Understanding User Sensemaking in Machine Learning Fairness Assessment Systems

Ziwei Gu*
Cornell University
zg48@cornell.edu

Jing Nathan Yan*
Cornell University
jy858@cornell.edu

Jeff Rzeszotarski
Cornell University
jeffrz@cornell.edu

ABSTRACT

A variety of systems have been proposed to assist users in detecting machine learning (ML) fairness issues. These systems approach bias reduction from a number of perspectives, including recommender systems, exploratory tools, and dashboards. In this paper, we seek to inform the design of these systems by examining how individuals make sense of fairness issues as they use different de-biasing affordances. In particular, we consider the tension between de-biasing recommendations which are quick but may lack nuance and "what-if" style exploration which is time consuming but may lead to deeper understanding and transferable insights. Using logs, think-aloud data, and semi-structured interviews we find that exploratory systems promote a rich pattern of hypothesis generation and testing, while recommendations deliver quick answers which satisfy participants at the cost of reduced information exposure. We highlight design requirements and trade-offs in the design of ML fairness systems to promote accurate and explainable assessments.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Algorithmic Bias; User Sensemaking; Interactive interfaces

ACM Reference Format:
Ziwei Gu, Jing Nathan Yan*, and Jeff Rzeszotarski. 2020. Understanding User Sensemaking in Machine Learning Fairness Assessment Systems. In *WWW'21: The Web Conference 2021(WWW'21), April 19 - 23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/xxx

1 INTRODUCTION

Biases are silently encoded into decision-making processes, potentially affecting a wide variety of domains including human resources[24], health care[1, 16] and policy making[41, 67]. With pervasive deployment of machine learning (ML) models in real-world tasks, such biases remain pervasive but may be masked by a veneer of computation. While ML models can reduce costs and improve the accuracy of decision-making, they are only as effective as their training data and model-builders. Due to these risks, researchers now investigate the social impact of ML biases empirically[7] and develop metrics and algorithms to detect and mitigate the bias theoretically[28, 36]. Recent research has synthesized these streams of work, leading to novel tools for interactively investigating bias issues in ML systems[9, 14, 23, 33, 75].

Such de-biasing systems generally offer affordances for exploring bias issues across different metrics. Many include recommendations as a way to bootstrap the process. Some are almost entirely automated. For example, IBM AIF 360 [9] offers recommendations - the user simply provides a model and dataset as inputs, and it will output biases with visualizations. Recommendation delivers answers quickly, but it risks reducing transparency and accountability. Additionally, recommendation may lead to overconfidence from inexperienced analysts and mask subtle, more pernicious biases. Recent work has exposed how exploratory interfaces may offer a different value proposition [75]. Instead of delivering instant results, users must explore to develop an understanding of possible issues. While this may grant a deeper understanding of bias that can be better transferred to future tasks, it requires more effort on the part of the analyst and may require more expertise.

It is critical for the ML community to develop effective, usable design patterns for interactive de-biasing. Moreover, a number of projects [13, 47, 59] identify how misalignment between folk definitions of fairness and existing statistical definitions can cause conflict, necessitating careful design practice to encourage proper reasoning about fairness that is well grounded in best practices. Yet, at present we lack knowledge about how the design of interactive de-biasing tools shapes how users reason about fairness. In order to develop more effective tools, it is crucial to understand the trade-offs between salient design features like recommendation and exploration.

In this paper, we investigate how various de-biasing tool affordances shape the way that individuals make sense of fairness issues in data. In particular, we focus on how individuals make use of machine recommendations and open exploratory interfaces. Through a think-aloud investigation, surveys, and semi-structured interviews we expose how process-level differences that result from these tool affordances can have dramatic downstream impacts on both outcomes and user experience.

Our work makes the following contributions:

- Through an analysis of existing data, we identify design issues in exploratory and recommender de-biasing systems.
- Using a think-aloud methodology, we observe participants using three different de-biasing tools in order to understand how interactive affordances shape their understanding of bias.

[*] both authors contributed equally to this paper

WWW'21, April 19 - 23, 2021, Ljubljana, Slovenia
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/xxx

- We synthesize the results of our think-aloud and highlight specific design considerations for the ML fairness community.

Through our investigation we find that while exploration leads to better outcomes due to its encouragement of iteration and branching investigations, it is often unsatisfying for users and has significant barriers to entry. On the other hand, we show that recommendations, while gratifying for users, risk misinterpretation and can exacerbate choice overload issues when working with complex data.

