lack of statistical significance does not diminish the observed effectiveness of the chatbot. Figure 5.8, which visualizes the number of interactions logged by the platform, demonstrates a clear distinction between users who were successful in bias mitigation and those who were not. Specifically, participants who engaged

more frequently with the chatbot component exhibited a higher number of interactions and, correspondingly, a greater ability to identify and reduce bias. This suggests that the chatbot's real-time feedback, adaptive responses, and a user-friendly interface were instrumental in guiding users through the complex task of bias mitigation.

Qualitative feedback further supports this conclusion. Successful users often cited the chatbot's ability to handle specific, nuanced queries about model behavior and feature importance as a key to their success. The chatbot's interactive nature allowed users to iteratively refine their understanding and approach to model steering, something that was less achievable through the manual model steering interface. However, some participants did highlight areas for improvement, particularly regarding the consistency of the chatbot's responses. For example, there were instances where users felt uncertain whether the chatbot had correctly implemented their input, pointing to the need for more transparent and confirmatory feedback mechanisms within the system.

Overall, the LLM chatbot facilitated a more accessible and effective model steering process, enabling users to engage with bias mitigation in a more intuitive and responsive manner than traditional methods.

Research Question 3

How do conversational explanations provide causal reasoning of model bias?

Conversational explanations were found to be highly effective in providing causal reasoning for model bias, which is a critical aspect of understanding and addressing bias in ML models. The study's results indicate that the FairML platform's ability to deliver causal explanations through its conversational interface significantly enhanced users' ability to comprehend the underlying causes of bias.

The qualitative data revealed that users who successfully identified and mitigated bias consistently cited the chatbot's explanations as instrumental in their understanding. For instance, when users queried the chatbot about why certain features were leading to biased outcomes, the chatbot was able to provide clear, logically structured explanations that outlined the causal relationships between those features and the model's predictions. This was particularly beneficial for participants who lacked deep technical expertise, as it allowed them to grasp complex causal mechanisms without needing to interpret technical jargon or complex mathematical representations.

Figure 5.1c from the results chapter shows that participants rated the usefulness of the chatbot highly, particularly in terms of its ability to clarify causal relationships. The chatbot's ability to articulate why certain biases were occurring, rather than just identifying their presence, was a key factor in this positive evaluation. This suggests that conversational explanations can play a crucial role in enhancing users' understanding of not just what biases exist, but why they exist and how they can be addressed.

Furthermore, the study highlighted that while the chatbot was effective in conveying causal reasoning, there is still room for improvement. Some participants noted that the

causal explanations could sometimes be too general or lacked the specificity needed for certain complex scenarios. This feedback suggests that future iterations of the FairML platform could benefit from refining the depth and accuracy of these explanations, potentially by integrating more sophisticated causal inference algorithms or by allowing users to explore causal relationships in greater detail through follow-up queries.

6.2 Limitations

This research, while providing valuable insights, is not without its limitations. These limitations must be carefully considered when interpreting the results and their broader implications.

1. The final user study included a relatively small sample of 15 participants. Although this sample provided critical data points within this specific user group, the findings should be interpreted with caution. The limited sample size may constrain the generalizability of the results, suggesting the need for further studies with more diverse and larger participant pools to validate these findings.

2. The FairML platform, utilized in this study, has inherent limitations.
   Firstly, the platform is not capable of autonomously identifying bias within datasets or models. Currently, no algorithms can definitively detect or classify bias with absolute certainty, which is a significant limitation in the field of bias detection.
   Additionally, the study utilized a single Large Language Model (LLM) without any fine-tuning. Specifically, the model selected was the most cost-effective option available from OpenAI. It is conceivable that alternative models, particularly those fine-tuned on specific datasets, could yield different or more effective results. Exploring a broader range of LLMs was outside the scope of this thesis, though it is an avenue worth pursuing in future research.

3. The study sample was relatively homogeneous, consisting of participants with similar backgrounds. This homogeneity could introduce bias, limiting the external validity of the findings. Future research should aim to include a more diverse participant pool to better capture how different user groups interact with the FairML platform.

4. Participants' familiarity with the system could have influenced their performance. Those who interacted with the platform multiple times may have shown improvement not necessarily due to the system's efficacy but rather due to increased familiarity. This potential learning effect should be accounted for in future studies to more accurately assess the system's effectiveness.

6.3 Future Work

The FairML platform has demonstrated its potential as a tool for bias mitigation in machine learning models. However, several avenues for future work could further enhance its capabilities and broaden its impact.

