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

Ariful Islam Anik
University of Manitoba, Winnipeg, Canada
aianik@cs.umanitoba.ca

Andrea Bunt
University of Manitoba, Winnipeg, Canada
bunt@cs.umanitoba.ca

ABSTRACT

Training datasets fundamentally impact the performance of machine learning (ML) 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 stakeholders. In this work, we explore the concept of data-centric explanations for ML systems that describe the training data to end-users. Through a formative study, we investigate the potential utility of such an approach, including the information about training data that participants find most compelling. In a second study, we investigate reactions to our explanations across four different system scenarios. Our results suggest that data-centric explanations have the potential to impact how users judge the trustworthiness of a system and to assist users in assessing fairness. We discuss the implications of our findings for designing explanations to support users' perceptions of ML systems.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); HCI design and evaluation methods; User studies.

KEYWORDS

Machine Learning Systems, Explanations, Training Data, Transparency, Trust, Fairness, User Expertise

ACM Reference Format:
Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445736

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 [66, 103], to product and content recommendations [4, 19, 41, 98], to informing decisions on matters with substantial individual and societal impacts, such as hiring [17, 39, 76], finance [29, 56], medicine [16, 40], and criminal justice [26, 43, 57]. Despite their importance and impact, such systems are often "black-box" by nature [82] and consequently are not transparent [24, 69, 81]. The lack of transparency can make it difficult for end-users to interpret and understand system outcomes [24, 69, 81]. The opacity of these systems can also hurt a user's ability to form meaningful trust relationships with machine learning systems [28, 75, 87, 89] and to hold the systems properly accountable for their decisions [11, 65].

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 system transparency [30, 38, 65, 69, 87, 88, 91] as well as to understand the impact of increased transparency on user perceptions of and interactions with such systems [5, 14, 20, 33, 58, 85, 104]. Much of this work, however, has focused on explaining outcomes and the properties of decisions to end-users [20, 30, 87, 88, 91], 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 way 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 [80] are ultimately reflected and aggravated in the deployed system [3, 105]. 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 our research questions of whether and how training dataset information could be communicated to end-users. What types of training data information might be available? How should such information be presented to end-users of varying backgrounds in machine learning? What impact could explanations that focus on training data have on perceived trust and fairness judgments of ML systems?

To answer our research questions, we first consulted prior work on training dataset documentation [44] to identify communicable information to end-users. We then used an iterative user-centered design process to develop prototype explanations that we refer to as *data-centric explanations.* The term "explanation" has been used broadly in the literature to characterize approaches to making systems more interpretable and transparent [5, 14, 79]. Our prototype

CHI '21, May 08–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445736

explanations aim to increase system transparency by describing the data used to train a system, which as described above, can fundamentally impact a system's behaviors.

In a study with 27 participants of various backgrounds, we investigated the impact of our data-centric explanations on participants' trust and fairness judgments across a range of four system scenarios. Our participants felt that the explanations helped them reflect on the training process, impacted their sense of trust in the system, and were particularly important for high-stakes systems. While the explanations received support from all expertise groups in our study, we noted subtle qualitative differences in how machine learning experts and non-experts approached the explanations. For example, some machine learning experts questioned whether the information would be understandable to those without ML training, whereas the non-experts felt the information was both clear and useful. Collectively, our findings establish data-centric explanations as a viable, promising approach to improving system transparency.

To summarize, our paper makes the following contributions: 1) We present data-centric explanations that focus on communicating information on training datasets to end-users. 2) We present study findings that show the potential for this type of data-centric explanation to influence users' perceptions of machine learning systems.

2 RELATED WORK

In this section, we 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, 64, 77, 98], healthcare applications [16, 18, 52, 93], finance [12, 29, 42, 45], hiring [39, 72], and criminal justice [94, 97, 102]. 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, 102] (the degree of influence of each input on the system output); *sensitivity based* [5, 87, 91] (how much the value of an input would have to differ to change the output); *demographic-based* [1, 5, 98] (aggregate statistics on the outcome classes for people in the same demographic categories as the decision subject); *case-based* [5, 14, 79] (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, 60, 96] (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. Our explanations 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, 61, 70]. Model-agnostic explanations, on the other hand, can explain decisions from a range of ML models [87, 88], 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, 79, 87]) 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, we design data-centric explanations that are model-agnostic and global.