2 RELATED WORK

The machine learning pipeline as a whole is filled with many challenges for researchers [39, 50, 52, 61] spanning from proposing powerful algorithms and models [64, 66], to interface design [27, 45]. More recently, ML has entered a variety of domains including marketing [21, 35], ethics [7], law, and policy[41, 67]. With the expansion of ML, interest has risen in how models might encode societal biases which then propagate forwards into decisions made by systems that have real world implications [32, 65] including perpetuating societal injustices [58]. As a result, increasing effort has been directed towards designing ML fairness assessment systems. These efforts involve deploying bias detection algorithms and prototyping effective interfaces for de-biasing, necessitating cross-disciplinary efforts among technical fields [37]. In the following sections we first review the theoretical basis for designing de-biasing tools, and then step through a variety of tools and techniques that have emerged.

2.1 Machine Learning Fairness

A wide variety of projects have proposed statistical definitions to quantify unfairness in data and ML models and corresponding algorithms to mitigate the issues [2, 10, 15, 20, 28, 36]. These metrics have different targets and application scenarios, and, unfortunately, can often be mutually exclusive [43]. Moreover, [32] benchmarked existing fairness-aware ML algorithms and proved that existing solutions correlate with each other and are sensitive to dataset composition. More recently, causality has been introduced as a means to generalize unfairness detection and add robustness [42, 51, 79].

The HCI community has also systematically investigated fairness from many different perspectives. Projects have identified the impact of machine learning fairness in different sectors of the society [7, 12, 34, 60, 73]. For example, Hamid et al. [34] and Schlesinger et al. [60] researched racial and gender bias in ML systems. Perceptions of bias also play a role in propagating negative effects. [73] highlighted how unfairness builds mistrust. Recently, HCI researchers have also conducted empirical investigations [13, 37, 47, 59]. Accountability is also a factor in ongoing work [48, 55, 77].

2.2 Sensemaking

Sensemaking can be defined as the process that humans employ to construct meaning from raw data [56, 70]. Many fields, such as visualization [18], information retrieval [6] and communication [71] consider aspects of sensemaking. In foundational work, Russel et al.[56] examined the process a team used to assemble educational materials, identifying how pieces of information are fit into larger schema. Pirolli and Card [54] extend this work, highlighting how individuals forage for data and sensemake schemas in an iterative process. In their notional model, iterative refinement emphasizes how understanding grows as part of a process and is not instantaneous. Recent projects have explored how to mitigate cognitive biases [57, 68] and promote more complete understanding of ML [29, 74].

2.3 Systems for ML Fairness

A variety of articles [22, 24, 60] have highlighted the importance of de-biasing systems in improving ML fairness. We identify two parallel threads in such systems, either focusing on recommendations

Recommendations: Recommender systems and toolkits have been developed to direct users towards mitigating bias in their ML pipelines. IBM AI Fairness 360 (AIF) [8] automatically detects unfairness issues and fixes the biases using integrated metrics and mitigation algorithms. Similarly, Fairlearn [14] provides automatic solutions for biased inputs, particularly for binary classification and regression tasks. Now, commercial platforms like Microsoft Azure integrate tools like Fairlearn into their offerings [1]. [33] is a human-in-loop system enabling practitioners to view automatically generated visualizations and correct models for bias. ML-fairness-gym [23, 30, 38, 49] contains a set of simple simulations that explore the potential long-run impacts of deploying machine learning-based decision systems in social environments, and maintain a learning agent that interacts with an environment.

Exploration: While many of the aforementioned toolkits can be used in an exploratory capacity, exploratory analytics is not their specific aim. Exploratory data analytics tools [62] often have affordances for experimenting and highlighting trends but ultimately giving users a sandbox to explore. Recommendations can assist, but it risks biases such as anchoring [68]. Silva[75] is an interactive system that employs a causal view to help users reason about fairness. Exploratory tools have seen use in other areas, including discovering classification error [19] and preventing false discoveries [44].

Balancing recommendations and exploration: While recommendations help to quickly reach actionable conclusions, they bring with them the risk of injecting unrealistic assumptions silently [25], lacking necessary accountability [77], and causing cognitive biases or misconceptions [17, 67, 69]. [76] highlight a few reasons why human analysts may mistrust recommendation systems, assuming they know they are using one at all [31]. When they are given the power to tune a recommendation, analysts had higher confidence in it [26]. Exploratory tools like Silva [75] and AnchorViz [19] might help to bridge this gap. Though there are a number of recent UX-focused studies of ML systems[3, 4, 19, 45, 52], to the best of our knowledge, there is no systematic study on how the design of de-biasing systems helps (or harms) the sensemaking of analysts engaged in de-biasing.