One area of future development involves improving the conversational interface. Although the current chatbot effectively assists users in model steering, there is room for enhancement. Incorporating more advanced natural language processing techniques could enable the chatbot to handle more complex queries and provide more nuanced explanations. Furthermore, future iterations of the chatbot could leverage machine learning models that adapt to user interactions, making the system more personalized and responsive to individual needs.

Another promising direction is the integration of real-time user feedback. By allowing users to rate the system's responses or provide comments on the clarity and usefulness of explanations, the platform could continuously evolve based on user input. This feedback loop would not only improve the system's accuracy and user satisfaction but also help identify areas where users commonly struggle, informing targeted improvements.

Longitudinal studies could also be conducted to understand the long-term effects of using the FairML platform. By tracking users over time, researchers could gain insights into how sustained interaction with the platform influences users' understanding of bias and their ability to mitigate it. This could also help in identifying any learning curves or potential fatigue in using the system.

Chapter 7

Conclusion

The research presented in this thesis set out to explore the potential of large language models (LLMs) and conversational interfaces to enhance the fairness of machine learning (ML) models by improving users' ability to understand and mitigate bias. The development and evaluation of the FairML platform, which integrated both manual and conversational model steering components, provided a robust environment to investigate these issues.

Key Findings:
The study demonstrated that conversational explanations significantly improved users' understanding of bias issues in ML models. The ability to interact with the model in natural language allowed participants to grasp complex bias-related concepts more effectively than through traditional manual methods. This was evidenced by the high ratings for understandability in the post-study questionnaires, as well as the qualitative feedback, which highlighted the clarity and accessibility of the chatbot's explanations.

The integration of LLM chatbots within the FairML platform facilitated more effective model steering, particularly in the context of bias mitigation. While the quantitative results did not show a statistically significant difference in performance between users who performed well and those who did not, the qualitative data indicated that users who engaged more frequently with the chatbot were more successful in identifying and reducing bias. This suggests that the real-time, adaptive nature of the chatbot played a crucial role in guiding users through the model steering process.

The study also found that conversational explanations were effective in providing causal reasoning for model bias, which is essential for understanding and addressing the root causes of biased outcomes. Participants appreciated the chatbot's ability to articulate why certain biases were occurring, which helped them to take more informed actions in mitigating those biases.

In conclusion, this thesis contributes to the growing body of research on explainable AI and bias mitigation by demonstrating the potential of LLM-based conversational interfaces to enhance the fairness of ML models. By making complex bias concepts more understandable and providing tools that facilitate effective model steering, the FairML platform represents a significant step towards more transparent and equitable AI systems. This work underscores the importance of user-centered design in AI, particularly in cre-

ating tools that empower all users to participate in the ongoing effort to make AI fairer and more just.

Appendix A

Recruitment

A.1 Recruitment Documents

Ik ben laatste jaar student master in de Toegepaste Informatica en doe mijn masterproef over bias in machine learning. Ik ben op zoek naar meerdere personen die met een applicatie willen interageren en nadien feedback willen geven over de gebruiksvriendelijkheid en de bias van de applicatie. Het onderzoek zal ongeveer 45 minuten duren en zal plaatsvinden in eind april. Dit onderzoek is goedgekeurd door het SMEC (dossiernr. G-2023-7463).

Bij interesse kan je mij contacteren via e-mail: yori.verbist@student.kuleuven.be

Informed consent

Title of the research: Making Machine Learning Fairer for Everyone: How LLMs Can Empower Lay Users

Name + contact details of supervisor and researcher(s):
Yori Verbist, yori.verbist@student.kuleuven.be | 0476405549 | Departement Computerwetenschappen
Aditya Bhattacharya, aditya.bhattacharya@kuleuven.be | +3216374865 | Departement Computerwetenschappen - Mens-Machine Interactie | Celestijnenlaan 200a - bus 2402 - 3001 Leuven
Verbert Katrien, katrien.verbert@kuleuven.be | +3216328286 | Departement Computerwetenschappen - Mens-Machine Interactie | Celestijnenlaan 200a - bus 2402 - 3001 Leuven

Goal and methodology of the research:
The goal of this research is to measure trust and satisfaction of an interactive AI model, where the user can adjust the importance features of the model themselves without any AI knowledge.

Duration of the experiment:
The duration of the experiment will last as long as the user wants to interact with the application. On average this will be around 45 minutes.