2.2 Evaluating the Impact of Explanations

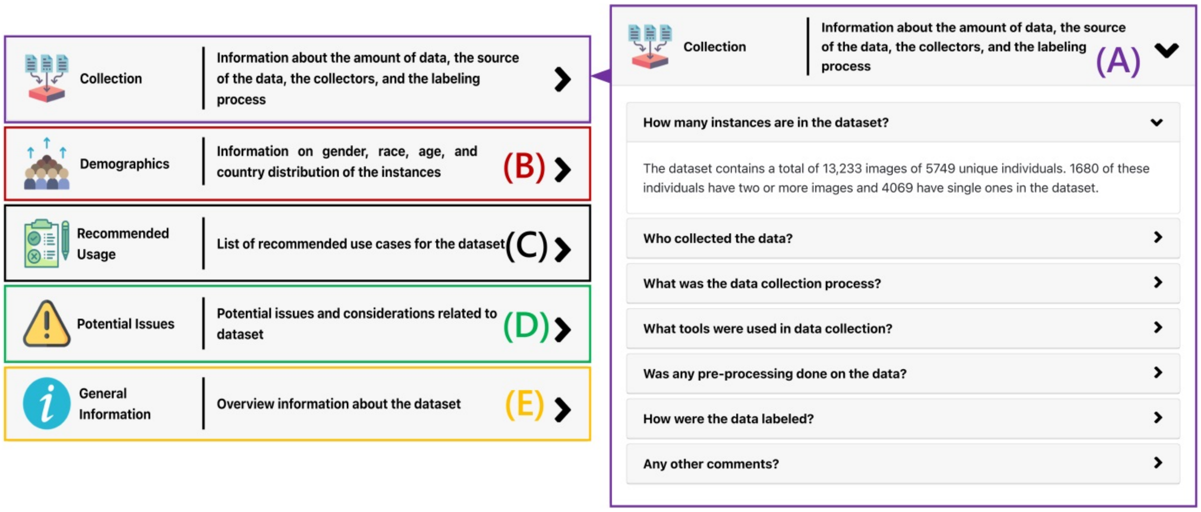
In parallel to developing different explanation approaches, numerous studies have investigated the impact of explanations on user perceptions of and interactions with machine learning systems [5, 14, 20, 33, 58, 59, 85, 101, 104].

Prior work has found that increased transparency through explanations can increase user acceptance of the systems [28, 49, 59, 100]. However, increased transparency does not always lead to increased trust. While many studies have found that explanations impact users' satisfaction and trust positively [9, 58, 62, 74, 84], some have found that explanations had no impact on trust [20, 28, 83], 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 [101], system failures [32, 37], soundness of the explanation [63], and the amount of information presented in the explanation [58]. 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 specific 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. We were 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



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 dataset

General Information — Overview information about the dataset

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 a total of 13,233 images of 5749 unique individuals. 1680 of these individuals have two or more images and 4069 have single ones in the dataset.

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?

Figure 1: Overview of the prototype of our data-centric explanations. The main screen, which lists the information categories, and provides a short description of each can be seen on the left. (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. To see the details of each category, please check the Supplementary Materials (Sup-D).

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, 57]. 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, 92, 99] and this approach is starting to gain traction in some organizations (e.g., [2, 78]). In our work, we investigate how to communicate training datasets to a different audience, namely end-users, as opposed to machine learning specialists. We also investigate how such information might impact end-users' perceptions of machine learning systems.