3 PRELIMINARY ANALYSIS

In order to understand how the design of interactive data de-biasing tools shapes their use, we first conducted a preliminary analysis using existing data generated during the evaluation of an interactive bias exploration tool. Our goal in this portion of our work was to identify seed candidates to later explore through more in-depth think-aloud studies. Given the degree to which task, dataset, and user group can affect the outcome of laboratory studies of interactive data tools, we analyze the activities and performance of users in a prior

[1] https://docs.microsoft.com/en-us/azure/machine-learning/concept-fairness-ml

comparative study in order to find key breakpoints, performance differences, areas of contention, and overarching themes that can help to inform our main investigation.

The subject of our preliminary investigation is an interactive debiasing tool, Silva [75], that uses a view of causal relationships among data variables in order to help users perform exploratory analyses of bias issues in their datasets. Silva has four major components: a Dataset Panel (where users select attributes and toggle attribute sensitivity), a Causal Graph view (visualizing causal relationships among attributes), a Table Group (displaying info on training data and user-formed groups), and a Fairness Dashboard (bar charts showing fairness values across models, metrics, and user-defined groups). In prior work on the tool, the authors of Silva found that the tool helped users to diagnose issues in datasets as well as or better than IBM AI 360 (AIF)[8]. Further, they found that though there were task-dependent changes, skill wasn't a factor [75].

In our analysis, we explore anonymized event log and unpublished performance data made available by the authors of Silva for further study. The event log dataset the authors collected contains information on participants' views of socially biased attributes before and after using both Silva and AIF [8], their activities when exploring Silva, and the time each participant spent using different components of Silva. 30 participants were asked to point out socially biased and socially acceptable attributes in two datasets before and after using either tools. We focus on three main subjects in our investigation:
Q1. How do both tools (Silva and AIF) shape participants' ideas, hypotheses, and goals during sessions?
Q2. How did participants make use of the de-biasing tools, and how might their use have been shaped by tool affordances?
Q3. How did the tools assist participants in reasoning about fairness in their datasets?

3.1 Shifts in Understanding During Tool use

One trend in the user performance dataset for [75] suggests that a fair number of participants re-thought their understanding of the dataset, exhibiting shifts towards different focus attributes (and in many cases arriving on the ground truth answers). Evidence for this emerged in self-reported responses in pre- and post-surveys. This suggests that the de-biasing tools, though their use, shaped participants' views. Not only did they accomplish the goal of identifying sensitive attributes, but their understanding of the space of data might also have changed. In order to gain some insight on this finding, we re-processed this dataset. We selected for participants who reported different views on attributes before and after using either tool.

Among those who changed their views after using the assigned de-biasing tools, we counted the number of ground truth biases they correctly identified in their post-survey through use of the systems. In general, both tools led participants to change their views: 14 participants using AIF 360 and 15 participants using Silva exhibited shifts in their understanding of sensitive attributes. However, aligning with what the authors of [75] found, among those who changed their opinions, there remained a significant difference (t(58)=2.2841 ρ<.0260) in the number of correct answers participants finally got between Silva's exploratory tool (M: 0.53, Std: 0.78) and AIF (M: 0.07, Std: 0.78). This provides some initial evidence that the interactive affordances in Silva in particular provided some kind of "secret ingredient" for improving performance (and might be informed by further investigation. Among the group of users who showed shifts in their understanding, we also could not detect differences based on self-reported skill, aligning with the prior analysis.

3.2 Mining Interaction Logs

We were able to obtain a variety of activity logs for participants using the Silva exploratory de-biasing tool [75]. These logs took the form of dwell times on specific components in the tool as well as usage counts. Event log series did not prove complete enough for full analysis. In examining the log data that we could access, we sought to identify which interface elements most directly related to changes in participant pre-/post-survey outcomes and total performance. These elements ought to be key areas of interest in further think-aloud study, as they hint at being broader leverage points for potential improvements in the de-biasing process.

