Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency

by

Md Ariful Islam Anik

A thesis submitted to The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
November 2020

Thesis Advisor
Andrea Bunt

Author
Md Ariful Islam Anik

Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency

Abstract

Training datasets fundamentally impact the performance of machine learning systems. Any biases introduced during training (implicit or explicit) are often reflected in the system's behaviors leading to questions about fairness and loss of trust in the system. Yet, information on training data is rarely communicated to the stakeholders. In this thesis, I explore the concept of data-centric explanations for machine learning systems that describe the training data to end-users. I design data-centric explanations that focus on providing information on training data. Through a formative study, I investigate the potential utility of such an approach and the data-centric information that users find most compelling. In a second study, I investigate reactions to the explanations across four different system scenarios. The results show that data-centric explanations can impact how users judge the trustworthiness of a system and can assist users in assessing fairness. I discuss the implications of the findings for designing explanations to support users' perception of machine learning systems.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to Almighty God for giving me the strength to pursue my degree away from home and writing this thesis.

I am eternally grateful to my advisor Dr. Andrea Bunt for her wonderful guidance, utmost support, patience, and constant encouragement during my MSc program here at the University of Manitoba. She gave me the freedom to follow my interest, showed great patience when I struggled, and guided me to reach the goal. Besides her guidance on my thesis work, I learned many things from her that helped me to grow as a person. I would also like to thank her for the financial support during my stay. I would also like to acknowledge the Computer Science Department for offering me the Entrance Award.

I would like to thank my committee members Dr. Carson Leung and Dr. Yang Wang for their precious time reading my proposal, checking on my progress, examining my thesis, and providing feedback at all the stages. I would also like to thank Dr. James Young and Dr. Celine Latulipe for their feedback on my thesis during the HCI lab presentations.

I want to express my utmost thanks to the HCI lab members. I feel really lucky to be a part of the lab. Thank you Stela, Patrick, Denise, Vlad, Dan, Adnan, Shahed, Ellie, Ananta for helping me adjust to the lab cultures and graduate life. My time in the lab

became more fun when Lorena, Rahat, and Mahya came into the lab. We had a lot of fun memories and I'll always remember our lunchtimes. Thank you guys for making me do fun stuffs with you. I would also like to thank Raquel, Tina, Chris, Taylor, Annalena, Lena, Jason, Agape for the friendly conversations and the laughs. I also want to thank my friends here in Winnipeg who helped me in many ways over the last two years.

Finally, I want to thank my parents, my siblings, and their spouses for their constant support. Being the youngest in the family, it has been hard staying away from them and my nieces and nephews, but their support kept going. Please always know that whatever I have achieved in my life or will hopefully achieve in my future, it is built on all of your sacrifices. I love you all.

Chapter 1

Introduction

Artificial Intelligence (AI) systems trained via data-driven machine learning (ML) algorithms have permeated society. ML systems are involved in a range of contexts, from targeted advertisements [67,104], to product and content recommendations [4,19,41,99], to informing decisions on matters with substantial individual and societal impacts, such as hiring [17,39,77], finance [29,57], medicine [16,40], and criminal justice [25,43,58]. Despite their importance and impact, such systems are often "black-box" by nature [83] and consequently are not transparent [24,70,82]. The lack of transparency can make it difficult for end-users to interpret and understand system outcomes [24,70,82]. The lack of transparency also can hurt a user's ability to form meaningful trust relationships with machine learning systems [27,76,88,90] and to hold the systems properly accountable for their decisions [11,66].

In light of the above consequences of opaque ML-based systems, there is a growing body of research in the AI and HCI research communities on Explainable AI, with the

goal of devising ways to increase transparency [30,38,66,70,88,89,92] as well as to understand the impact of increased transparency on user perceptions of and interactions with such systems [5,14,20,33,59,86,105]. Much of this work, however, has focused on explaining outcomes and the properties of decisions to end-users [20,30,88,89,92], for example, by explaining factors that influence a system's behaviors, or by relating behaviors to information in an end-user's profile. While valuable, such approaches rarely communicate information on the manner in which the system was trained. Since machine learning algorithms look at the patterns in the training data, the quality and underlying characteristics of training datasets are fundamental to system performance [15]. For example, if the training dataset is not representative of the target population, certain groups can be disadvantaged [3], and any biases in the training data [81] are ultimately reflected and aggravated in the deployed system [3,106]. For example, when a popular word embedding tool was trained on a corpus of Google News articles, implicit gender biases in article coverage caused the system to learn similarly biased word associations (e.g., doctors are men and nurses are women) [6].

Prior work has shown that industry practitioners are well aware of the importance of the training datasets, often revisiting datasets when they notice problems with the systems [51]. Training information, however, is typically not made available to end-users once systems are deployed. This leads to my research questions of whether and how training dataset information could be communicated to end-users. With the goal to make machine learning systems more transparent to end-users, in this thesis, I design explanations that focus on communicating training data information to end-users (I call

them as data-centric explanations) and explore how these explanations can impact end-users' perception of the systems.

1.1. Research Questions

My thesis seeks to answer the following research questions:

1) What types of training data information might be available to communicate to end-users?

2) How should such information be presented to end-users of varying backgrounds in machine learning?

3) What impact could data-centric explanations have on end-users' perceived trust and fairness judgments of ML systems?

1.2. Methodology and Approach

I approached my research questions by i) exploring communicable training dataset information and designing a prototype data-centric explanation to present them, ii) exploring the feasibility of such explanations in a concept exploration study and using the feedback to improve the explanations, and iii) investigating the utility of the explanations across a range of decision-making scenarios in a user study. What follows is a summary of each of these thesis components.

1.2.1. Prototyping Data-centric Explanations

With the goal to find communicable information about training datasets that could be included in data-centric explanation, I first consulted prior work on training dataset

documentation [44] to identify communicable information. I then used an iterative user-centered design process to develop a prototype data-centric explanation.

1.2.2. Concept Exploration Study and Prototype Refinement

I used the prototype data-centric explanation in a concept exploration study where I interviewed 17 participants. I used the prototype explanation to ground the discussions with the participants and elicited feedback on the explanations along with what they know about machine learning system workflows and their outlook on the data-centric explanation. I used the study findings to improve the data-centric explanation.

1.2.3. Investigating the Utility of the Explanations

Finally, in a study with 27 participants of various backgrounds, I investigated the impact of the explanations on trust and fairness judgments across a range of four system scenarios. My findings indicate that participants felt that the data-centric explanations helped them reflect on the training process, impacted their trust in the system, and were particularly important for high-stakes systems. While the explanations received support from all expertise groups in my study, I noted subtle qualitative differences in how machine learning experts and non-experts approached the explanations. For example, machine learning experts questioned whether the information would be understandable to those without machine learning training, whereas the non-experts felt the information was both clear and useful.

1.3. Contributions

In summary, this thesis makes the following contributions:

1) I identify communicable information about training datasets and present them as data-centric explanations from machine learning systems.
2) I present study findings that show the potential for data-centric explanations to influence users' perception of machine learning systems.

The remainder of this thesis is organized in five chapters: Chapter 2 summarizes prior work related to this thesis, Chapter 3 introduces data-centric explanations, Chapter 4 describes the exploratory study for concept exploration and prototype refinement, Chapter 5 presents the second study as well as the findings from the study, and Chapter 6 concludes the thesis.

Chapter 2

Related Work

In this chapter, I review prior work on different approaches to designing explanations in machine learning systems, the effect of explanations on end-users' perceptions of machine learning systems, and approaches to documenting training datasets.

2.1. Approaches to Explanations in Machine Learning Systems

In the field of Explainable AI, a myriad of research has aimed at increasing system transparency of machine learning systems. Popular domains in this body of work include recommender systems [34,65,78,99], healthcare applications [16,18,52,94], finance [12,29,42,45], hiring [39,73], and criminal justice [95,98,103]. Explanations in all these domains have aimed to make the systems more interpretable and to explain the outcomes to the end-users.

Prior work has explored a range of explanation approaches including: *input influence* [5,30,103] (the degree of influence of each input on the system output); *sensitivity based* [5,88,92] (how much the value of an input would have to differ to change the output); *demographic-based* [1,5,99] (aggregate statistics on the outcome classes for people in the same demographic categories as the decision subject); *case-based* [5,14,80] (using an example instance from the training data to explain the outcome); *white-box* [20] (showing the internal workings of an algorithm); and *visual explanations* [50,61,97] (explaining the outcomes or the model through a visual analytics interface). Except for case-based explanations, most of these approaches have focused on explaining the decision-process or the decision factors. The data-centric explanations that I design represent a new approach by focusing on the training data, rather than the features or individual decisions of the systems.

Prior work has also categorized explanations across two key dimensions. One pertains to their degree of specificity [36], categorizing an explanation as either *model-specific* or *model-agnostic*. Model-specific explanations pertain to a particular model and can only explain that model's decisions [16,62,71]. Model-agnostic explanations, on the other hand, can explain decisions from a range of ML models [88,89], enabling a greater degree of generality. A second dimension relates to explanation scope in the sense of supporting end-users in understanding either individual decisions (i.e., *local explanations* [35,80,88]) or the system as a whole (i.e., *global explanations* [1,30]). Local explanations justify individual decisions, whereas global explanations describe how the whole model works. In comparison to local explanations, global explanations have been found to induce more confidence in understanding the model and as being helpful for fairness

judgments [33]. Motivated by this prior work, I design data-centric explanations that are model-agnostic and global.

2.2. Evaluating the Impact of Explanations

In parallel to developing different explanation approaches, numerous studies have investigated the impact of explanations on user perceptions of and interactions with the systems [5,14,20,33,59,60,86,102,105].

Prior work has found that increased transparency through explanations can increase user acceptance of the systems [27,49,60,101]. However, increased transparency does not always lead to increased trust. While many studies have found that explanations impact users' satisfaction and trust positively [9,59,63,75,85], some have found that explanations had no impact on trust [20,27,84], suggesting gaps between the focus of the explanations and user needs. Further, the impact of explanations on trust can depend on the stated accuracy of the system [102], system failures [32,37], soundness of the explanation [64], and the amount of information presented in the explanation [59]. These mixed results motivate further research to understand when and why different types of explanations impact trust.

Prior work has also evaluated the impact of explanations on helping users judge the fairness of machine learning systems. Binns et al. explored people's perception of justice in automated decision-making for four different explanation approaches (*input influence, case-based, demographic-based, sensitivity based*), finding that all explanations had the potential to help people to evaluate fairness in the system's decisions [5]. In a different study, Dodge et al. experimented with the same four explanation approaches on a single

machine learning model [33]. They found that certain explanation approaches were more suited to helping users identify particular fairness issues. For example, they found that global explanations (*input influence, demographic-based*) helped enhanced fairness perceptions of the model more than the other approaches, and could also help users identify model-wide fairness issues. I was motivated by this work to investigate whether a *global* data-centric explanation approach can also support fairness judgments.

2.3. Documenting Training Data

Without a standardized way to document datasets, it is hard for anyone to determine the quality of a dataset and whether or not it is a good fit for a machine learning system [44]. Further, unintentional misuse of datasets or using problematic datasets to train models of high-stakes applications can lead to systematic discrimination by the systems [6,10,58]. To address this problem, Gebru et al. proposed the concept of providing a datasheet for each dataset to document, for example, its motivation, creation, composition, intended uses, distribution, and maintenance [44]. The authors primarily designed this documentation for direct dataset users, i.e., those who develop machine learning systems, suggesting that dataset creators should make this documentation available to increase the transparency of the datasets. Many machine learning researchers have begun adopting this procedure when releasing their datasets [21,93,100] and this approach is starting to gain traction in some organizations (e.g., [2,79]). In this work, I investigate how to communicate training datasets to potential end-users, and how such information might impact their perceptions of machine learning systems.

2.4. Summary

In this chapter, I described prior research that has explored different approaches to explanation in machine learning systems and how they impact end-users' perceptions of the systems. In my thesis, I leverage prior work on dataset documentation to investigate how this information can be communicated to end-users through explanations from machine learning systems. This thesis extends prior work on explanations from machine learning systems by presenting a new approach that focuses on the training data, rather than the features or individual decisions of the systems. This thesis further extends prior work on evaluating the impact of explanations on end-users' perceptions of machine learning systems by investigating the impact of data-centric explanations.

Chapter 3

Data-centric Explanations

In this chapter, I describe how I approached designing data-centric explanations for machine learning systems. I start this chapter by describing how I decided on the content of the explanations and conclude it by describing my design process.

3.1. Content of the Explanations

This thesis focuses on designing data-centric explanations that provide information on a system's training data and training process. My first step was to examine what type of information might be captured during the training process. To this end, I leveraged the datasheets as proposed by Gebru et al. [44], where the authors proposed a standardized in-depth documentation of datasets (called datasheets) that contains information on the motivation, creation, composition, intended uses, distribution, maintenance information

on the dataset. However, I could not proceed with the information in the datasheets directly as the volume of the information was fairly big. Moreover, since this information was primarily designed for machine learning specialists, some of the information might be too complex and less relevant from an end-user's perspective. Therefore, I selected the information I thought would be valuable for end-users in an iterative way.

To select information to present in the explanation, I went through the sample questions about datasets that Gebru et al. [44] provided in their datasheets and selected questions that seemed most relevant from the perspective of an end-user. I categorized the selected questions into several groups based on their relevance to better focus the information. I went through several iterations of selecting and categorizing information receiving peer-feedback from HCI specialists to solicit their input on whether or not they felt I had relevant information in the explanation. In the end, I ended up with five categories of information. Each category contained relevant questions about the dataset that (when answered) could provide more insights into the collection, demographics, usage, issues, and other important information about the dataset. The categories and some sample questions for each category are listed in Table 1.