3 DATA-CENTRIC EXPLANATIONS

Our data-centric explanations, shown in Figure 1, provide users with information on a system's training data. Our first step in designing the explanations was to get a sense of the type of information that might be captured during the training process. To this end, we leveraged Gebru et al.'s datasheets [44], which provide a standardized, in-depth documentation of datasets. As described in the previous section, Gebru et al.'s datasheets were designed for machine learning specialists rather than system end-users. To transform this information into an appropriate form for end-users, we used an iterative user-centered design process. We explored and piloted several presentation styles, including the use of infographics, flowcharts, and different other ways to segment the information. Samples of our early sketches and low-fidelity prototypes are included in the Supplementary Materials (see Sup-A). We also piloted information subsets to find a balance between being comprehensive and overwhelming.

Informed by pilot studies, we settled on including five different categories of information (collection, demographics, recommended usage, potential issues, and general information) in our prototype explanations, as shown in Figure 1. Within each category, the prototype employs the question-based approach used by Gebru et al. [44] to provide users with specific answers to questions about training data. Figure 1 shows the questions for the collection category. Our pilot participants indicated that this format made the volume of information manageable since they could focus on the topics they were most interested in. The question-based approach has also been shown to work well in a range of prior work on explainable AI [71, 73, 86]. The prototype was built using web technologies (HTML, CSS, and JavaScript).

4 STUDY 1: INITIAL CONCEPT EXPLORATION AND PROTOTYPE REFINEMENT

In our first, exploratory study, we used semi-structured interviews to learn about what participants generally know about machine learning systems and their workflows, and how they would feel about potential data-centric explanations from the systems. We used an earlier version of the prototype shown in Figure 1, which is provided in the Supplementary Materials (see Sup-B), to ground discussions.

4.1 Participants

We recruited 17 participants (10 men, 7 women) by putting posters around a university campus and by reaching out to personal contacts. We 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. We targeted most of our recruiting on those outsides of the machine learning field (15 of our participants), however, we also included two experts for the sake of contrast. The average age for our 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 our institutional research ethics board.

4.2 Study Method and Procedure

Our study consisted of in-person semi-structured interviews, covering questions on participants' existing knowledge and experience with machine learning systems, their ideas on algorithmic fairness, and their thoughts on data-centric explanations. We began the study sessions with demographics and background questions. We asked our participants about their experiences of receiving decisions from a range of decision-making systems (e.g., ad recommendation, automated hiring, criminal justice system). We then discussed the role that data plays in the decisions from these systems. For most participants, this topic came naturally into the conversation, for others we initiated it. We then transitioned to the information categories we created for our explanations. For each category, we asked about participants' existing knowledge of the category and what they would be interested in learning more about. We then showed them a prototype (see Sup-B in the Supplementary Materials) and asked for their feedback. We also asked participants about the potential for this type of explanation to influence their judgments of fairness and their trust in the system. To conclude the session, participants rated each piece of information in the prototype explanation on two 5-point Likert scale items: one for understandability and one for usefulness. We 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. Our discussions with participants revealed insights on why data-centric explanations are worth pursuing. All our participants reacted positively to the idea of having data-centric explanations and were interested in having more information about training data. We also asked participants if they are aware of fairness issues in machine learning and gave some examples of existing fairness problems. We were 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

After we discussed data-centric explanations with our 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. We saw some initial indications that user expertise might impact attitudes towards data-centric explanations. While we found that both expert and non-expert participants had positive things to say about having data-centric explanations, our 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 Q&A Format Well Received, But Some Answers Need More Depth. When discussing the prototype, most participants 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

During the interviews, we also probed on specific design issues, including soliciting participants' ideas on potential presentation styles other than the question-based approach presented in Figure 1. Like in our pilots, study participants liked the question-based approach, feeling that it helped guide their focus to information that they were most interested in.

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.

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

Based on the above feedback, we revised the prototype to include more depth in the answers when possible, which resulted in us adding more detail to almost all answers in the prototype. For example, in the early version of the prototype, the answer to the question "*how many instances are in the dataset?*" was "*13,233 face images of 5749 individuals*". Figure 1A shows the additional detail that we added on how the instances consisted of single/multiple images of individuals. As a second example, for the question "*how the data was labeled?*", the original answer was "*Image labels were obtained from external sources*". We expanded this answer to include information on the labeling process and data labelers as follows: "*Each image is accompanied by a label indicating the facial expression of the person in an image. The expression in each image in the dataset was determined by an operator by looking at the face images and the context of the photographs*".