- ☐ I understand what is expected of me during this research.
- ☐ I know that I will participate in the following trials or tests:
  Answering multiple questionnaires
  Using the application
- ☐ I know that my participation may be associated to risks or discomforts:
  Not applicable
- ☐ I or others can benefit from this research in the following ways:
  Getting first-hand experience with how AI models work
- ☐ My participation offers a contribution to the scientific research. I know that I will not receive any further reward or compensation for my participation.
- ☐ I understand that my participation to this study is voluntary. I have the right to stop participating at any time. I do not have to give a reason for this and I know that it will not have any negative repercussions for me.

  In the context of the GDPR the collected data will be processed under public interest as the legal basis. When you end your participation the data that were already collected can still legally be included in the research and do not need to be deleted by KU Leuven.
- ☐ The results of this study can be used for scientific goals and may be published. My name will not be published. The confidentiality of the data will be protected in all stages of the research.
- ☐ I would like to be informed about the results of this research. The researchers may contact me for this purpose using the following e-mail address.

Drawn up in duplicate.

_____

- For questions and for the execution of my rights (access to my data, rectification of the data, …) after my participation I know that I can contact:
  yori.verbist@student.kuleuven.be
  More information with regard to privacy in research can be found at https://kuleuven.be/privacy/en/. With further questions about privacy issues I can contact the data protection officer: dpo@kuleuven.be

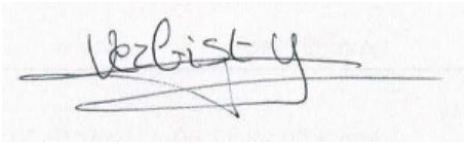
- This study has been reviewed and approved by the Social and Societal Ethics Committee (SMEC) of KU Leuven (G-2023-7463). In case of complaints or other concerns with regard to the ethical aspects of this research I can contact SMEC: smec@kuleuven.be

- I know that I can contact the individuals/organizations below if I would experience any discomfort or difficulties as a result of some of the subjects that were the topic of this research:
  Yori Verbist yori.verbist@student.kuleuven.be

I have read and understood the information in this document and I have received an answer to all my questions regarding this research. I give my consent to participate.

Date:
Name and signature of the participant

Name and signature of the researcher
Yori Verbist



Drawn up in duplicate.

Information about what will be asked during participation

Information about what will be asked during participation You will be asked to work with the application in which the user can see which parameters the AI model uses to make a certain decision, and can adjust these parameters themselves in case the user thinks they are unfair. Afterwards, you will be asked what the user thinks of the application in terms of user-friendliness. The use of the application will take about 45 minutes (the user may also work with it longer if they wish) and filling in the feedback will take a maximum of 5 minutes.

Information about the processing of your personal data

Information about the processing of your personal data As part of your participation in the research “Making Machine Learning Fairer for Everyone: How LLMs Can Empower Lay Users”, personal data about you will be collected and processed. This processing will be done in accordance with the General Data Protection Regulation (GDPR). In this letter, we would like to give you more information about the use and storage of this data.

In the information letter for participation in the research, you will find more information about the categories of data that will be collected about you during this research. These include personal data such as:

- Identification data:
    - Email address
    - Age
    - Gender
    - Experience with AI
    - Eduction level

Use of your personal data

Only personal data that is necessary for the purposes of this research will be collected and processed.

Your data will be pseudonymized in the context of this research. This means that data that can identify you, such as email addresses, will be disconnected from the other data of the research and replaced by a unique, random identifier. In this way, it is no longer immediately visible which data comes from which specific person. Only the researcher can link the data back to a specific person via the unique code. However, this will only happen in exceptional cases, for example if you exercise your right to access, rectification or erasure of your data. Also, in the scientific output of this research, such as publications, you will not be identified. The general interest is used as the legal basis for the processing of your data. This means that the research will lead to an increase in knowledge and insight that benefits society (directly or indirectly).

Your data will be stored by the researchers during the course of the research (this is no longer than a month). After this period, the personal data will be permanently deleted if they are no longer necessary for the execution of the research.

Your rights

You always have the right to ask for more information about the use of your data. In addition, you can exercise the right of access, the right to rectification, and the right to erasure of your data, as long as these rights do not make the purposes of the research impossible or seriously hinder them.

If you want to exercise any of these rights, you can contact the researchers using the contact details at the end of this letter.