Category Name	What type of information it contains	Sample questions
Collection	Information on the data collection and associated process	- How many instances are in the dataset? - Who collected the data?
Demographics	Information on the distribution of different demographics	- Gender distribution of the instances? - Race distribution of the instances?
Recommended Usage	List of recommended use cases of the dataset	- Suggested use cases for the dataset? - Where you should not use the dataset?
Potential issues	Information on the potential issues and the concerns about the dataset	- What errors have been identified? - Does the dataset contain sensitive information?
General Information	Overview information about the dataset	- When was the dataset released? - Have any updates been provided?

Table 1: Categories of information to be presented in data-centric explanations

3.2. Designing Data-centric Explanations

Once I had settled on the information that I wanted to include in data-centric explanations, I used an iterative, user-centered design process to generate an initial prototype. Although I pared down the datasheet information significantly, I still had a lot of information to present. Therefore, first, I started with low-fidelity prototyping to explore ways to present the information compactly. I sketched a number of different presentation styles, including creating flowcharts of information, providing summarized information for each of the categories, providing information in a q&a style. Informed by pilot testing, I settled on the question-based approach used by Gebru et al. [44] in the datasheets as I found that this approach helped participants to target the information they were most interested in the explanation. Others have used this question-based

approach in explainable AI and found that this works well to answers potential questions a user might have about the systems [72,74,87]. I created the initial prototype using a prototyping tool called inVision [54].

Figure 1 depicts an overview of the initial prototype. In the figure, we can see the categories I created and a high-level idea on what type of information they contain (Figure 1: A). As an example, the figure provides the details of the collection category (Figure 1: B). All the questions under the collection category along with the answers to one sample question are depicted in the figure. For every category, the prototype contained a set of questions that were answered in the explanations. Additional screenshots of the prototype and the questions for each category can be found in Appendix A.

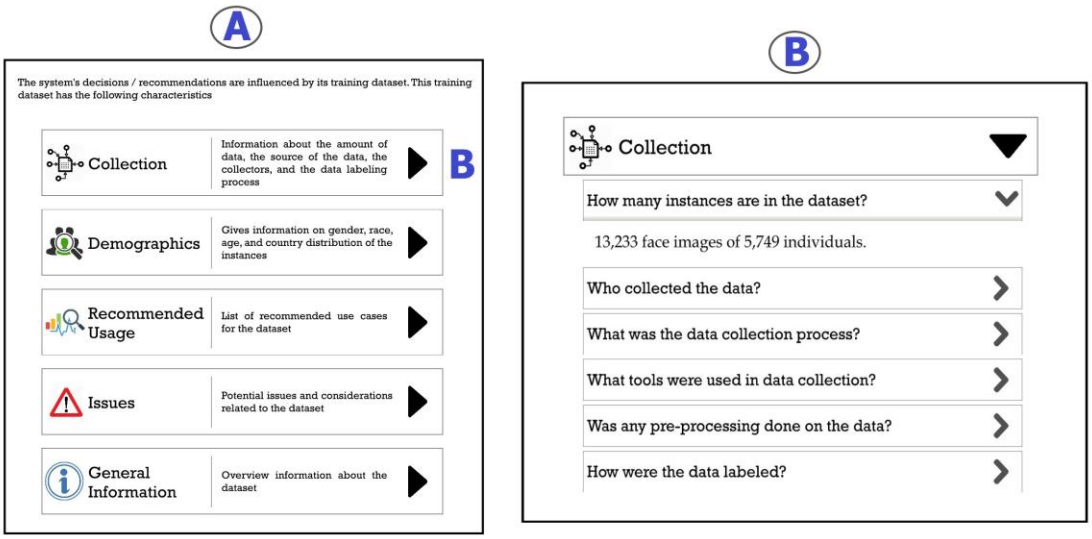


Figure 1: Overview of the initial prototype of data-centric explanation.

3.3. Summary

In this chapter, I described the process I followed to design data-centric explanations that focus on providing information on the training data to end-users. I used prior work on dataset documentation to decide on the content of data-centric explanations and sketching to settle on an initial approach. The next chapter will use this initial version of the explanation in an exploratory study.

Chapter 4

Study 1: Initial Concept Exploration and Prototype Refinement

In my first, exploratory study, I used semi-structured interviews to learn about what people generally know about machine learning systems and their workflows, and how they would feel about potential data-centric explanations from the systems. I used the prototype shown in Figure 1 to ground discussions. In this chapter, I describe the study design, the procedure, and the findings from the exploratory study.

4.1. Participants

I recruited 17 participants (10 men, 7 women) by putting posters around a university campus, and by reaching out to personal contacts (Appendix B). I recruited participants from a range of technical and non-technical backgrounds: five participants self-identified as non-technical and 12 self-identified as a technical person. I targeted most of my recruiting on those outsides of the machine learning field (15 of the participants), however, I also included two experts for the sake of contrast. The average age for the participants was 28 years (SD = 8.83) with ages ranging from 19 to 57. Participants received $20 for their participation. This study was approved by the institutional research ethics board (See Appendix C and Appendix D for certificates).

4.2. Study Method and Procedure

I conducted semi-structured interview sessions with participants. The interview covered questions on participants' existing knowledge and experience with machine learning systems, their ideas on algorithmic fairness, and their thoughts on data-centric explanations. I began the study sessions by asking participants to sign a consent form (Appendix E) and some initial demographics and background questions. I asked the participants about their experiences of receiving decisions from a range of decision-making systems (e.g., ad recommendation, automated hiring, criminal justice system). I then discussed the role that data plays in the decisions from these systems. For most participants, this came naturally into the conversation, and for others, I initiated the topic. I then transitioned to the information categories I created for data-centric explanations. For each category, I asked what they know and what type of information they would be interested in learning about. I then showed them the prototype and asked

for their feedback. I also asked participants about the potential of these explanations to be helpful in their judgment of fairness and their trust in the system. To conclude the session, participants rated each piece of information in the explanation on two 5-point Likert scale items: one for understandability and one for usefulness. I audio-recorded the interview sessions and later transcribed them for data analysis.

4.3. Findings

4.3.1. Data-centric explanations seen as worth pursuing

My discussions with participants revealed insights on why data-centric explanations are worth pursuing. I asked participants if they are aware of fairness issues in machine learning and gave some examples of existing fairness problems. I was surprised that more than half of the participants (9/17) lacked knowledge of fairness issues. For example, the following participant indicates that computers are accurate, which implied that they would also be fair:

> *"Since it is a computer [program], it should be fair. Because [...] computers are very accurate in most of the things. So, I believe [they were fair]." – P4*

Participants mentioned a range of reasons for fairness problems, most of which focused on the data. Several participants (11/17) mentioned that a lack of effort in the data collection process, data providers overlooking important things when collecting and feeding data to the system can lead to fairness problems. For example, P15, who was aware of machine learning but did not have any working experience with it mentioned:

> *"I think another problem could be if you have not enough data put in. So, the facial recognition [example], if you have a society where 90% of the people are*

> *white and 10% of the people are non-white, then if you work from pictures, then you have only 10% of pictures that you feed are from the non-white people which means the program has much less data to work on. […] therefore, the system automatically becomes biased" – P15*

After I discussed data-centric explanations with the participants and showed them the prototype, participants talked about how these types of explanations could be helpful for their trust in the system. Participants discussed how the explanations gave them ideas about the inner workings of the system and the effort of providing the explanations generally left a positive impression. Almost all participants (16/17) felt that the explanations had the potential to impact their trust in the system.

> *"Absolutely, [having] this information increases my trust, unless there is missing information or error in the data. Then I am not gonna trust the system." – P9*

4.3.2. Value of explanations questioned by Machine Learning Experts

I saw some initial indications that user expertise might impact attitudes towards data-centric explanations. While I found that both expert and non-expert participants had positive things to say about having data-centric explanations, the two expert participants also expressed some reservations. They were concerned that the data-centric explanations would not be understandable to non-experts in machine learning and would trigger additional questions. For example, one expert participant with experience in building machine learning systems mentioned that,

> *"I am afraid that the general public might not understand what some of the information means [talking about pre-processing of the data]. It may trigger additional questions for the users, and they will forward these questions to administrators." – P2*

The same participant further mentioned that providing information on the issues could cause people to complain regardless of actual system fairness.

> *"But, some of the things in the issues may be triggering. As long as they have a tab for issues, [people are] always going to say that this dataset is not working. […] So, as a part of the explanation to the user, maybe it is not a good idea to have issues." – P2*

4.3.3. Data-centric explanations are positively received but need more depth

When discussing the prototype, most participants liked the Q&A format, and felt that the information was useful and comprehensive. Most of the participants (14/17) felt that the prototype covers enough information to be helpful.

> *"I think the explanation pretty much asked all the questions here about the [dataset]. Like, I pretty much saw everything I wanted to see for the dataset. Like in the demographics, I saw many distributions." – P5*

A few participants (3/17), however, felt that the information provided in the prototype was a bit shallow and it lacked depth to be useful. Therefore, they want more detailed information.

> *"I feel like the answers […] are way too short and not detailed enough. […] It probably needs to be bit more detailed and technical." – P15*

4.4. Improving Data-centric Explanations

I use feedback from my first study to improve the prototypes for data-centric explanations. I used the same explanation styles (question-based) in the updated prototype since it was well-received by the participants and I did not receive feedback that participants were overloaded with too much information. Yet, I looked for opportunities to streamline the content that did not receive as much positive feedback. First, I used participants' questionnaire responses on the understandability and usefulness of each item to decide which seemed less important. I discarded items if the median score for usefulness was less than 3 (the 5-pt scale), unless the understandability score was also less than 3, in which case I simplified and rephrased the descriptions. I ended up discarding 3 questions (information on the dataset creators, funding source, and maintenance information) from the original prototype where participants indicated they understood the information but rated it low on usefulness. Based on participants' feedback, I also added information about the data labelers and increased the depth of the information when possible. I used web programming (HTML, CSS, JavaScript) to generate the updated prototype.

Figure 2 shows the updated version of the prototype. The main screen, which lists the information categories, and provides a short description of each can be seen on the left. (Figure 2: A) shows an example of the Q&A format for the Collection category. (B), (C), (D), and (E) show the placeholders and short descriptions for the other four categories which expand to reveal the detailed information. Additional screenshots of the prototype and the expanded version of each category can be found in Appendix F.

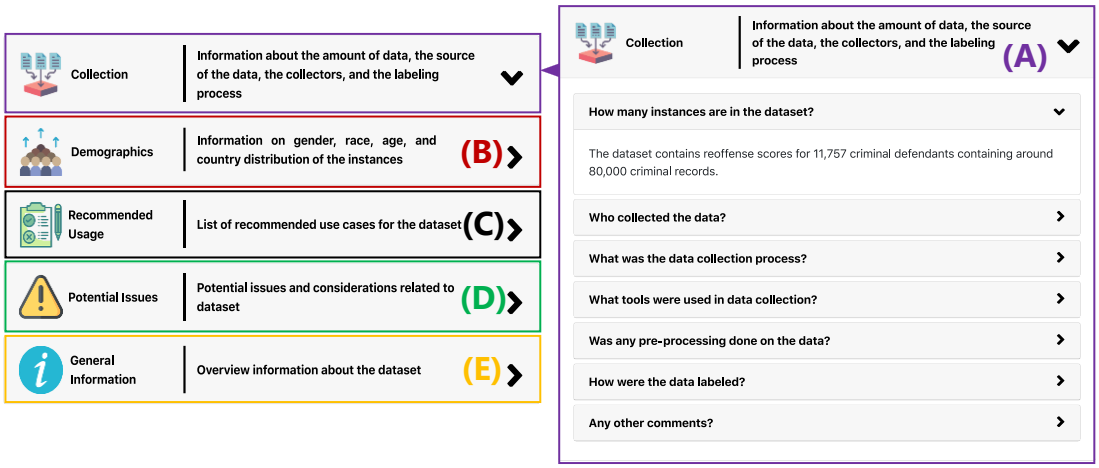


Figure 2: Improved prototype of data-centric explanation.

4.5. Summary

In this chapter, I described my exploratory study that gathered insights on data-centric explanations. The feedback showed positive attitudes toward data-centric explanations from the participants. In the next chapter, I investigate the utility of the updated data-centric explanations in a second user study.

Chapter 5

Study 2: Investigating the Utility of the Explanations across a Range of Scenarios

My first study showed some support for the concept of data-centric explanations and I was able to use the feedback to refine the prototype. In this chapter, I describe my second study, where I investigated how the data-centric explanations impact trust and fairness judgments across a range of potential automated systems scenarios and training data characteristics. Given some of the expertise differences that I observed in my exploratory study, I was also interested in understanding potential differences in participant's perceptions related to their backgrounds in machine learning. I then describe the findings from this study and discuss the implications of the findings.

5.1. Participants

To explore the role of user expertise in machine learning, I sought to include a range of backgrounds in the study. Specially, I recruited participants across three potential expertise dimensions, which I defined as follows

i. ***Expert:*** People who have prior ML experience (e.g., took at least one ML course)

ii. ***Intermediate:*** People from a Computer Science or Engineering background, but no specific ML experience

iii. ***Beginner:*** People from non-engineering or CS backgrounds, without prior experience with ML

I recruited 30 participants for the study by posting advertisements on different online platforms (e.g., Reddit, Twitter) and through snowball sampling. Three participants did not complete the full study (i.e., they did not view all explanations presented to them), leaving me with 27 participants (15 men, 12 women). Participants were between 18 and 54 years old (mean: 28.7, SD = 8.9). Participants had a range of educational backgrounds. For example, 7 participants had completed high school, 9 had completed an undergraduate degree, and 11 had completed a professional or a master's degree. Among the participant pool, I had 9 experts (5 men, 4 women), 8 intermediates (5 men, 3 women), and 10 beginners (5 men, 5 women) according to my definitions above. Participants received $20 for their participation. The study was approved by the institutional research ethics board (see Appendix G for certificates).