We also used questionnaire responses to look for opportunities to either clarify the answers or discard the questions altogether. We ended up discarding 3 questions (information on the dataset creators, funding source, and maintenance information) from our original prototype, where participants indicated they understood the information but rated it low on usefulness. The prototype in Figure 1 depicts the revised version.

5 STUDY 2: INVESTIGATING THE UTILITY OF THE EXPLANATIONS ACROSS A RANGE OF SCENARIOS

Our first study showed some support for the concept of data-centric explanations and we were able to use the feedback to refine our prototype to the version depicted in Figure 1. In our second study, we investigate how our 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 we observed in our exploratory study, we were also interested in understanding potential differences in participant's perceptions related to their backgrounds in machine learning.

5.1 Participants

To explore the role of user expertise in machine learning, we sought to include a range of backgrounds in our study. Specially, we recruited participants across three potential expertise dimensions, which we 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 and non-CS backgrounds, without prior experience with ML

We recruited 30 participants for our 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 us 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 our participant pool, we had 9 experts (5 men, 4 women), 8 intermediates (5 men, 3 women), and 10 beginners (5 men, 5 women) according to our definitions above. Participants received $20 for their participation. The study was approved by our research ethics board.

5.2 Study Design

Our 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. We also included a second, within-subjects factor, where we manipulated characteristics of the training data presented in the explanations. In our study, participants interacted with our 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. Following a growing body of research showing that users respond positively to having explanations present [5, 14, 20], and many similar studies without a control condition [33, 58, 63], we did not include a no-explanation condition in our study. Instead, our design prioritized the breadth of scenarios over comparison to a control condition.

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 1 (for more details on the scenarios as presented to participants, see Sup-C in the Supplementary Materials). 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.

To help generate realistic data-centric explanations for each scenario, we consulted reference datasets for bail decisions [67], labeled faces in the wild [53], graduate admissions [106], and speaker recognition [23]. We adjusted this information as needed, for example, to generate potential red flags. For missing information, we either generated fictitious data in a manner consistent with the other explanations or listed the information as being "unknown".

5.4 Study Procedure

Our study sessions took place online, using a video-conferencing platform of the participant's choice. We began the study session by asking participants some introductory demographic questions, including questions on their experiences with computer systems and machine learning. We then presented the four scenarios to our participants, one at a time using Qualtrics [107]. 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

Table 1: Scenarios used in the study.

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

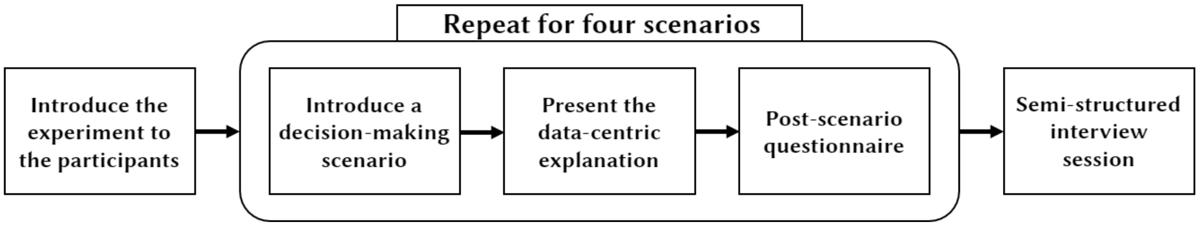

Repeat for four scenarios
Introduce the experiment to the participants
Introduce a decision-making scenario
Present the data-centric explanation
Post-scenario questionnaire
Semi-structured interview session

Figure 2: Study procedure

explanation communicated information on the training dataset information to them and whether or not they found it helpful. One full explanation example can be found in the Supplementary Materials (see Sup-D). Our pilot testing 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 the Supplementary Materials (see Sup-E), 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. We adapted existing scales to measure trust [55] and fairness [5, 25]. As shown in Figure 2, 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. We randomized the order of the scenarios across participants to mitigate potential order effects.