Silva participants made use of all of the different affordances provided in the tool. As expected, the most novel feature of the tool, the causal graph view, received the most dwell time of all of the tools (M: 394.5s, Std: 797.9s, total session: 1200s) and the highest number of interaction events (M:11.0:, Std: 5.69) compared to the mean total operations performed by participants (M: 13, Std: 6.20) in the case of the simpler Berkeley dataset. The more complex adult dataset involved even more time spent interacting with causal graph (M: 464.2, Std: 504.0) and a concordant increase in the number of interaction operation (M: 34.45, Std: 35.23). While there was too much variability in the time data to observe a difference between datasets, we did observe a significant increase in causal graph interface actions between the simple and complex dataset (t(58) = 3.5991, $\rho < 0.0007$). This emphasizes the role that the causal view played in participants' understanding of the data, as mediated by the tool. Though, ironically, further controlled study would be needed to fully demonstrate a causal relationship between data complexity and causal graph use. This is a strong motivator for examining the precise role that the interactive causal graph plays through further think-aloud study.

One other key feature of the Silva tool was the ability to group attributes, save them, and examine how metrics changed across models. For example, the participants might have thought that certain attributes or data might be more biased. By grouping points, they might test this assumption. Taking this idea that grouping of points is a potential signal for inference and hypothesis testing, we examined the creation of groups in the Silva user study logs. We found that in the case where data were complex (adult dataset, M: 5.4, Std: 4.92), Silva users created significantly more groups compared to the simpler Berkeley dataset (M: 2.0, Std: 0.71, t(58) = 3.7463, $\rho < .0004$). However, it is hard to tell whether the creation of groups resulted in improved outcomes for participants. A think-aloud study would allow us to test our operating assumption that group making matches up to participant hypothesis generation and testing, and to gather suggestive evidence about the efficacy of interactive grouping tools in de-biasing. We have provided additional event log summary statistics in the Appendix of this document.

3.3 Hypothesis Testing and Inference

Finally, we re-examine the quotes and self-report feedback provided by participants in the pre-existing study dataset. While in the initial work the authors discussed general themes in findings, in this investigation we focused on the kinds of reasoning that participants conducted during their investigation. We operate primarily on inference, as this was not a specific goal of the survey instrument. Inferring from quotes, we can find a few suggestions to investigate further in an additional study.

One overarching theme in the qualitative data is that most participants reported that the causal graph was useful to them in making sense of the relationships between sensitive attributes. One reports, "the causal graph helped me to see that there was not a direct dependency between sex and admisson [sic]," in exploring the Berkeley dataset. This is suggestive evidence that the participant was testing a hypothesis that they had developed using their own intuitions and expertise. There was confusion as well, with one participant reporting feelings of being overwhelmed by the abundant and sometimes inconsistent information shown. They reported, "it was tough comparing across so many groups and to my surprise different metrics are changing in different directions". Effect attribution and response to changing interface state seems to be a persistent challenge. On the other hand, one participant said that "Silva made me rethink about other biased situations..." and that "the results contradicted [their] intuition". Another participant formed over 10 groups and reported frequently asking what-if questions in the form of "let me see what's gonna happen if I exclude age". In sum, it appears that exploratory affordances may have promoted reasoning about sources of bias and hypothesis-testing based on prior knowledge and inferences from exploration, though direct evidence is limited in this dataset.

3.4 Discussion

Revisiting our initial questions in this investigation, we did find evidence that interactive exploratory affordances did promote engagement, and in some cases may have led to better outcomes. Qualitative data suggests that interactivity helped participants to leverage their own background knowledge and then adapt it to fit the specific situation of a dataset (which may or may not match those intuitions). While we saw evidence of participants reasoning using metrics, we lack data to state for certain how the metrics influenced their decision-making. One key break-point we observed in the experimental data was how tool performance differed, and that the different affordances in the tools might have played a role in determining that outcome. While we saw evidence of exploration tools promoting testing, we did not receive the same self-report feedback in the case of AIF, which provided more direct answers. On the other hand, the Silva users made use of the interactive affordances for a prolonged period of time before they achieved those higher outcomes.

While this investigation helped to inform our perspective on two different approaches for helping users de-bias their data, it doesn't provide insight into the mechanisms at play in this task. Why precisely do participants use the causal view so much? How does use of that tool translate into improved performance (assuming it does at all)? Do participants' mental models of bias issues match the ground truth, or are they just optimizing specific metrics because they are shown as bars that can always be pushed higher? Additional study is needed in order to understand these issues.

In order to find more direct evidence to understand how core design elements shape users' sensemaking process, we conducted a series of think-aloud experiments among AIF 360 [8], Google What-If [33], and a version of Silva [75] identical to that used during the generation of the data analysed in this section. In the following section we discuss this new investigation.