5.2. Study Design

My study design included two main factors:

i. Participant Expertise: Expert vs. Intermediate vs. Beginner

ii. Training Data Characteristics: Red Flags vs. Balanced

The first factor, participant expertise, was as defined in the previous section. I also included a second, within-subjects factor, where I manipulated characteristics of the training data presented in the explanations. In the study, participants interacted with the explanations in the context of four different scenarios, representing a range of possible use cases for automated systems. In two scenarios, the explanations showed training data with clear red flags. In the other two scenarios, the explanations depicted relatively balanced training data.

5.3. Automated System Scenarios and Data-centric Explanations

Participants interacted with four different explanations, which collectively covered a range of automated system application scenarios. These scenarios are listed in Table 2 (for more details on the scenarios as presented to participants, see Appendix H). Explanations for two of the scenarios (Predictive Bail Decisions and Facial Expression Recognition) showed obvious red flags in the training data. For example, the demographics distributions (e.g., gender, race) were fairly imbalanced, the sample sizes were fairly small, and prior issues were mentioned with the datasets. For the remaining two scenarios (Automated Admission Decisions and Automated Speech Recognition), the explanations presented relatively balanced training data.

Scenario	Overview of the scenario
Predictive Bail Decision	A system that calculates re-offense risk for a defendant and recommends bail decisions.
Facial Expression Recognition	A system that recognizes the facial expression of a person from a given image.
Automated Admission Decision	A system that assesses student application and recommends admission decisions.
Automated Speech Recognition	A system that recognizes the identities of individuals from speech clips.

Table 2: Scenarios used in the study

To help generate realistic data-centric explanations for each scenario, I consulted reference datasets for bail decisions [68], labeled faces in the wild [53], graduate admissions [107], and speaker recognition [23]. I adjusted this information as needed, for example, to generate potential red flags. For missing information, I either generated fictitious data in a manner consistent with the other explanations or listed the information as being "unknown". Information presented on the explanation for an example scenario can be found in Appendix I.

5.4. Study Procedure

The study sessions took place online, using a video-conferencing platform of the participant's choice. I began the study session by asking participants to sign a consent form (Appendix J) and some introductory demographic questions, including questions on their experiences with computer systems and machine learning (Appendix K). I then presented the four scenarios to the participants, one at a time using Qualtrics [108]. After seeing each scenario description, participants were presented the data-centric

explanation (which would open in a different window) and asked to go through the explanation to explore the degree to which the explanation communicated information on the training dataset information to them and whether or not they found it helpful. The pilot testing for the study revealed that participants need some initial direction on what to do with the explanation once opened. After the participants were done looking at the data-centric explanation for a scenario, they completed a questionnaire consisting of Likert-scale questions (7pt scale). The questionnaire, which can be found in Appendix L, aimed to measure trust in the system, perceptions of system fairness, as well as how much the explanations helped them to get ideas about the system and reflect on the data. I adapted existing scales to measure trust [56] and fairness [5,28]. As shown in Figure 3, this process was repeated for all four scenarios. Participants on average spent 30 min 51 sec (*SD= 13 min 44 sec*) looking at the explanations for the four scenarios and providing responses to the questionnaires. I randomized the order of the scenarios across participants to mitigate potential order effects.

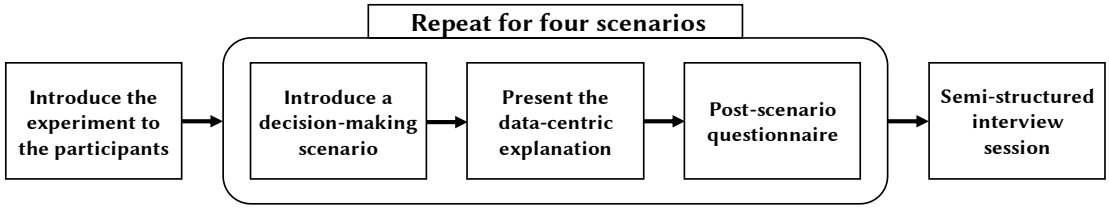


Figure 3: Study procedure

I concluded the study session with a 40-60 min semi-structured interview (sample questions can be found in Appendix M), where I solicited further information from participants on their experiences with machine learning systems, and their perceptions of the data-centric explanations. Throughout the interviews, I probed on issues

surrounding trust, fairness, and characteristics of the system scenarios and training data. The entire session took approximately 90 minutes.

5.5. Data Collection and Analysis

I collected both quantitative data (from the post-scenario questionnaires) and qualitative data (from the post-session interviews). For the quantitative data, I used the non-parametric Kruskal-Wallis H test to analyze the impact of Expertise (a between-subject factor with 3 levels) and the Wilcoxon signed-rank test to analyze the impact of Training Data Characteristics (a within-subject factor with 2 levels). I used $p=0.05$ as the threshold for statistical significance. To analyze the interview data, I first transcribed all the interview sessions. I then conducted bottom-up affinity diagramming [26] on participant quotes from the interview transcripts. My advisor and I were involved in the data analysis. I created the initial affinity diagrams, coding the resulting clusters using an open coding scheme. We then collaboratively looked for themes in the coded data. I did several iterations of this analysis, revisiting the raw data frequently.

5.6. Findings from Questionnaire Data

I first provide an overview of how expertise and training data characteristics impacted participants' perceptions of the systems and the data-centric explanations according to the questionnaire data.

5.6.1. Impact of Expertise

As Table 3 illustrates, Expertise did not significantly impact any of the measures in the questionnaire. I found that irrespective of participants' backgrounds in machine learning or technology, participants rated the explanations highly in terms of getting ideas about

the data and reflecting on the training process. For the other measures, the scores were in the medium range (e.g., around 4-5 on a 7-point Likert-scale) for each expertise level. As I will show in the next section, the scores were low for scenarios where the explanations revealed potential problems with training data and high for scenarios with more balanced training data.

	Scale range	Beginner	Intermediate	Expert	H	Sig
Trust in the system	6.00-42.00	28.00 (7.13)	26.75 (8.00)	31.50 (7.00)	2.146	0.342
Fairness in the system	4.00-28.00	17.75 (8.50)	17.50 (5.50)	21.50 (3.00)	2.089	0.352
Perception of fair training	1.00-7.00	5.00 (2.38)	3.75 (1.63)	5.50 (1.00)	3.636	0.162
Comfort in the system	1.00-7.00	3.75 (2.75)	4.00 (1.88)	5.00 (2.25)	1.622	0.444
Ideas about the data by the explanation	1.00-7.00	6.00 (1.00)	5.75 (1.38)	6.00 (1.25)	1.796	0.407
Refelct on the training process by the explanation	1.00-7.00	6.00 (1.13)	5.75 (3.00)	6.00 (1.25)	0.218	0.897

Table 3: Median (IQR) values for the Likert-Scale questionnaire responses by Expertise level. Since some measures combine multiple questionnaire items, I also provide the scale range (Low-High).

5.6.2. Impact of Training Dataset Characteristics

I also analyzed the questionnaire responses to see if characteristics of the training data impacted participants' perceptions of the system and the utility of the explanations. Table 4 shows that participants had significantly more trust in the system, felt that the system was more fair, and were more comfortable with the system when the explanations indicated relatively balanced training data than when the explanations showed some

potential red flags. Training Data Characteristics, however, did not significantly impact participants' perceived utility of the explanations. Participants rated the explanations highly in terms of giving them a sense of the data and helping them reflect on the nature of the training process, regardless of whether or not the explanations revealed potential problems.

	Scale range	Balanced training data	Training data with red flags	Z	Sig
Trust in the system	6.00-42.00	31.00 (7.00)	26.50 (9.00)	3.635	0.00028
Fairness in the system	4.00-28.00	22.50 (6.50)	17.50 (7.50)	3.945	0.00008
Perception of fair training	1.00-7.00	5.00 (2.00)	4.50 (3.00)	2.652	0.008
Comfort in the system	1.00-7.00	5.00 (2.00)	4.00 (2.00)	2.538	0.011
Ideas about the data by the explanation	1.00-7.00	6.00 (0.50)	6.00 (1.50)	-0.265	0.791
Refelct on the training process by the explanation	1.00-7.00	6.00 (1.00)	6.00 (1.50)	-0.619	0.536

Table 4: Median (IQR) values for the Likert-Scale questionnaire responses according to Training Data Characteristics. Since some measures combine multiple questionnaire items, I also provide the scale range (Low-High).

5.7. Interview Findings

I now present findings from the interviews that provide further insights into how and why the explanations impacted participants' trust and sense of fairness. I also describe commonalities and differences that I observed across the different expertise groupings. In the quotes below, "E" =Expert, "I" = Intermediate, and "B" = Beginner.

5.7.1. Data-centric explanations impact trust in the system

All 27 participants, regardless of machine learning expertise or technical background, indicated in the interviews that the data-centric explanations impacted the degree to which they trusted the systems described in the scenarios.

For a small group of participants (5/27), the mere presence of the explanations was enough to have positive impacts on their levels of trust. These participants saw the explanations as an effort made by the organization to ensure transparency, which ultimately improved their confidence that the systems themselves were trustworthy.

> *"I actually trust [the systems] more now that I have [seen the explanations]. Because, now that I have read it, I think the explanations were transparent. I trust these explanations and they are trying to tell the truth of how they got everything. So yeah, I'd trust it more because they released this information" – P7-I*

The remaining participants (22/27) reported that the specific contents of the explanations impacted their trust. As the following quote illustrates, these participants described how they used the information presented in the explanations to assess whether or not they *should* trust the systems.

> *"Well, I appreciate the disclosure [through the explanation]. Systems like this would get high marks for being transparent. However, just being honest about your specs, doesn't mean that they're necessarily good specs. So, it's good that they reveal that they had a 3.7% margin of error, [but] that's a very high margin of error with something as facial recognition. That's unacceptable. So, is it good*

or bad? I mean, yes, it is good. But it doesn't make me necessarily trust the system more. It depends on the information they are providing." – P30-B

Some participants (10/27) reported that their trust was particularly positively impacted by the data-centric explanations when the amount of data, who is behind the system, and the errors in the data labeling seemed favorable.

"I find the sample size [to be really important]. So, nothing else really matters unless you have a good amount of data. You could say oh, gender was completely equal, however, the sample size [is] of 100. Well, I can't really trust it until your sample size is in a great amount" – P15-B

"I'll trust [a system] more when it has more data points. I mean, the more data points it uses, the more I feel like it would be fair. The more equal distribution [of the] demographics, the fairer, I think. And when I know who's behind it." – P26-I

"The most useful piece of information I found [in the] explanation was the error rate because [if] a system has a zero percent failure rate, I would almost 100% trust that system making the decision." – P14-B

A couple of participants explicitly mentioned that the data-centric explanations revealed problems in the systems that they would not have been aware of otherwise:

"I think if I did not have the explanations, the results would seem more reliable. Because, I had no idea about the distribution of gender, country, and [others]. I had no idea how the data [was] collected, by whom, or by computer or manually. Also, I had no idea about the percentage of errors that were in the collected data.

So, I think the explanations helped me to have a more in-depth idea about the evaluation and the results." – P3-I

One participant also mentioned that they generally expect these types of systems to be sophisticated and accurate, but that the information in explanations suggested otherwise. In the quote below, the participant describes how they were surprised to see Mechanical Turk being used for data processing – they had assumed this type of processing would have been done by an algorithm. This lack of perceived sophistication impacted their trust negatively.

"[Without the explanation], I probably would feel more trust, more confident in the system, just because I would not have a question on how the data is associated with the results. And I wouldn't think they used Amazon Mechanical Turk [in data processing]. I would just kind of feel like oh, they must have come up with something really nifty computer algorithm that did [the preprocessing]." – P26-I

5.7.2. Training data demographics perceived as most influential

Nearly all of the participants (25/27) found value in the demographic information of the data-centric explanations (Figure 2: B) and two-thirds of the participants (18/27) mentioned training data demographics as the most influential aspect of the explanations.

"Demographic information is the most helpful because it basically just gives a broad overview of what data has been used to train the system." – P29-B

Several participants (12/27) reported that the distribution in the demographics helped them to paint a clear picture of the potential biases in the training data. I noticed that

regardless of expertise, participants were able to identify biases in the data from the demographic information.

> *"And I found the bar graphs [in the demographics] a good kind of thumbnail representation, it was more meaningful to see it that way. Because you could immediately spot over inherent bias, whether it was mainly white people, or mainly men, or mainly one country or so on." – P30-B*

> *"By means of these demographics [distributions], I can analyze the results of the system better. For example, I can see the distribution that more than 90% are male, so I can conclude that, these results would be more reliable to the men than the women. It's kind of biased." – P3-I*

> *"[What I understood from] the overall explanation is whether the data will be [able] to give accurate results. […] Taking the example of the admission one, more of the candidates are from Canada. So, I can assume that the model you will train will be biased towards the Canadian students. So, the chances of errors, I can easily predict [that] from the data and the visualization as well." – P12-E*

Two expert participants mentioned that they could situate their own demographic within the distribution to gauge whether or not the system would work for them.

> *"If I look at the [explanation] after I am rejected for admission and I look at like okay, so they are using this particular [dataset] to reject or accept any particular student. Then I would look at the demographics section and on that section I would decide, if I'm from India and the data set contains only 1% of the Indians, so, there will be something in your mind like okay, their model is not trained or*

they do not have the data related to the Indians. So, that may be the case." – P12-E

Along with the demographic information, several participants (14/27) mentioned collection information (Figure 2: A) as an important component of the explanations as it gave information on the sample size, and how the data was gathered:

"Collections was an obvious choice [for being the most important information] because I would like to know how the data was collected, who the collectors were, what was the labeling process. Because data forms the base of everything that the machine learning system has, that will define how it was collected, how it was graded, how it was labeled, how it was classified. [So] that gives you a full overview, like how the data was put into the machine learning model." – P16-B

5.7.3. Data-centric explanations are more important in high stakes scenarios compared to low stakes

Participants discussed how the stakes and the importance of the systems impacted their perception of data-centric explanations. All participants wanted the explanations to be available when dealing with high-stakes scenarios, mentioning that these systems contribute to life-impacting decisions, with consequences of biased systems being more severe.