We concluded the study session with a 40-60 min semi-structured interview, where we solicited further information from participants on their experiences with machine learning systems, and their perceptions of the data-centric explanations. Throughout the interviews, we 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

We collected both quantitative data (from the post-scenario questionnaires) and qualitative data (from the post-session interviews). For the quantitative data, we 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). We used p=0.05 as our threshold for statistical significance. To analyze our interview data, we first transcribed all the interview sessions. We then conducted bottom-up affinity diagramming [27] on participant quotes from the interview transcripts. Two researchers were involved in the data analysis. The researcher who conducted the interviews created the initial affinity diagrams, coding the resulting clusters using an open coding scheme. The two researchers then collaboratively looked for themes in the coded data. We did several iterations of this analysis, revisiting the raw data frequently.

6 FINDINGS

We 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. We then present findings from our interviews that provide further insights into how and why the explanations impacted their trust and sense of fairness.

6.1 Questionnaire Data

6.1.1 Impact of Expertise. As Table 2 illustrates, Expertise did not significantly impact any of the measures in our questionnaire for any of the training dataset characteristics.

The data indicate 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 shown in Figure 3, the scores were comparatively low for scenarios where the explanations revealed potential problems with training data (red flag scenarios) and high for scenarios with more balanced training data.

6.1.2 Impact of Training Dataset Characteristics. We 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 3 shows that participants had significantly more trust in the system, felt that the system was

Table 2: Median (IQR) values for the Likert-Scale questionnaire responses by Expertise level. Since some measures combine multiple questionnaire items, we also provide the scale range (Low-High).

	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
Reflect 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

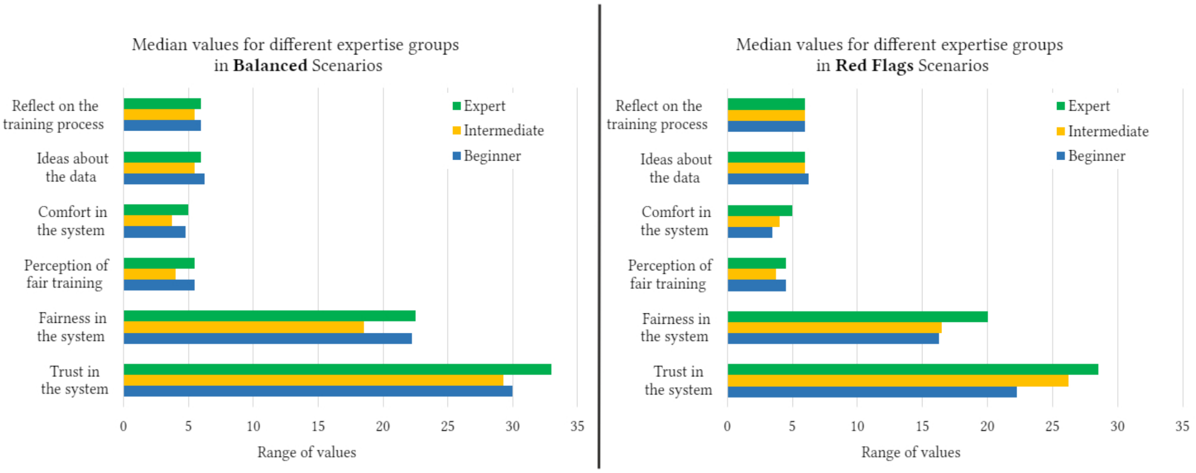


Figure 3: Median values for each of the measures for the expertise groups across balanced and red flags scenario. The scale range for each of the measures is same as described in Table 2.

Table 3: Median (IQR) values for the Likert-Scale questionnaire responses according to Training Data Characteristics. Since some measures combine multiple questionnaire items, we also provide the scale range (Low-High).