4 THINK ALOUD STUDY

In order to understand how the design of interactive de-biasing tools shapes the data exploration and de-biasing process, we need more insight into the process that individuals follow and their ongoing state of mind while they use a tool. For this reason, we conducted a study across three de-biasing tools employing a think-aloud methodology. In a think-aloud study participants are encouraged to vocalize their thoughts as they complete a task. This stream of consciousness gives some insight into participants' inner cognitive processes. If the task is especially intensive, as de-biasing is, then it is more likely that the participant will not have spare attentional resources to control how they vocalize (for example, due to self-censorship). This means the data will be even more informative as to their inner state.

We conduct a think aloud study using three different tools. We make use of a web-accessible version of the Silva tool as it was when [75] was written, as well as IBM AIF[9] and Google What-if tool [33]. In thinking about the ways that tools might influence behavior, as informed by our preliminary investigation, we considered how automation and recommendation might serve to short-circuit the exploration process. While this has a beneficial effect of improving efficiency, it might come at the cost of reduced understanding of the entire dataset due to lack of exploration. Further, it may be at higher risk of suffering from unexpected flaws in algorithmic recommendation/optimization systems. We use the Silva tool as one pole in this spectrum, as it forces users to explore manually. At the other end of the spectrum, AIF mainly provides recommendations through an automated user experience. We have introduced Google What-if as a moderate, semi-automated tool.

Participants in the study used two of the three tools to complete two different de-biasing tasks (counter-balanced for order effects across participants). Within each task we prompted users to vocalize their thoughts, taking video footage which we later converted into scripts. We also gathered event log data and conducted a pre- and post-survey. Following completion of a session, for some participants we also conducted a semi-structured interview to learn more about their process retrospectively. In this section we will outline the general experimental procedure, describe our resulting sessions, and discuss how we processed the qualitative think-aloud data. In the next section we examine themes that emerged from our analysis.

4.1 Methodology

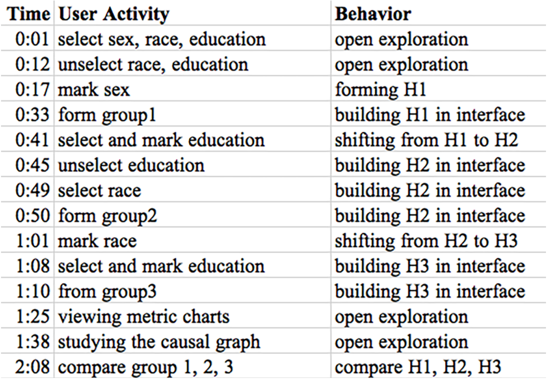
4.1.1 De-biasing Tools. In section 2 we took a deeper look into the spectrum of available de-biasing tools. For the purposes of this paper, we consider the aforementioned three tools that spread across a continuum from fully automated (AIF)Bellamy et al. [9] to semi-automated (What-if)Google [33] to manual exploration (Silva)Yan et al. [75]. Our intention in this investigation is not to be exhaustive. Rather, we aim to find salient differences across this spectrum of de-biasing interface design affordances. While recommendation-based tools are relatively common, there has been growing interest in interactive tools [75]. We hope to build on this thread by elaborating on the mechanisms that justify exploration's increased effort.

AI Fairness 360 (AIF) is an interactive web tool that steps users through the process of checking and remediating bias. Since we focused only on bias detection, we directed users to the "Check bias metrics" step of the tool and left out the "Mitigate" and "Compare" steps. Google What-If (What-if) is an interactive tool for understanding the behavior of a black-box classification or regression ML model, with built-in support for fairness evaluation. It allows its users to examine, evaluate, and compare machine learning models in a variety of ways, including editing datapoints, comparing counterfactuals, experimenting using confusion matrices and ROC curves, and testing algorithmic fairness constraints. Similar to AIF, we limit participants to the performance and fairness detection modules in What-if. Finally, Silva is an interactive tool for exploring potential sources of unfairness in datasets or machine learning models. Silva uses interactive elements to direct user attention to relationships between attributes through a global causal view, provides interactive recommendations, presents intermediate fairness results, and visualizes metrics. Screenshots of the interfaces for each de-biasing tool can be found in the Appendix.