"I think that's very important to know [the explanation], especially in the higher stakes situation. Because like I say for bail or like whether you should convict someone or something like that, it can really affect someone's life whereas

> *recommendation, if they keep recommending me the wrong things, I'm annoyed, but my life isn't greatly affected." – P19-E*

> *"[The] Amazon recommendation where you bought such and such, it's such a simple thing [and] the result of following Amazon's recommendation doesn't hurt anybody except me and my wallet a little bit. The stakes are so low. Who cares, right? But in this case, it's about admitting a student in a university or not. You're affecting their future. Same with the criminals [in predictive bail]. You're affecting their future. So, yeah that's why [I would be more interested in the explanation for these two scenarios]." - P26-I*

Some participants (11/27) also mentioned that the importance of the system would impact how carefully or deeply they would look at the explanations.

> *"I would like to have the option [to have the explanation for every system]. […] For higher sensitive applications, I would definitely look at the [explanation] and read carefully." – P27-I*

For low-stakes situations (e.g., social media, ads, video recommendations), the majority of the participants (22/27) did not feel that the explanations were necessary, however, some participants (5/27) reported they would still like to have the explanation available, or at least a simplified version of it. These findings support results from prior work showing that explanations might not be valued for low-impact systems [8].

5.7.4. Data-centric explanations help participants' assess fairness but are not enough

Most of the participants (21/27), again regardless of expertise, mentioned that the data-centric explanations helped them judge the systems' fairness at least to some extent. Participants mentioned that knowing the diversity in the data from the demographics (Figure 2 : B), and whether there are any fatal flaws in the system from the error information (Figure 2 : D) were most helpful in this regard. The quotes below illustrate both a beginner and an expert perspective. While the expert quote uses more ML terminology, both speak to similar issues.

> *"Looking at like how much data they have, how many people they pull that information from and where they're from, and stuff to make sure it's diverse enough would help me know that's fair. And then even looking at the errors would help me know that's fair too." – P20-B*

> *"If I'm looking at the information you have provided in the explanations, I may doubt the fairness of the system. Because, in all the training data, the categories in them were not equal. For example, if gender is really important for the training set, I would like to have an equal number of males and females." – P2-E*

Some participants (6/27), on the other hand, indicated that the data-centric explanations were not sufficient to judge fairness. Three of these participants, all of whom were experts, wanted information on the decision process, including the factors affecting the system's decisions. The other 3 participants (1 beginner, 2 intermediates) did not have concrete ideas of what they thought was missing.

5.7.5. Many commonalities across expertise groups, but with nuanced differences

I found many similarities in how the different expertise groups responded to the data-centric explanations. Regardless of participants' ML training or technical background, the data suggest that the data-centric explanations impacted participants' trust in the machine learning systems described in the scenarios. Further, participants in all expertise categories reported that they could identify potential biases in the data from the demographic information presented in the explanations and all were eager to have the explanations available in higher stakes situations. Participants, again regardless of their expertise, felt that the explanations helped them to judge system fairness to at least some extent.

I did, however, see some nuanced differences in how participants' machine learning backgrounds impacted how they felt about the data-centric explanations. As I reported above, some expert participants wanted information on the decision factors in addition to the data-centric explanations to judge system fairness, whereas the non-expert participants did not have specific requests for additional information.

Interestingly, some experts (4/9) felt that the explanations would be more useful for those with machine learning expertise. The following quote illustrates this sentiment:

> *"I think if a user does not have any machine learning background or cs background, they will find it hard at first, [...] because they will not be clear about the training data, like what is called training data, how it is trained, [this]*

will go over their head. [So] it would be complicated at the beginning, but in the long run, they will adjust with it." – P8-E

No intermediates or beginners, however, expressed this concern. In fact, other than one participant who mentioned that the explanations were a little difficult to follow given that English was not their first language, all the participants reported the explanations to be easy to understand.

I did not observe any obvious differences in the study data between intermediate and beginner participants. One reason is likely related to my expertise definitions, where I distinguished between beginner and intermediate participants based on their Engineering / Computer Science background. I found, however, that some beginners were more knowledgeable about machine learning than intermediates based on workplace interactions or from the news media.

5.7.6. Amount of explanation content not overwhelming, but could be streamlined

Despite being rather lengthy, only one participant mentioned that the explanations contain too much information. However, many felt that the recommended usage (Figure 2 : C) and general information (Figure 2 : E) components had limited utility. Some participants (8/27) mentioned that the general information component is the least important because the information is too broad. Similarly, some participants (7/27) mentioned that the recommended usage component was also less important because it lacks enough detail to be helpful. Thus, while the explanation could be streamlined, participants did not find the quantity of information presented to be overwhelming.

5.8. Discussion

I now discuss the implication of the findings from the study, how they relate to the results from prior works on trust and fairness, and how expertise and prior experiences impact participants' attitudes to data-centric explanations.

5.8.1. Relation to findings from prior works on trust and fairness

The results from the study indicate that data-centric explanations have the potential to help people develop an informed sense of trust in machine learning systems. Participants' trust was impacted positively when the training seemed balanced and negatively when the explanations revealed problems. Like prior work, I found that participants cared most about the explanations for high-stakes system scenarios [8]. Future work should investigate other system traits that might impact explanation utility, such as system failures [32,37] and the stated accuracy of a system [102].

Participants indicated that the explanations also impacted their sense of system fairness, but to a lesser extent. They felt less confident in judging fairness without more information on the decision process. This indicates that data-centric explanations could serve as complements to established explanation approaches that explain the outcomes and the properties of a decision [20,30,88,89,92]. How users might prioritize data-centric explanations vs. feature-oriented explanations is an important area of future work. I also acknowledge that fairness is a social and ethical concept, and that perceptions of fairness are multi-dimensional and context-dependent [46,47,69]. While I measured fairness using widely used prior scales [5,55], a more comprehensive treatment of this construct is needed. Specific metrics for fairness that have been proposed in prior work [22,48] could serve as a useful starting point in this direction.

5.8.2. Potential mismatch on expectation of and capabilities of the end-users

Findings from this study suggest a potential mismatch between machine learning designers' expectations and end-users' interests and capabilities. Some participants with experience in building machine learning systems expressed concerns about the data-centric explanations being too complicated for end-users, yet I did not observe the non-experts having difficulty with the information. It would be interesting to explore the issue further. For example, are machine-learning practitioners underestimating the capabilities and interests of their target user populations? How do these preconceptions influence the information that machine learning practitioners are willing to release about the systems they create?

5.8.3. Impact of Expertise and prior experiences

For the non-expert participants, I observed individual differences with respect to existing positions on algorithmic decisions and machine learning systems. For example, a couple of participants expressed general distrust towards machine learning systems, while some other participants seemed to have inherent trust, feeling that computers are rarely wrong. I found these participants less receptive to the data-centric explanations, suggesting the potential for confirmation bias. This is in line with prior findings that users' individual prior positions on machine learning fairness and personal characteristics (e.g., locus of control [91], need for cognition [13], visual literacy [7]) can have a significant influence on their perceptions of explanations from the system [33,78].

5.9. Summary

In this chapter, I have described my second study where I investigated the utility of data-centric explanations in a range of automated decision scenarios. The study results

showed that participants have significantly more trust, are more confident about the systems' fairness, and have more comfort with the systems when the explanation showed that training data was relatively balanced. I also found nuanced differences among the expertise groups in their perception of the explanations.

Chapter 6

Conclusions

In this thesis, I presented data-centric explanations that focus on providing end-users with information on the training data of machine learning systems. I designed data-centric explanations using a user-centered process, gathering feedback on an initial prototype in a concept exploration study with 17 participants. I conducted a second study where I investigated the utility of the explanation across four automated system scenarios. I showed that data-centric explanations helped people to get insights into the systems, reflect on the training data, and influence their assessments of trust and fairness. My work is an important step forward in the general direction of aiming to bridge the gap between those who create machine learning systems and those affected by them.

6.1. Contributions

This thesis makes two contributions. I first contribute a design for data-centric explanations for machine learning systems that focus on communicating training dataset information to the end-users. My explanation approach is novel in that it considers the training data, rather than the features or individual decisions of machine learning systems.

I also contribute findings from two user studies. First, I contribute the findings from my concept exploration study where I investigated the feasibility of the idea of data-centric explanations and found positive attitudes to data-centric explanations. I also contribute the findings from this second study where I found that data-centric explanations make the systems more transparent to end-users, and can impact participants' trust and sense of fairness in the system. I further contribute by identifying the commonalities and the differences among people with different levels of machine learning expertise in their outlook to the data-centric explanations.

6.2. Limitations and Future Research Directions

This thesis is a step towards promoting transparency into machine learning systems by communicating training dataset information through explanations from machine learning systems. The findings from my second study show that data-centric explanations can support users' trust and fairness judgment of machine learning systems to some extent. Further, I found that user expertise and prior experiences of participants impact their outlook toward data-centric explanations. However, there are many

potential avenues for further exploration of data-centric explanations that could provide additional insights.

First, my second study's scenario-based approach, a commonly used method to study user perceptions of machine learning systems [5,47,74,96,105], allowed participants to reflect on a range of potential scenarios that were grounded in real-world examples. However, given that the scenarios were hypothetical and did not impact the participants personally, they likely lacked the consequences and the significance of real-world decisions. Further, since the explanations were not generated by already existing documentation from actual machine learning models, the explanations themselves might have lacked some degree of ecological validity. Prior work has suggested that simulating explanations can impact the generalizability of study findings [5]. Future work is needed to understand how users might respond to the explanations under conditions where they have more direct interactions with real systems and/or the systems' outputs. Future work should also explore the generalizability of my findings to a larger sample.

I found some initial insights on how participants' expertise and prior experiences impact their perception of data-centric explanations from machine learning systems. Future work should investigate ways to characterize these types of differences more systemically for data-centric explanations. Along these lines, future work should also explore ways to better characterize prior machine learning knowledge and experience. To help recruit a range of participants, I used a simple objective measure of technical background to include what I categorized as both novices and intermediates. While this approach did seem to help diversify the sample, I did not see clear differences between these two groups in their attitudes towards the explanations. I suspect that prior exposure

to machine learning concepts (e.g., from the media) might be a more informative distinguishing characteristic. Future work could, therefore, consider developing and using a more comprehensive pre-screening questionnaire.

The study scenario asked users to take on the role of the end-user of a machine learning system – somebody who would be directly interacting with the systems' output. Moving forward, it would be interesting to explore other potential audiences for these types of data-centric explanations. One potential audience could be journalists, who have often criticized machine learning systems for their black-box nature [68,98], and prior research has argued that journalists play a vital role in communicating information on algorithms to the general public [31]. It would also be interesting to explore the impact on those who make system acquisition decisions in companies or organizations, to see whether explanations on training data might influence their ultimate purchasing decisions.

Bibliography

1. Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue: Personalized Recommendation of Tourist Attractions. *APPLIED ARTIFICIAL INTELLIGENCE: Special Issue on Artificial Intelligence for Cultural Heritage and Digital Libraries* 17, 8–9: 687–714.

2. M. Arnold, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, R. K.E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. Natesan Ramamurthy, and A. Olteanu. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4–5. https://doi.org/10.1147/JRD.2019.2942288

3. Solon Barocas and Andrew Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3: 671. https://doi.org/10.15779/Z38BG31

4. Claudio Biancalana, Fabio Gasparetti, Alessandro Micarelli, Alfonso Miola, and Giuseppe Sansonetti. 2011. Context-aware movie recommendation based on signal processing and machine learning. *ACM International Conference Proceeding Series*: 5–10. https://doi.org/10.1145/2096112.2096114

5. Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's reducing a human being to a percentage"; perceptions of justice in algorithmic decisions. *Conference on Human Factors in Computing*

Systems - Proceedings 2018-April: 1–14. https://doi.org/10.1145/3173574.3173951

6. Tolga Bolukbasi, Kai Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4356–4364.

7. Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean Daniel Fekete. 2014. A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics* 20, 12: 1963–1972. https://doi.org/10.1109/TVCG.2014.2346984

8. Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. *International Conference on Intelligent User Interfaces, Proceedings IUI*: 169–178. https://doi.org/10.1145/2166966.2166996

9. Andrea Bunt, Joanna McGrenere, and Cristina Conati. 2007. Understanding the utility of rationale in a mixed-initiative system for GUI customization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4511 LNCS: 147–156. https://doi.org/10.1007/978-3-540-73078-1_18

10. Joy Buolamwini and Timnit Gebru. 2017. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classificatio. *Conference on Fairness, Accountability, and Transparency* 10: 1889–1896.

https://doi.org/10.2147/OTT.S126905

11. Jenna Burrell. 2015. How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms. *SSRN Electronic Journal*, June: 1–12. https://doi.org/10.2139/ssrn.2660674

12. Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. Explainable Machine Learning in Credit Risk Management. *Computational Economics*: 1–21. https://doi.org/10.1007/s10614-020-10042-0

13. John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment* 48, 3: 306–307. https://doi.org/10.1207/s15327752jpa4803_13

14. Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. *International Conference on Intelligent User Interfaces, Proceedings IUI* Part F1476: 258–262. https://doi.org/10.1145/3301275.3302289

15. Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. *Studies in Applied Philosophy, Epistemology and Rational Ethics* 3: 43–57. https://doi.org/10.1007/978-3-642-30487-3_3

16. Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* -

KDD '15: 1721–1730. https://doi.org/10.1145/2783258.2788613

17. Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5: 124–127. https://doi.org/10.1257/aer.p20161029

18. Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2016. Interpretable Deep Models for ICU Outcome Prediction. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2016: 371–380.