Contact Information

KU Leuven acts as the data controller for this research. More specifically, only the researchers: Yori Verbist. For questions or to exercise your rights, you can contact via:

yori.verbist@student.kuleuven.be

For further questions and considerations about the processing of your personal data, you can contact Toon Boon, the data protection officer for scientific research at KU Leuven (dpo@kuleuven.be). Please clarify which research this concerns by mentioning the title and the names of the researchers.

If you, after having contacted the data protection officer, would like to file a complaint about how your information is handled, you can contact the Belgian Data Protection Authority (www.gegevensbeschermingsautoriteit.be).

Appendix B

Questionnaires

B.1 Pre-Study Questionnaire

Table B.1: Questions asked in the pre-study questionnaire, with the possible answers given in the second column.

Pre-Study Questionnaire Questions	
What is your age?	Number
What is you gender?	Female/Male/Prefer not to say/Other
What is your education level?	High school/Bachelors/Masters/PhD
How good is your knowledge of AI?	Beginner/Intermediate/Advanced/Expert

B.2 Post-Study Questionnaires

Table B.2: Post-study questionnaire that asked questions about the usefulness of the system regarding the manual model steering part. All questions used a 5-point Likert scale.

	Post-Study Usefulness Questionnaire Questions
	Manual Model Steering
Q1	I think that I would like to use this system frequently.
Q2	I found the system unnecessarily complex.
Q3	I thought the system was easy to use.
Q4	I think that I would need the support of a technical person to be able to use this system.
Q5	I found the various functions in this system were well integrated.
Q6	I thought there was too much inconsistency in this system.
Q7	I would imagine that most people would learn to use this system very quickly.
Q8	I found the system very cumbersome to use.
Q9	I felt very confident using the system.
Q10	I needed to learn a lot of things before I could get going with this system.

Table B.3: Post-study questionnaire that asked questions about the usefulness of the system regarding the chatbot model steering part. All questions used a 5-point Likert scale.

	Post-Study Usefulness Questionnaire Questions **Chatbot Model Steering**
Q1	I think that I would like to use this system frequently.
Q2	I found the system unnecessarily complex.
Q3	I thought the system was easy to use.
Q4	I think that I would need the support of a technical person to be able to use this system.
Q5	I found the various functions in this system were well integrated.
Q6	I thought there was too much inconsistency in this system.
Q7	I would imagine that most people would learn to use this system very quickly.
Q8	I found the system very cumbersome to use.
Q9	I felt very confident using the system.
Q10	I needed to learn a lot of things before I could get going with this system.

Table B.4: Post-study questionnaire that asked open questions about the usefulness of the system.

	Post-Study Usefulness Questionnaire Open Questions
Q1	In which cases would you prefer to use the manual steering over the chatbot interaction?
Q2	Do you have suggestions on how we could improve the manual model steering interaction?
Q3	In which cases would you prefer to use the chatbot interaction over the manual model steering?
Q4	Do you have suggestions on how we could improve the chatbot model interaction?

Table B.5: Post-study questionnaire that asked questions about the bias of the system. All questions used a 5-point Likert scale.

	Post-Study Usefulness Questionnaire Questions
Q1	The prototype supports my ability to asses bias effectively.
Q2	The interface provides the information I need to assess bias effectively.
Q3	The interface provides a sufficient amount of detail needed to assess bias effectively.
Q4	The interface provides the functionality I need to assess bias effectively.
Q5	The interface supports the way I reason when making a decision.

Table B.6: Post-study questionnaire that asked questions about the perceived understandability of the system. All questions used a 5-point Likert scale.

	Post-Study Perceived understandability Questionnaire Questions
Q1	I know what will happen the next time I use the system because I understand how it behaves.
Q2	I understand how the system will assist me with decisions I have to make.
Q3	Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
Q4	It is easy to follow what the system does.
Q5	I recognize what I should do to get the advice I need from the system the next time I use it.

Table B.7: Post-study questionnaire that asked questions about the workload of the system. All questions used a 10-point Likert scale.

Post-Study Workload Questionnaire Questions	
Manual Model Steering	
Q1	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
Q2	How much physical activity was required? Was the task easy or demanding, slack or strenuous?
Q3	How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
Q4	How successful were you in performing the task? How satisfied were you with your performance?
Q5	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Q6	How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Table B.8: Post-study questionnaire that asked questions about the workload of the system. All questions used a 10-point Likert scale.