	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
Reflect on the training process by the explanation	1.00-7.00	6.00 (1.00)	6.00 (1.50)	-0.619	0.536

fairer, 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.

6.2 Interview Findings

Findings from our interviews provide insights into how and why the data-centric explanations impacted participants' trust and their judgment of fairness in the systems. We also describe commonalities and differences that we observed across the different expertise groupings. In the quotes below, "E" = Expert, "I" = Intermediate, and "B" = Beginner.

6.2.1 *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.

A small group of participants (5/27) felt that 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's good. But it doesn't make me necessarily trust the system more. It depends on the information they're providing." – P30-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 wouldn't 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

6.2.2 *Training Data Demographics Perceived As Most Influential.*

Nearly all our participants (25/27) found value in the demographic information of our data-centric explanations (Fig-1: 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 get a clear picture of the potential biases in the training data. We noticed that regardless of expertise, these 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

> "[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, most 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, our participants considered two other categories important. The first was collection information (Fig-1: A), mentioned by several participants (14/27) as providing key information on sample size and how the data was gathered:

> "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 [that the] gender was completely equal, however, the sample size [is] of 100 people. I can't really trust it until your sample size is

> great amount. And then after that, you know, I look at other stuff. But first, I want to look at sample size." – P15-B

> "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

Other participants (16/27) indicated that error information (Fig-1: D) impacted their confidence in the system:

> "Let's say there's no errors or not much errors [in the data]. That definitely makes me more confident. I think that the one with the bail decisions, that [had a] kind of pretty high error rate. And so that definitely gives you less confidence in the system." - P19-E

6.2.3 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.

> "[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 our 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].

6.2.4 Data-Centric Explanations Help Participants' Assess Fairness But Are Not Enough. Most of our 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 (Fig-1: B), and whether there are any fatal flaws in the system from the error information (Fig-1: 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.

6.2.5 Many Commonalities Across Expertise Groups, But With Nuanced Differences. We saw many similarities in how the different expertise groups responded to our explanations. Regardless of participants' ML training or technical background, our qualitative data suggest that the data-centric explanations impacted participants' sense of trust in the machine learning systems described in our 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.

We did, however, see some nuanced differences in how participants' machine learning backgrounds impacted how they felt about our data-centric explanations. As we 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 our participants reported the explanations to be easy to understand.

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

7 DISCUSSION

Our results from a study with 27 participants suggest that data-centric explanations have the potential to help people develop an informed sense of trust in machine learning systems. In our study, participants' trust was impacted positively when the training data seemed balanced and negatively when the explanations revealed problems. Like prior work, we 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 [101].

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 our explanations could serve as complements to established explanation approaches that explain the outcomes and the properties of a decision [20, 30, 87, 88, 91]. How users might prioritize data-centric explanations vs. feature-oriented explanations is an important area of future work. We also acknowledge that fairness is a social and ethical concept, and that perceptions of fairness are multi-dimensional and context-dependent [46, 47, 68]. While we measured fairness using scales used in prior work [5, 54], 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.

Our qualitative findings 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 we 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?

For some of the non-expert participants, we 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. We 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 [90], need for cognition [13], visual literacy [7]) can have a significant influence on their perceptions of explanations from the system [33, 77]. 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, we used a simple objective measure of technical background to include what we categorized as both novices and intermediates. While this approach did seem to help diversify our sample, we did not see clear differences between these two groups in their attitudes towards the explanations. We 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.

Our study's scenario-based approach, a commonly used method to study user perceptions of machine learning systems [5, 47, 73, 95, 104], allowed participants to reflect on a range of potential scenarios that were grounded in real-world examples. A limitation of this method, however, is that because 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 based on existing documentation from actual machine learning models, the explanations themselves might have lacked some degree of ecological validity. Given that prior work has suggested that simulating explanations can impact the generalizability of study findings [5], additional studies are needed to understand how users might respond to the explanations in-situ, where they have more direct interactions with real systems and/or the systems' outputs.