19. Lin Chen, Rui Li, Yige Liu, Ruixuan Zhang, and Diane Myung Kyung Woodbridge. 2018. Machine learning-based product recommendation using Apache Spark. *2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017* - : 1–6. https://doi.org/10.1109/UIC-ATC.2017.8397470

20. Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Conference on Human Factors in Computing Systems - Proceedings*: 1–12. https://doi.org/10.1145/3290605.3300789

21. Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen Tau Yih, Yejin Choi, Percy

Liang, and Luke Zettlemoyer. 2020. QUAC: Question answering in context. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*: 2174–2184. https://doi.org/10.18653/v1/d18-1241

22. Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. 1–13. Retrieved from http://arxiv.org/abs/1810.08810

23. Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxceleB2: Deep speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2018-Septe, ii: 1086–1090. https://doi.org/10.21437/Interspeech.2018-1929

24. Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Washington Law Review* 89, 1: 1–33.

25. Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks . It' s actually not that clear. *The Washington Post*: 1–7.

26. Juliet Corbin and Anselm Strauss. 2008. Strategies for qualitative data analysis. *Basics of Qualitative Research. Techniques and procedures for developing grounded theory* 3.

27. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. In *User Modeling and User-Adapted Interaction*, 455–496.

https://doi.org/10.1007/s11257-008-9051-3

28. Russell S. Cropanzano, Maureen L. Ambrose, Jason A. Colquitt, and Jessica B. Rodell. 2015. Measuring Justice and Fairness. *The Oxford Handbook of Justice in the Workplace*: 187–202. https://doi.org/10.1093/oxfordhb/9780199981410.013.8

29. Anupam Datta. 2017. Did Artificial Intelligence Deny You Credit? *The Conversation*. Retrieved January 20, 2019 from http://theconversation.com/did-artificial-intelligence-deny-you-credit-73259

30. Anupam Datta, Shayak Sen, and Yair Zick. 2017. Algorithmic Transparency via Quantitative Input Influence. 71–94. https://doi.org/10.1007/978-3-319-54024-5_4

31. Nicholas Diakopoulos. 2015. Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3, 3: 398–415. https://doi.org/10.1080/21670811.2014.976411

32. Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1: 114–126. https://doi.org/10.1037/xge0000033

33. Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K.E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. *International Conference on Intelligent User Interfaces, Proceedings IUI*: 275–285. https://doi.org/10.1145/3301275.3302310

34. Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2018. Explaining

recommendations by means of user reviews. In *CEUR Workshop Proceedings*.

35. Donal Doyle, Alexey Tsymbal, and Pádraig Cunningham. 2003. A Review of Explanation and Explanation in Case-Based Reasoning. *Dublin, Trinity College Dublin, Department of Computer Science, TCD-CS-2003-41*: 41.

36. Mengnan Du, Ninghao Liu, and Xia Hu. 2020. Techniques for interpretable machine learning. *Communications of the ACM* 63, 1: 68–77. https://doi.org/10.1145/3359786

37. Mary T. Dzindolet, Linda G. Pierce, Hall P. Beck, and Lloyd A. Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1: 79–94. https://doi.org/10.1518/0018720024494856

38. Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. *International Conference on Intelligent User Interfaces, Proceedings IUI*: 211–223. https://doi.org/10.1145/3172944.3172961

39. Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Jacques, Meysam Madadi, Xavier Baro, Stephane Ayache, Evelyne Viegas, Yagmur Gucluturk, Umut Guclu, Marcel A.J. Van Gerven, and Rob Van Lier. 2017. Design of an explainable machine learning challenge for video interviews. *Proceedings of the International Joint Conference on Neural Networks* 2017-May: 3688–3695. https://doi.org/10.1109/IJCNN.2017.7966320

40. Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen

M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639: 115–118. https://doi.org/10.1038/nature21056

41. Ea Eyjolfsdottir, Gaurangi Tilak, and Nan Li. 2010. MovieGEN: A Movie Recommendation System. *Computer Science Department, ….* Retrieved from http://www.cs.ucsb.edu/~nanli/projects/CS265-MovieGEN.pdf

42. Gerald Fahner. 2018. Developing Transparent Credit Risk Scorecards More Effectively : An Explainable Artificial Intelligence Approach. c: 7–14. Retrieved from https://www.thinkmind.org/index.php?view=article&articleid=data_analytics_2018_1_30_60077

43. Alex Fefegha. 2019. Racial Bias and Gender Bias Examples in AI systems So here it goes : Racial Bias. 1–14. Retrieved January 17, 2019 from https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1

44. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for Datasets. Retrieved from http://arxiv.org/abs/1803.09010

45. Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable Credit Application Predictions With Counterfactual Explanations. 1–9. Retrieved from

http://arxiv.org/abs/1811.05245

46. Ben Green and Lily Hu. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. *the ICML 2018 Debates Workshop*.

47. Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*: 903–912. https://doi.org/10.1145/3178876.3186138

48. Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, Nips: 3323–3331.

49. J. L. Herlocker, J. A. Konstan, and J. Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 241–250. https://doi.org/10.1145/358916.358995

50. Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*: 1–13. https://doi.org/10.1145/3290605.3300809

51. Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miroslav Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do

industry practitioners need? *Conference on Human Factors in Computing Systems - Proceedings*: 1–16. https://doi.org/10.1145/3290605.3300830

52. Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Müller, Robert Reihs, and Kurt Zatloukal. 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. 1–34. Retrieved from http://arxiv.org/abs/1712.06657

53. Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*.

54. InVisionApp Inc. 2020. InVision | Digital product design, workflow & collaboration. Retrieved October 1, 2019 from https://www.invisionapp.com/

55. J. A.Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of applied psychology 68*, 386–399.

56. Jiun-Yin Jian, Ann M Bisantz, Colin G Drury, and James Llinas. 1996. United States Air Force Research Laboratory Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1: 53–71.

57. Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. 2012. Consumer Credit Risk Models Via Machine-Learning Algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1568864

58. Lauren Kirchner, Surya Mattu, Jeff Larson, and Julia Angwin. 2016. Machine Bias.

Propublica 23: 1–26. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

59. Rene F. Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. *Conference on Human Factors in Computing Systems - Proceedings*: 2390–2395. https://doi.org/10.1145/2858036.2858402

60. Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300641

61. Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*: 5686–5697. https://doi.org/10.1145/2858036.2858529

62. Todd Kulesza, Margaret Burnett, Weng Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to personalize interactive machine learning. *International Conference on Intelligent User Interfaces, Proceedings IUI* 2015-Janua: 126–137. https://doi.org/10.1145/2678025.2701399

63. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. *Conference on Human Factors in Computing Systems - Proceedings*: 1–10. https://doi.org/10.1145/2207676.2207678

64. Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*: 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

65. Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. *Conference on Human Factors in Computing Systems - Proceedings*: 1–12. https://doi.org/10.1145/3290605.3300717

66. Paul B. de Laat. 2018. Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy and Technology* 31, 4: 525–541. https://doi.org/10.1007/s13347-017-0293-z

67. Anísio Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani, and Berthier Ribeiro-Neto. 2006. Learning to advertise. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 549–556. https://doi.org/10.1145/1148170.1148265

68. Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2020. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

69. Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data and Society* 5, 1: 1–16. https://doi.org/10.1177/2053951718756684

70. Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4: 611–627. https://doi.org/10.1007/s13347-017-0279-x

71. Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics* 9, 3: 1350–1371. https://doi.org/10.1214/15-AOAS848

72. Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376590

73. C C S Liem, Markus Langer, Andrew Demetriou, Annemarie M F Hiemstra, A S Wicaksana, M Ph. Born, and Cornelius J. König. 2018. *Explainable and interpretable models in computer vision and machine learning*. https://doi.org/10.1007/978-3-319-98131-4

74. Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. *UbiComp 2009: Ubiquitous Computing*: 195. https://doi.org/10.1145/1620545.1620576

75. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 2119. https://doi.org/10.1145/1518701.1519023

76. Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3: 31–57. https://doi.org/10.1145/3236386.3241340

77. Gideon Mann and Cathy O'Neil. 2016. Hiring Algorithms Are Not Neutral. *Harvard Business Review*. Retrieved August 4, 2020 from https://hbr.org/2016/12/hiring-algorithms-are-not-neutral

78. Martijn Millecamp, Cristina Conati, Nyi Nyi Htun, and Katrien Verbert. 2019. To explain or not to explain: The effects of personal characteristics when explaining music recommendations. *International Conference on Intelligent User Interfaces, Proceedings IUI*: 397–407. https://doi.org/10.1145/3301275.3302313

79. Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Figure 2: 220–229. https://doi.org/10.1145/3287560.3287596

80. Conor Nugent and Pádraig Cunningham. 2005. A case-based explanation system for black-box systems. *Artificial Intelligence Review* 24, 2: 163–178. https://doi.org/10.1007/s10462-005-4609-5

81. Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2017. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *SSRN Electronic Journal*: 1–47. https://doi.org/10.2139/ssrn.2886526

82. Frank Pasquale. 2015. *The Black Box Society*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

83. Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Luca Pappalardo, Salvatore Ruggieri, and Franco Turini. 2018. Open the Black Box Data-Driven Explanation of Black Box Decision Systems. 1, 1: 1–15. Retrieved from http://arxiv.org/abs/1806.09936

84. Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. Retrieved from http://arxiv.org/abs/1802.07810

85. Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. *International Conference on Intelligent User Interfaces, Proceedings IUI* 2006: 93–100. https://doi.org/10.1145/1111449.1111475

86. Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*: 1–13. https://doi.org/10.1145/3173574.3173677

87. Ashwin Ram. 1993. AQUA: Questions that Drive the Explanation Process. *Georgia Institute of Technology*.

88. Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 97–101. https://doi.org/10.18653/v1/n16-3020

89. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. https://doi.org/10.1145/2858036.2858529

90. Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *IJCAI International Joint Conference on Artificial Intelligence* 0: 2662–2670. https://doi.org/10.24963/ijcai.2017/371

91. J. B. Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs* 80, 1: 1–28. https://doi.org/10.1037/h0092976

92. Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus Robert Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11: 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

93. Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. 2018. Baselines and a datasheet for the Cerema AWP dataset. *arXiv preprint arXiv:1806.04016*. Retrieved from http://arxiv.org/abs/1806.04016

94. Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff,

Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": Supporting clinical decision-making with deep learning. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 99–109. https://doi.org/10.1145/3351095.3372827

95. Eduardo Soares and Plamen Angelov. 2019. Fair-by-design explainable models for prediction of recidivism. 3–7. Retrieved from http://arxiv.org/abs/1910.02043

96. Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. Human perception of fairness: A descriptive approach to fairness for machine learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 2459–2468. https://doi.org/10.1145/3292500.3330664

97. Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting black-box classifiers using instance-level visual explanations. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA 2017*: 1–6. https://doi.org/10.1145/3077257.3077260

98. Caroline Wang, Bin Han, Bhrij Patel, Feroze Mohideen, and Cynthia Rudin. 2020. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. 1–58. Retrieved from http://arxiv.org/abs/2005.04176

99. Yuanyuan Wang, Stephen Chi Fai Chan, and Grace Ngai. 2012. Applicability of demographic recommender system to tourist attractions: A case study on TripAdvisor. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2012*: 97–101.

https://doi.org/10.1109/WI-IAT.2012.133

100. Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2020. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*: 1358–1368. https://doi.org/10.18653/v1/d18-1166

101. L. Richard Ye and Paul E. Johnson. 1995. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly: Management Information Systems* 19, 2: 157–172. https://doi.org/10.2307/249686

102. Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*: 1–12. https://doi.org/10.1145/3290605.3300509

103. Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 180, 3: 689–722. https://doi.org/10.1111/rssa.12227

104. Yong Zhang, Hongming Zhou, Nganmeng Tan, Saeed Bagheri, and Meng Joo Er. 2017. Targeted Advertising Based on Browsing History. Retrieved from http://arxiv.org/abs/1711.04498

105. Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. 2020. Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and*

Transparency: 295–305. https://doi.org/10.1145/3351095.3372852

106. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*: 2979–2989. https://doi.org/10.18653/v1/d17-1323