Post-Study Workload Questionnaire Questions	
Chatbot Model Steering	
Q1	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
Q2	How much physical activity was required? Was the task easy or demanding, slack or strenuous?
Q3	How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
Q4	How successful were you in performing the task? How satisfied were you with your performance?
Q5	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Q6	How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Bibliography

[1] *Assumptions of the Mann-Whitney U test — Laerd Statistics*. URL: https://statistics.laerd.com/statistical-guides/mann-whitney-u-test-assumptions.php.

[2] Aditya Bhattacharya et al. "EXMOS: Explanatory Model Steering through Multifaceted Explanations and Data Configurations". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642106. URL: https://doi.org/10.1145/3613904.3642106.

[3] Shiva Borzooei et al. "Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study". In: 281.4 (Oct. 2023), pp. 2095–2104. DOI: 10.1007/s00405-023-08299-w. URL: https://doi.org/10.1007/s00405-023-08299-w.

[4] John Brooke. "SUS: A quick and dirty usability scale". In: *Usability Eval. Ind.* 189 (Nov. 1995).

[5] Francisco Maria Calisto and Jacinto C. Nascimento. *NASA-TLX Survey*. en. 2018. DOI: 10.13140/rg.2.2.26978.79044. URL: http://rgdoi.net/10.13140/RG.2.2.26978.79044.

[6] Jonathan Dodge et al. "Explaining models: an empirical study of how explanations impact fairness judgment". In: *Proceedings of the 24th international conference on intelligent user interfaces*. 2019, pp. 275–285.

[7] Norman Geoff. "Likert scales, levels of measurement and the "laws" of statistics". In: *Advances in Health Sciences Education* 15.5 (Feb. 10, 2010), pp. 625–632. DOI: 10.1007/s10459-010-9222-y. URL: https://pubmed.ncbi.nlm.nih.gov/20146096/.

[8] Lijie Guo et al. "Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules". In: *27th International Conference on Intelligent User Interfaces* (2022). URL: https://api.semanticscholar.org/CorpusID:247585155.

[9] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[10] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[11] Ninareh Mehrabi et al. “A Survey on Bias and Fairness in Machine Learning”. In: *CoRR* abs/1908.09635 (2019). arXiv: 1908.09635. URL: http://arxiv.org/abs/1908.09635.

[12] Michelle M. Mello and Neel Guha. “Understanding Liability Risk from Using Health Care Artificial Intelligence Tools”. In: *New England Journal of Medicine* 390.3 (2024), pp. 271–278. DOI: 10.1056/NEJMhle2308901. eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMhle2308901. URL: https://www.nejm.org/doi/full/10.1056/NEJMhle2308901.

[13] Atul Rawal et al. “A Quantitative Comparison of Causality and Feature Relevance via Explainable AI (XAI) for Robust, and Trustworthy Artificial Reasoning Systems”. In: *Artificial Intelligence in HCI: 4th International Conference, AI-HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part I*. Copenhagen, Denmark: Springer-Verlag, 2023, 274–285. ISBN: 978-3-031-35890-6. DOI: 10.1007/978-3-031-35891-3_17.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”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*. 2016, pp. 1135–1144.

[15] Patrick Schramowski et al. *Making deep neural networks right for the right scientific reasons by interacting with their explanations*. 2024. arXiv: 2001.05371 [cs.LG].

[16] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[17] Nima Shahbazi et al. “Representation Bias in Data: A Survey on Identification and Resolution Techniques”. In: *ACM Computing Surveys* 55.13s (July 2023), 1–39. ISSN: 1557-7341. DOI: 10.1145/3588433. URL: http://dx.doi.org/10.1145/3588433.

[18] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[19] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[20] Stefano Teso et al. *Leveraging Explanations in Interactive Machine Learning: An Overview*. 2022. arXiv: 2207.14526 [cs.LG].

[21] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[22] Michael L. Waskom. “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[23] Simone Stumpf Aisha Naseer Daniele Regoli Yuri Nakao Lorenzo Strappelli and Giulia Del Gamba. "Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness". In: *International Journal of Human–Computer Interaction* 39.9 (2023), pp. 1762–1788. DOI: 10.1080/10447318.2022.2067936.

[24] Cheng Zhang et al. *Understanding Causality with Large Language Models: Feasibility and Opportunities*. 2023. arXiv: 2304.05524 [cs.LG].

KULeuven Computerwetenschappen
Celestijnenlaan 200A
3000 LEUVEN, BELGIË
tel. + 32 16 32 77 00
fax + 32 16 32 79 96
www.cs.kuleuven.be