A second key limitation of our study is its lack of behavioral, objective data, with our study instead relying on Likert-scale self-reports and qualitative interview data. In moving towards more ecologically valid studies that elicit behavioral data in addition to subjective responses, a challenge will be sufficient availability of and access to real-world systems with documented datasets. We hope that our findings will motivate more dataset creators to document their datasets and more ML developers to make this information available. Finally, while our study did not reveal significant quantitative differences in our questionnaire data based on participant expertise, it is possible that some of these effects might reach significance with a larger sample. Future work should therefore also explore the generalizability of our findings to a larger number of participants.

Our study scenarios asked participants to take on the role of the end-user of machine learning systems – 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 [67, 97], 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.

8 CONCLUSION

We presented data-centric explanations that focus on providing end-users with information on the training data of machine learning systems. From a study with 27 participants of different backgrounds and machine learning expertise, we showed that data-centric explanations can help people to get insights into the system, reflect on the training data, and influence their assessments of trust and fairness. Our 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.

ACKNOWLEDGMENTS

We thank the anonymous reviewers and University of Manitoba HCI Lab members for their valuable feedback on the paper. We also thank the study participants for contributing their time. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[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. *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*: 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. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* 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. *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*: 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. In *User Modeling 2007*, 147–156.

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classificatio. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 10: 1889–1896. https://doi.org/10.2147/OTT.S126905

[11] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1: 2053951715622512. https://doi.org/10.1177/2053951715622512

[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. *Proceedings of the 24th International Conference on Intelligent User Interfaces*: 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. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: 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] 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

[26] 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.

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

[28] 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

[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. *Transparent Data Mining for Big and Small Data*: 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. *Proceedings of the 24th International Conference on Intelligent User Interfaces*: 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 Iii, and Kate Crawford. 2018. Datasheets for datasets. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. 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. *Proceedings of the machine learning: the 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*: 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. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: 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? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: 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] J. A.Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of applied psychology 68*, 386–399.

[55] 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.

[56] Amir E Khandani, Adlar J Kim, and Andrew W Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34, 11: 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

[57] 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

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

[59] 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. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: 1–14. https://doi.org/10.1145/3290605.3300641

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

[61] Todd Kulesza, Margaret Burnett, Weng Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*: 126–137. https://doi.org/10.1145/2678025.2701399

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

[63] 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

[64] 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. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: 1–12. https://doi.org/10.1145/3290605.3300717

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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

[70] 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

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

[72] Cynthia C S Liem, Markus Langer, Andrew Demetriou, Annemarie M F Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph. Born, and Cornelius J König. 2018. Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer International Publishing, Cham, 197–253. https://doi.org/10.1007/978-3-319-98131-4_9

[73] 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

[74] 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

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

[76] 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

[77] 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. *Proceedings of the 24th International Conference on Intelligent User Interfaces*: 397–407. https://doi.org/10.1145/3301275.3302313

[78] 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

[79] 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

[80] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2: 13.

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

[82] 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

[83] 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

[84] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. *Proceedings of the 11th International Conference on Intelligent User Interfaces* 2006: 93–100. https://doi.org/10.1145/1111449.1111475

[85] 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

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

[87] 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

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

[89] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, {IJCAI-17}* 0: 2662–2670. https://doi.org/10.24963/ijcai.2017/371

[90] 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

[91] 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

[92] 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

[93] 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

[94] 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

[95] 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 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 2459–2468. https://doi.org/10.1145/3292500.3330664

[96] 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

[97] 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

[98] 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

[99] 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

[100] 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

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

[102] 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

[103] Yong Zhang, Hongming Zhou, Nganmeng Tan, Saeed Bagheri, and Meng Joo Er. 2017. Targeted Advertising Based on Browsing History. *CoRR* abs/1711.0. Retrieved from http://arxiv.org/abs/1711.04498

[104] 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. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 295–305. https://doi.org/10.1145/3351095.3372852

[105] 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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. https://doi.org/10.18653/v1/D17-1323

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

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