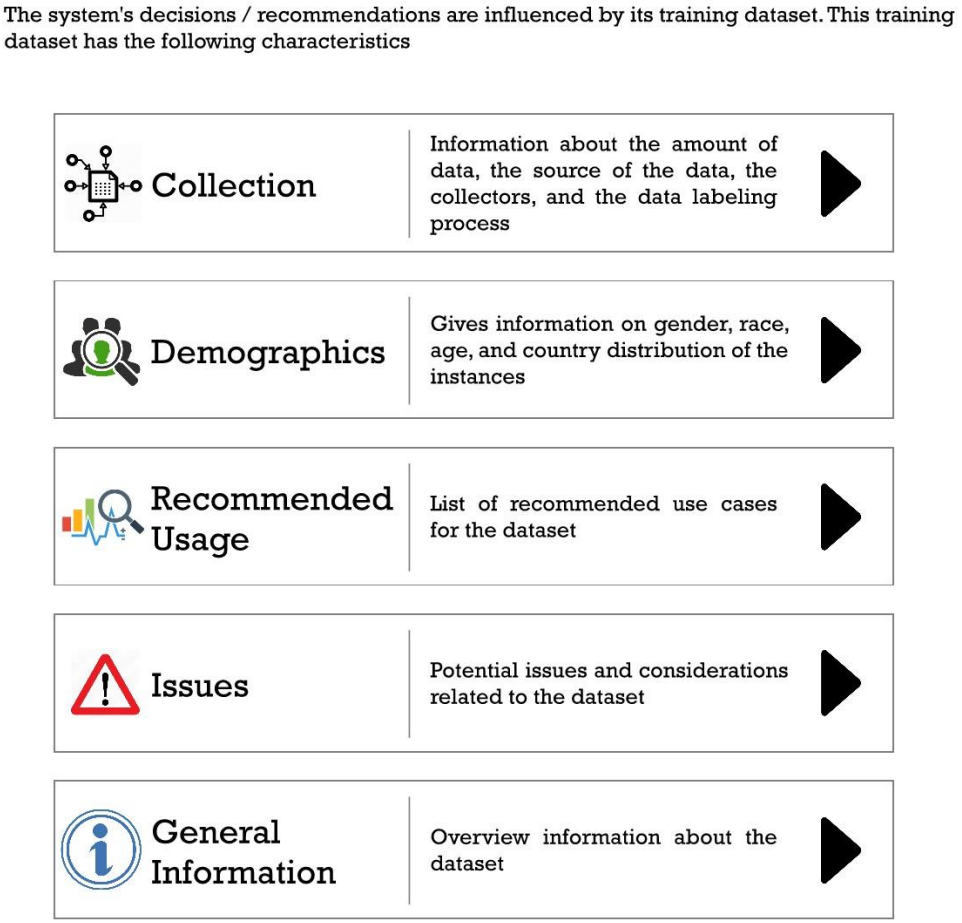
107. Graduate Admission 2. Retrieved September 2, 2020 from https://kaggle.com/mohansacharya/graduate-admissions

108. Qualtrics. *Qualtrics*. Retrieved August 5, 2020 from https://www.qualtrics.com/core-xm/survey-software/

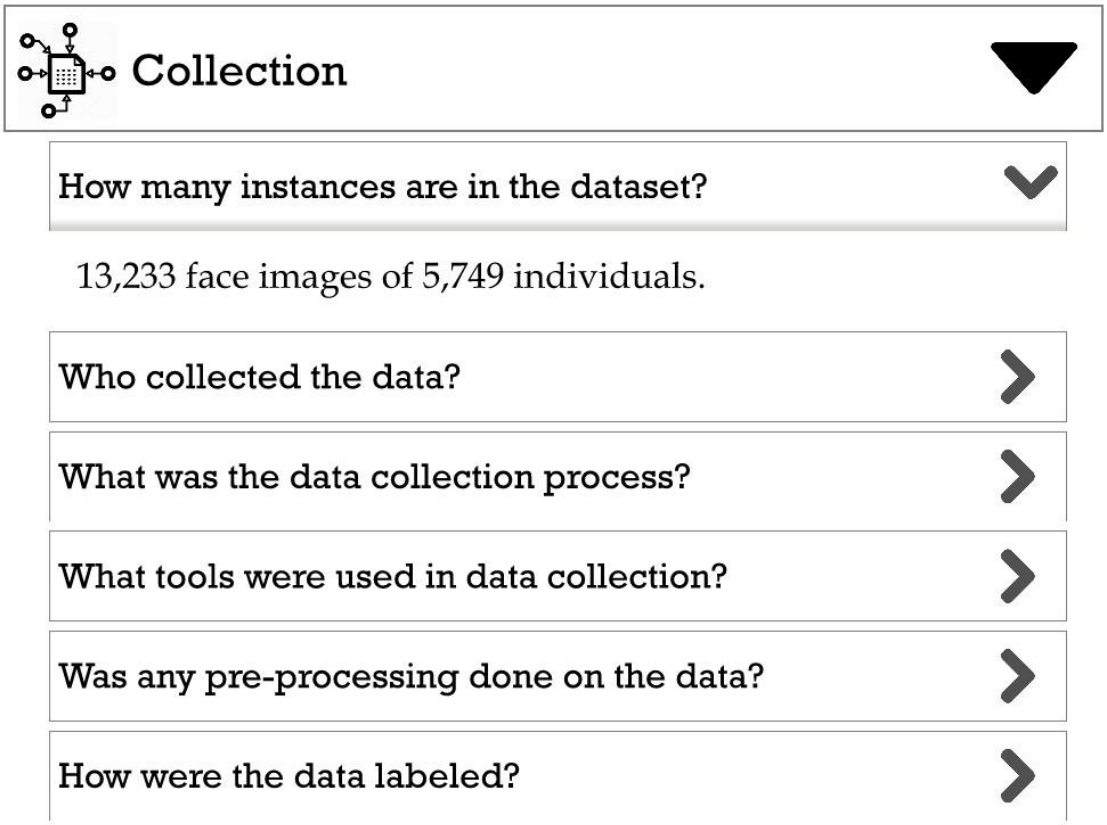
Appendix A

Initial Prototype of Data-centric Explanations

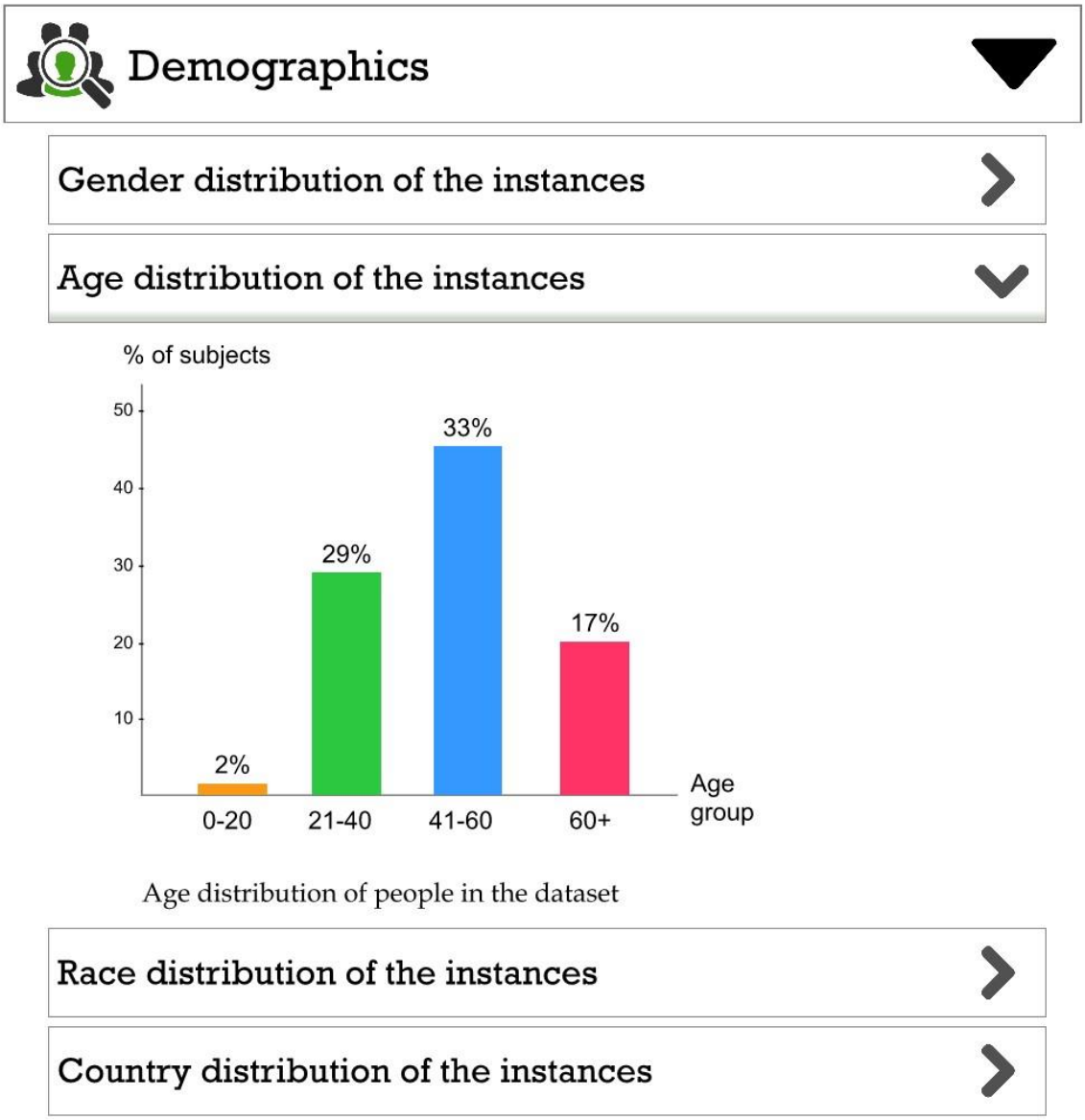
Overview of the prototype



Collection Information

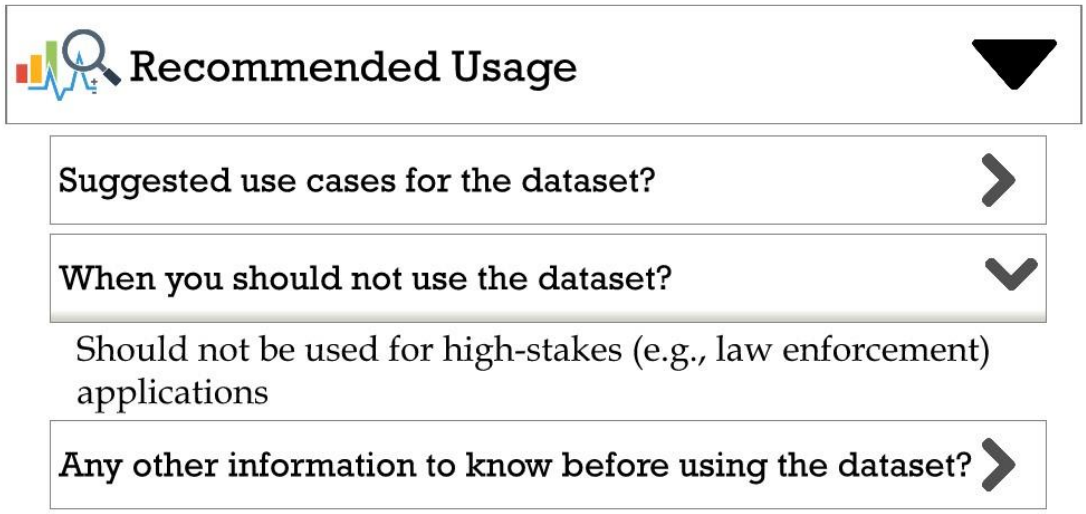


Demographics

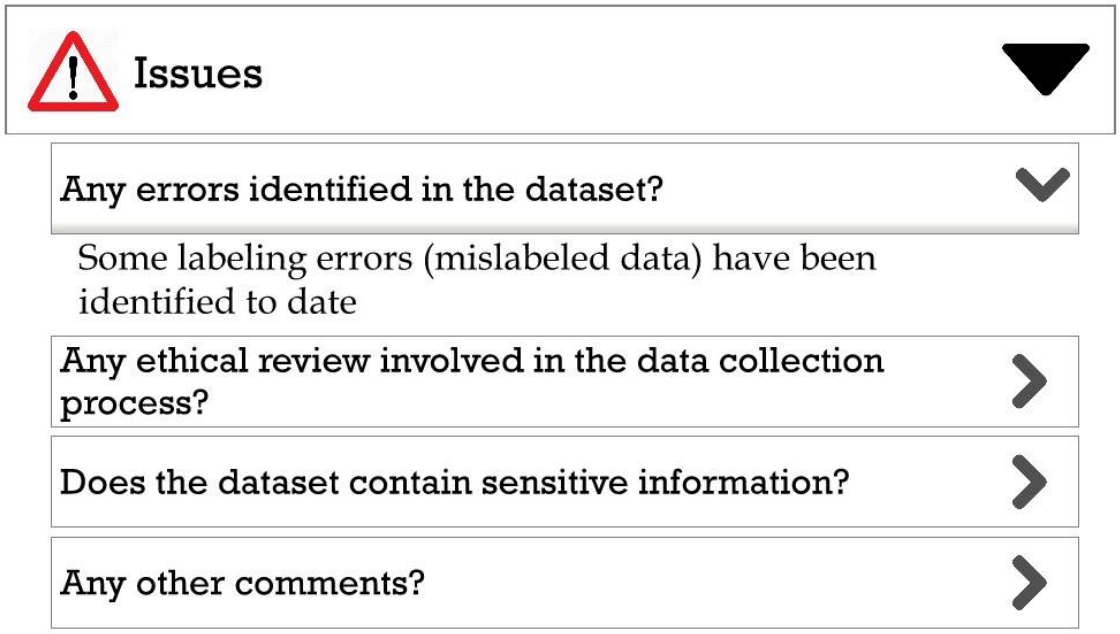


Age distribution of people in the dataset

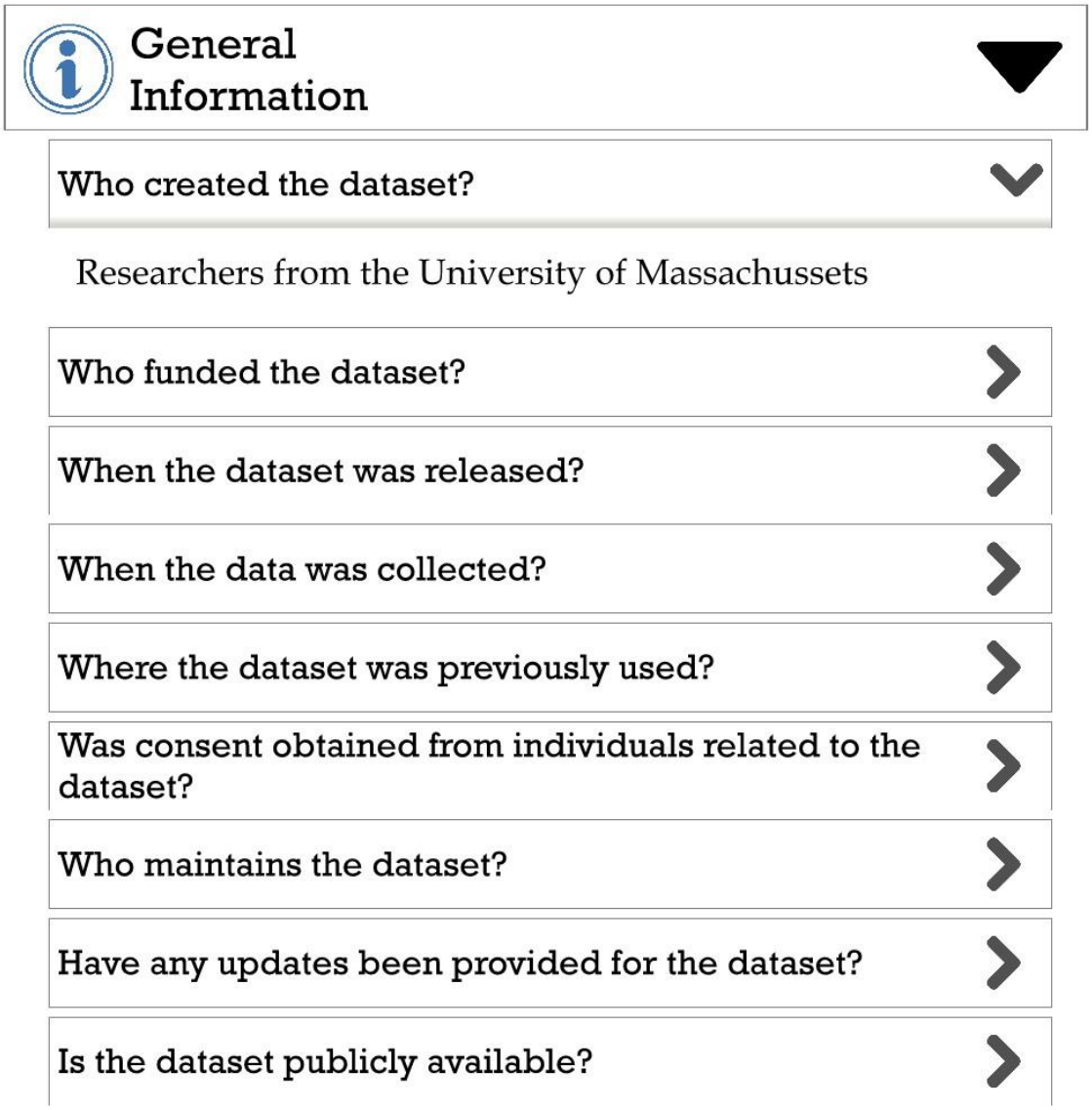
Recommended Usage



Issues Information



General Information



Appendix B

Poster Advertising the Study

University of Manitoba

HCI Lab
UNIVERSITY OF MANITOBA

Do you have prior experience of interacting with machine learning systems?

Input Decision Making Systems Output

We are looking for people to participate in a study to share their opinions on ways a machine learning system could explain how it was trained. The study will take approximately 60-90 minutes of your time and you will get $20 cash or gift card (at your choosing).

If you are more than 18 years old and you are interested in participating in our study, please contact -

Md Ariful Islam Anik
aianik@cs.umanitoba.ca

Research approved by University of Manitoba Joint Faculty Research Ethics Board. The Research Ethics Board can be reached by phone (204) 474-6791 or email humanethics@umanitoba.ca

Appendix C

Research Ethics Board Approval

UNIVERSITY OF MANITOBA | Research Ethics and Compliance

Human Ethics
208-194 Dafoe Road
Winnipeg, MB
Canada R3T 2N2
Phone +204-474-7122
Email: humanethics@umanitoba.ca

PROTOCOL APPROVAL

TO: **Andrea Bunt**
Principal Investigator

FROM: **Julia Witt, Chair**
Joint-Faculty Research Ethics Board (JFREB)

Re: **"Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems"**

Effective: November 13, 2019 **Expiry:** November 13, 2020

Joint-Faculty Research Ethics Board (JFREB) has reviewed and approved the above research. JFREB is constituted and operates in accordance with the current *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.*

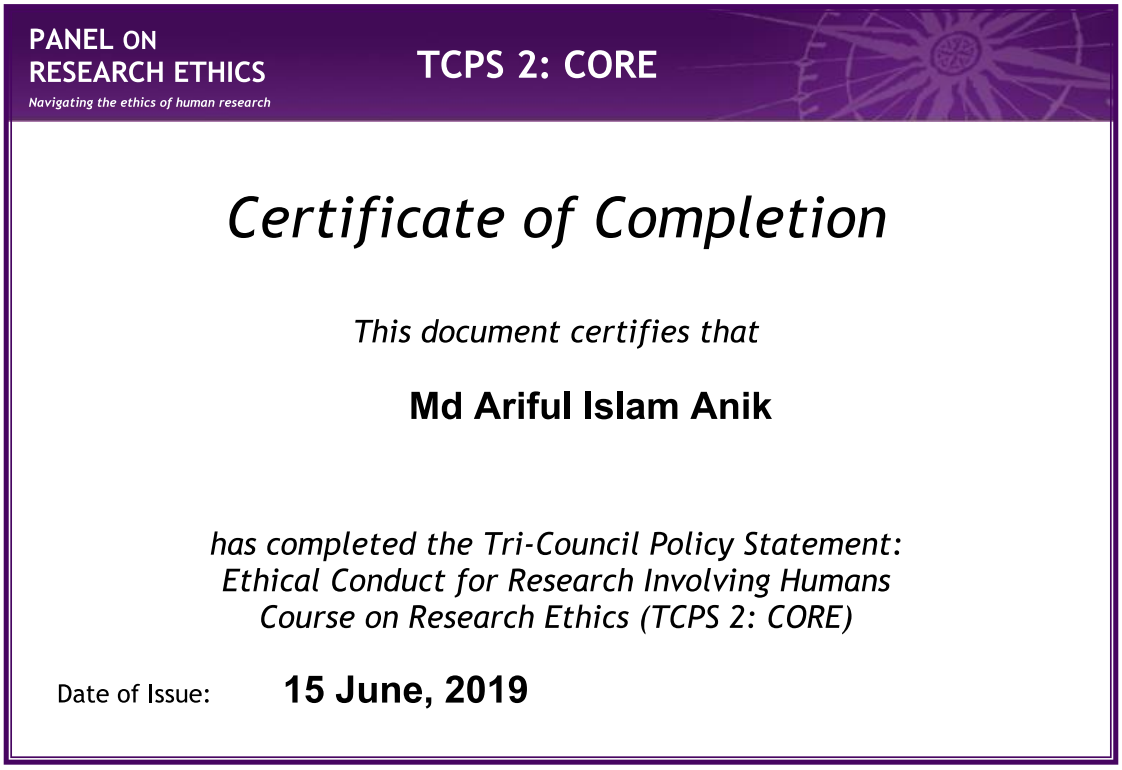
This approval is subject to the following conditions:

1. Approval is granted for the research and purposes described in the application only.
2. Any modification to the research or research materials must be submitted to JFREB for approval before implementation.
3. Any deviations to the research or adverse events must be submitted to JFREB as soon as possible.
4. This approval is valid for one year only and a Renewal Request must be submitted and approved by the above expiry date.
5. A Study Closure form must be submitted to JFREB when the research is complete or terminated.
6. The University of Manitoba may request to review research documentation from this project to demonstrate compliance with this approved protocol and the University of Manitoba *Ethics of Research Involving Humans*.

Funded Protocols:
- **Please e-mail a copy of this Approval, identifying the related UM Project Number, to the Research Grants Officer at researchgrants@umanitoba.ca**

Appendix D

TCPS 2: CORE Certificate



PANEL ON RESEARCH ETHICS
Navigating the ethics of human research

TCPS 2: CORE

Certificate of Completion

This document certifies that

Md Ariful Islam Anik

has completed the Tri-Council Policy Statement:
Ethical Conduct for Research Involving Humans
Course on Research Ethics (TCPS 2: CORE)

Date of Issue: **15 June, 2019**

Appendix E

Consent Form for the First Study

UNIVERSITY OF MANITOBA

DEPARTMENT OF COMPUTER SCIENCE

Winnipeg, Manitoba
Canada R3T 2N2
(204) 474-8313
FAX: (204) 474-7609

Research Project Title: Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems

Researchers:

Dr. Andrea Bunt, Associate Professor, Department of Computer Science, University of Manitoba,

Md Ariful Islam Anik, Graduate Research Assistant, Department of Computer Science, University of Manitoba,

Research Sponsored by Natural Sciences and Engineering Research Council of Canada

Please take the time to read this carefully and to ensure you understand all the information.

This consent form, a copy of which will be left with you for your records and reference, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, you should feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

You are invited to participate in a research study on the topic of evaluating the impact of explanations on users' fairness judgment of data-driven decisions in machine learning systems. The goal is to explore whether end-users find it valuable to have explanations on the way that machine learning systems are trained, and what specific information those explanations should contain. Your participation in this research study will involve participating in an interview, as well as looking at some example explanations, and providing feedback on them. If you have any questions or concerns at this time or any time during the project, please feel free to ask the researcher for clarification.

The benefits of participating in this research are that you may gain a greater understanding of how machine learning works. Further, the knowledge you will contribute to this study will be used to improve the way that machine learning systems explain their decisions. The risks of this study are no greater than in everyday life.

Participation in this study is voluntary and will take approximately 60-90 minutes of your time. You will receive a $20 compensation in the form of cash or a gift card (at your choosing) after signing the consent form. You are free to withdraw from the interview any time up to the end of the interview and/or refrain from answering any questions you prefer to omit. Even by withdrawing, you will keep your compensation.

We wish to record our discussions with you by using a hand-held digital audio recorder. The audio recording will assist our data analysis by allowing us to review the discussion and the study session in detail. Any information you choose to contribute in our discussion is completely confidential and will be used for anonymized research analysis. We may use anonymized quotes

for purposes of dissemination. Your name will not be included or in any other way associated with the data presented in the result of this study. By signing this consent form, you agree that you understand this and that we may use the recorded audio for data analysis purposes only. Unfortunately, if you do not wish to be audio recorded then you cannot participate in this study.

Data collected during this study will be retained for a maximum period of three years in a locked cabinet or in a password-protected computer in a locked office or laboratory in the EITC building at the University of Manitoba, to which only researchers associated with this project (Dr. Andrea Bunt, Md Ariful Islam Anik) have access. The data will be destroyed by December 2022. In addition, the University of Manitoba may look at research records to see that the research is being done in a safe and proper way. We intend to present results as academic publications and a thesis which will be published in MSpace. Once published, the results of the study will be made available to the public for free at http://hci.cs.umanitoba.ca/. Again, no personal information about you will be included.

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research study. By doing this you also confirm that you are of the age of majority in Canada (18 years or more). In no way does this waive your legal rights nor release the researchers, sponsors, or involved institutions from their legal and professional responsibilities.

This research has been approved by the University of Manitoba Joint Faculty Research Ethics Board. If you have any concerns or complaints about this project, please contact Dr. Andrea Bunt at or the Human Ethics Coordinator at A copy of this consent form has been given to you to keep for your records and reference.

☐ I wish to receive a summary of the findings.

☐ I wish to receive a copy of the interview transcript of the audio recording to confirm its accuracy.

Please write your email address if you checked the box above:

Participant's email address: _____

Participant's Signature: _____ Date _____

Researcher's Signature _____ Date _____

Appendix F

Updated Prototype of Data-centric Explanations

Overview of the prototype

The system's decisions/recommendations are influenced by its training dataset and how that is used to train the system. The training dataset for this predictive bail decision system has the following characteristics.

Collection | **Information about the amount of data, the source of the data, the collectors, and the labeling process**

Demographics | **Information on gender, race, age, and country distribution of the instances**

Recommended Usage | **List of recommended use cases for the dataset**

Potential Issues | **Potential issues and considerations related to the dataset**

General Information | **Overview information about the dataset**

Collection

Collection

Information about the amount of data, the source of the data, the collectors, and the labeling process

How many instances are in the dataset?

The dataset contains reoffense scores for 11,757 criminal defendants containing around 80,000 criminal records.

Who collected the data?

What was the data collection process?

What tools were used in data collection?

Was any pre-processing done on the data?

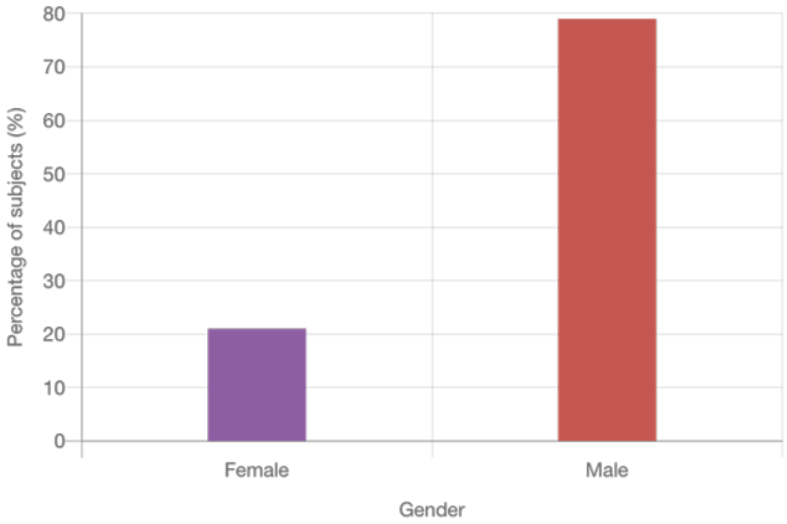
How were the data labeled?

Any other comments?

Demographics

Demographics | **Information on gender, race, age, and country distribution of the instances**

Gender distribution of the instances?



Gender distribution of people in the dataset

Age distribution of the instances?

Race distribution of the instances?

Country distribution of the instances?

Recommended Usage

Recommended Usage | **List of recommended use cases for the dataset**

Suggested use cases for the dataset?

The dataset was suggested to be used in determining re-offense risk of individuals in the USA (preferably in Florida).

Where you should not use the dataset?

Any other information to know before using the dataset?

Potential Issues

Potential Issues | **Potential issues and considerations related to the dataset**

Any errors identified in the dataset?

Sometimes people's names or dates of birth were incorrectly entered in some records – which led to incorrect matches between an individual's re-offense score and his or her criminal records. In a random sample of 400 cases, there was an error rate of 3.75%.

Any ethical review involved in the data collection process?

Does the dataset contain sensitive information?

Any other comments?

General Information

General Information | **Overview information about the dataset**

When the dataset was released?

When was the data collected?

The dataset contains criminal record data for 2013 and 2014 which were collected in April 2016.

Where the dataset was previously used?

Was consent obtained from individuals related to the dataset?

Have any updates been provided for the dataset?

Is the dataset publicly available?

Any other comments?

Appendix G

Research Ethics Approval for the Second Study

University of Manitoba | Research Ethics and Compliance

Human Ethics - Fort Garry
208-194 Dafoe Road
Winnipeg, MB R3T 2N2
T: 204 474 8872
humanethics@umanitoba.ca

AMENDMENT APPROVAL

April 27, 2020

TO: **Andrea Bunt**
Principal Investigator

FROM: **Julia Witt, Chair**
Joint-Faculty Research Ethics Board (JFREB)

Re: **Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems**

Joint-Faculty Research Ethics Board (JFREB) has reviewed and approved your Amendment Request received on **April 22, 2020** to the above-noted protocol. JFREB is constituted and operates in accordance with the current *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.*

This approval is subject to the following conditions:

1. Approval is given for this amendment only. Any further changes to the protocol must be reported to the Human Ethics Coordinator in advance of implementation.
2. Any deviations to the research or adverse events must be submitted to JFREB as soon as possible.
3. Amendment Approvals do not change the protocol expiry date. Please refer to the original Protocol Approval or subsequent Renewal Approvals for the protocol expiry date.

Appendix H

Automated System Scenarios

Scenario	*Scenario Description*
Predictive Bail Decision	The law enforcement department of your area decided to automate the process of assessing a defendant's re-offense risk (how likely they are to commit a crime again) to help the judge make bail decisions. The automated system uses machine learning algorithms to calculate the re-offense risk for a defendant and recommends whether or not to make a positive bail decision. The system was trained using the data of previous bail decisions that were made manually by judges. It uses the training data to calculate the re-offense risk for new defendants and recommend bail decisions. Suppose you are appointed as a judge and your role is to provide the final bail decision based on the recommendation from the system. The system provides you this explanation with every recommendation.
Facial Expression Recognition	To provide better accessibility, a company developed an automated face recognition system where the system recognizes the facial expression of the person from a given image. The system is trained on a publicly available dataset of images with faces. When given a new image, the system uses computer vision algorithms to identify the person's facial expression based on the training dataset. Suppose you are an end-user of the system and you are looking at the captions for some of the results determined by the tool. The system provides you the following explanation for the system-generated results.

Automated Admission Decision	The University of X decided to use an automated system to help the administrator to give admission decisions to new graduate applicants. Their system assesses applicants based on a machine learning model that profiles students based on the application materials and gives a recommendation for each of the applicants. The system is trained on the data of the previous admission decisions made for applicants to the university. It uses the training data to recommend the admission decision to current applicants. Suppose you are an administrator at the university and your role is to provide the final decision on admission based on the recommendation provided by the system. The system provides you this explanation with every recommendation.
Automated Speech Recognition	A company that hosts media content developed an automated speech recognition tool that can recognize the identities of individuals from their speech. The system is trained on an audio-visual dataset that consists of short clips of human speech extracted from different videos. When the system is given a new clip of human-speech, it uses machine learning algorithms to recognize identities from voice based on the training dataset. Suppose you are an end-user of the system and you are looking at the identities recognized by the automated tool. The system provides you the following explanation for every recognition.

Appendix I

Sample Information Presented in Explanations (for Predictive Bail Decisions)

Categories	Questions	Predictive Bail Decisions
Collection: *Information about the amount of data, the source of the data, the collectors, and the labeling process*	How many instances are in the dataset?	The dataset contains reoffense scores for 11,757 criminal defendants containing around 80,000 criminal records.
	Who collected the data?	Data were collected by Propublica (an independent, non-profit newsroom that produces investigative journalism) through a public request to the sheriff's office.
	What was the data collection process?	Data about the defendants were obtained from the Boward County Sheriff's Office in Florida through a public request. Criminal records were collected from the Boward County Clerk's office website in April 2016.
	What tools were used in data collection?	Data were collected manually.
	Was any pre-processing done on the data?	Data instances were discarded if defendants were assessed at parole, probation or other stages in the criminal justice system. Only data for people who

		were assessed at the pretrial stage were kept in the dataset.
	How were the data labeled?	Each of the defendants was labeled according to the risk of reoffending. The score for each defendant ranged from 1 to 10, with ten being the highest risk. Scores 1 to 4 were labeled as "Low"; 5 to 7 were labeled "Medium"; and 8 to 10 were labeled "High."
	Any other comments?	The dataset uses the same race classifications used by the Broward County Sheriff's Office, where they identified defendants as black, white, Hispanic, Asian and Native American.
Demographics: *Information on gender, race, age, and country distribution of the instances*	Gender distribution of the instances?	79% Male (9336) 21% Female (2421)
	Age distribution of the instances?	Less than 25: 20.76% 25 to 45: 56.55% More than 45: 22.69%
	Race distribution of the instances?	African-American: 49.54% Asian: 0.5% Caucasian: 34.74% Hispanic: 9.35% Native American: 0.34% Other: 5.62%

	Country distribution of the instances?	No distribution is given for country in the dataset since it was collected and primarily used in USA.
Recommended Usage: *List of recommended use cases for the dataset*	Suggested use cases for the dataset?	The dataset was suggested to be used in determining re-offense risk of individuals in the USA (preferably in Florida).
	Where you should not use the dataset?	The dataset should not be used in any other purpose other than calculating re-offense risks.
	Any other information to know before using the dataset?	N/A
Potential Issues: *Potential issues and considerations related to the dataset*	Any errors identified in the dataset?	Sometimes people's names or dates of birth were incorrectly entered in some records – which led to incorrect matches between an individual's re-offense score and his or her criminal records. In a random sample of 400 cases, there was an error rate of 3.75%.
	Any ethical review involved in the data collection process?	Unknown. However, the data were fetched with permission from the Sheriff's office.
	Does the dataset contain sensitive information?	Yes. The dataset contains criminal records of individuals.
	Any other comments?	No.
General Information: *Overview*	When the dataset was released?	2016.
	When was the data collected?	The dataset contains criminal record data for 2013 and 2014 which were collected in April 2016.

information about the dataset	Where the dataset was previously used?	The dataset was previously used only in research purposes.
	Was consent obtained from individuals related to the dataset?	No, individual consents were not taken. However, the data were taken with permission from the sheriff's office.
	Have any updates been provided for the dataset?	No.
	Is the dataset publicly available?	Yes. The dataset can be found here *(link removed)*
	Any other comments?	No.

Appendix J

Consent Form for the Second Study

UNIVERSITY OF MANITOBA

DEPARTMENT OF COMPUTER SCIENCE

Winnipeg, Manitoba
Canada R3T 2N2
(204) 474-8313
FAX: (204) 474-7609

Research Project Title: Explanations in Support of Fairness Judgments of Data-Driven Decisions in Machine Learning Systems

Researchers:

Dr. Andrea Bunt, Associate Professor, Department of Computer Science, University of Manitoba,

Md Ariful Islam Anik, Graduate Research Assistant, Department of Computer Science, University of Manitoba,

Research Sponsored by Natural Sciences and Engineering Research Council of Canada

Please take the time to read this carefully and to ensure you understand all the information.

This consent form, a copy of which will be left with you for your records and reference, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, you should feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

You are invited to participate in a research study on the topic of evaluating the impact of explanations on users' fairness judgment of data-driven decisions in machine learning systems. The goal is to explore whether end-users find it valuable to have explanations on the way that machine learning systems are trained, and what specific information those explanations should contain. Your participation in this research study will involve interacting with our explanations in four given scenarios presented in an online study platform (Qualtrics) and providing feedback on them. We will ask you to complete a short questionnaire after each scenario and ask you to participate in an interview. If you have any questions or concerns at this time or any time during the project, please feel free to ask the researcher for clarification.

The benefits of participating in this research are that you may gain a greater understanding of how machine learning systems are trained. Further, the knowledge you will contribute to this study will be used to improve explanations from machine learning systems. The risks of this study are no greater than in everyday life.

Participation in this study is voluntary and will take approximately 60-90 minutes of your time. You will receive a $20 (Canadian dollars) compensation in the form of cash or a gift card (at your choosing) after signing the consent form. You are free to withdraw from the study any time up to the end of the study and/or refrain from answering any questions you prefer to omit. Even by withdrawing, you will keep your compensation.

We wish to record our discussions with you by using a digital audio recorder. The audio recording will assist our data analysis by allowing us to review the discussion and the study session in detail. Any information you choose to contribute in our discussion is completely confidential and will be used for anonymized research analysis. We may use anonymized quotes for purposes of dissemination. Your name will not be included or in any other way associated with the data presented in the result of this study. By signing this consent form, you agree that you understand this and that we may use the recorded audio for data analysis purposes only. Unfortunately, if you do not wish to be audio recorded then you cannot participate in this study.

Data collected during this study will be retained for a maximum period of three years in a locked cabinet or in a password-protected computer in a locked office or laboratory in the EITC building at the University of Manitoba, to which only researchers associated with this project (Dr. Andrea Bunt, Md Ariful Islam Anik) have access. The data collected through the online study platform (Qualtrics) will be stored in their secured server, which is only accessible from a password protected account of one of the researchers (Md Ariful Islam Anik). All data will be destroyed by December 2022. In addition, the University of Manitoba may look at research records to see that the research is being done in a safe and proper way. We intend to present results as academic publications and a thesis which will be published in MSpace. Once published, the results of the study will be made available to the public for free at http://hci.cs.umanitoba.ca/. Again, no personal information about you will be included.

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research study. By doing this you also confirm that you are of the age of majority in Canada (18 years or more). In no way does this waive your legal rights nor release the researchers, sponsors, or involved institutions from their legal and professional responsibilities.

This research has been approved by the University of Manitoba Joint Faculty Research Ethics Board. If you have any concerns or complaints about this project, please contact Dr. Andrea Bunt at or the Human Ethics Coordinator at

A copy of this consent form has been given to you to keep for your records and reference.

☐ I wish to receive a summary of the findings.

☐ I wish to receive a copy of the interview transcript of the audio recording to confirm its accuracy.

Please write your email address if you checked the box above:

Participant's email address: _____

Participant's Signature: _____ Date _____

Researcher's Signature: _____ Date _____

Appendix K

Initial Questionnaire Used in the Second Study

Participants were ask some questions on their background and experience with machine learning systems.

1. How old are you?
2. How do you identify yourself?
3. Which country are you from?
4. What is your educational background?
5. Have you ever taken any Machine Learning (or related) course?
6. What is highest level of education you have completed?

Participants provided their agreement with each of the statements on a 7-point Likert scale with values from 1 (least agreement) to 7 (highest agreement).

1. I am confident using computers
2. I understand how computer algorithms work
3. I can make use of computer programming to solve a problem
4. I understand how Amazon recommends products for me to purchase
5. I understand why I see relevant ads in social media

Appendix L

Questionnaire Used after Each Scenario

Participants provided their agreement with each of the statements on a 7-point Likert scale with values from 1 (least agreement) to 7 (highest agreement).

1. I am confident in the system
2. The system has integrity
3. The system is dependable
4. The system is reliable
5. I can trust the system
6. I am familiar with the system
7. The system is free of bias
8. The system upholds ethical and moral standards
9. The system's explanations are reasonable
10. I would agree with the system's decision based on the explanation.
11. The system was trained in a fair way.
12. I would feel comfortable using the system's decision
13. The explanation gives me ideas about the data used in the system
14. The explanation helps me to reflect on whether the training process was fair

Appendix M

Semi-Structured Interview Sample Questions

1. Do you have any experience of receiving decisions from similar systems in your life?
2. Assuming you have, do you understand the process of how these computer systems make decisions?
3. Do you think these explanations would have benefitted you in those contexts?
4. What did you get from these explanations? What's the high-level idea?
5. What do you think the explanations are communicating to you?
6. Did you find it easy to understand the explanations? Was the information easy to digest? Was it easy to navigate?
7. Did the explanation helped to reflect on the training process of the system?
8. Is it possible to agree with the system based on these explanations?
9. What information you found noteworthy?
10. What information could help/helped to increase your confidence in the system?
11. Which category/information in the explanations were most helpful for you? Why?
12. Is there something that does not help you or you felt of less important to know?
13. Based on the explanations, can you make a judgment about the system?
14. What else would you want in the explanations?
15. Do you think these explanations can affect your trust in the system?

16. Do you think these explanations will help you to reason with the decision?
17. Do you think the stakes of the decision has any impact on your judgment?
18. Do you think the stakes of the decision has any impact on how you perceive the explanations?
19. Do you think the need of this explanation is dependent on the stakes of the decision?