4.1.2 Datasets. Our study employed three public datasets: Adult Census Income (Adult)[2]Zemel et al. [78], Berkeley Admission 1973 (Berkeley) Bickel et al. [11], and COMPAS Recidivism Risk Score (COMPASS)Larson et al. [46][3]. These three datasets have been widely studied by AI fairness researchers, resulting in established ground truth data on attribute sensitivity and bias. Each of these datasets contains features that will be familiar to many participants. They offer opportunities for them to leverage their own life experiences and domain expertise. At the same time, they span a breadth of complexities. As we observed in the preliminary experiment, dataset complexity seems to play a role even when participant skill does not result in performance differences.

4.1.3 Participants. We recruited 13 university undergraduate students to participate in this study via a university participant pool. Of them, 12 successfully finished the task. 5 identified as female, 7 as male, and 0 as non-binary. We included a pre-screen to filter out participants who already had deep exposure to bias concepts and the three datasets. As mentioned previously, participants completed two dataset tasks using two different tools. We randomly assigned conditions across the 12 participants so that each tool and dataset received even exposure and order was counter-balanced. This results in a semi-within-subjects design. Specific participants allocation details can be found in the Appendix.

4.1.4 Protocol. Due to social distancing constraints as a result of the COVID-19 pandemic, all studies were conducted virtually through an online meeting tool. Though we were initially leery of conducting remote usability studies, in the past asynchronous [5] and synchronous [53] usability studies have proved effective, albeit more resource intensive. We deployed all de-biasing tools online using a Bokeh server applet running on the Heroku cloud platform. While not incredibly high in specification, the server proved powerful enough to be responsive to users. Within each user study we conducted two sessions. In each session the participant first watched a tutorial video introducing the tool (of comparable length for each system), completed a pre-survey with demographic and background questions, and then prepared their assigned tool. They were then given a task and asked to complete it. Afterwards, participants completed a per-task final survey to reflect on their findings. Finally, participants completed a post-survey. Participants were offered the opportunity to volunteer to participate in a semi-structured interview about their experiences during the study.



Time	User Activity	Behavior
0:01	select sex, race, education	open exploration
0:12	unselect race, education	open exploration
0:17	mark sex	forming H1
0:33	form group1	building H1 in interface
0:41	select and mark education	shifting from H1 to H2
0:45	unselect education	building H2 in interface
0:49	select race	building H2 in interface
0:50	form group2	building H2 in interface
1:01	mark race	shifting from H2 to H3
1:08	select and mark education	building H3 in interface
1:10	from group3	building H3 in interface
1:25	viewing metric charts	open exploration
1:38	studying the causal graph	open exploration
2:08	compare group 1, 2, 3	compare H1, H2, H3

Figure 1: Sample of an activity log coded by a researcher

Within each de-biasing task participants read a short description of the dataset and associated models. On their pre-task survey, participants reported the attributes they thought might lead to unfairness (replicating the pre/post attribute data referenced in our preliminary post hoc analysis of the [75] dataset). Participants were then instructed to use the tool to find sources of bias inside of the model or dataset. As participants used the tool, we probed using a think-aloud methodology (outlined below). In the process of completing the task, participants were exposed to data variabled and reasoned about bias. Through a post-task survey, we asked them to report what sensitive attributes they observed and to reflect on how their thinking changed. After participants finished using both tools, they compared their experience with the two tools across different stages of their exploration (searching for information, searching for evidence, generating hypotheses, reevaluation, etc.) in a post-survey. Participants were encouraged to include detailed examples of how an element of a tool contributed (or did not) to their overall understanding of the data. Finally, we asked about their tool preference.

As participants used a tool to complete a task, the interviewers prompted them to vocalize their current thoughts and activities, following a think-aloud protocol [40]. For example, one participant might explain why they were creating a group, calling out what new insight they were hoping to discover. When the participant was not vocalizing, the experimenter reminded them to speak. The interviewers also encourage participants specifically to voice what attributes they were investigating and their goals. Though interviewers frequently encouraged participants to speak, they were not permitted to give hints or suggestions at any point in the exploration process.

Participant sessions were screen-captured and stored in a secured computer hard drive. As participants were required to share their screen at the start of the study, these videos contained audio and

[2] https://archive.ics.uci.edu/ml/datasets/adult
[3] https://github.com/propublica/compas-analysis/

video streams of both the participant and their current window. Following completion of all sessions, experimenters transcribed the video dialog into anonymized transcripts. Timestamps in the transcript were paired with views of their computer screen for analysis.

4.2 Analysis

Once the scripts were generated, two members of the research team went through each video independently. In the first pass of analysis, researchers tagged the video with timestamped codes for the specific analysis actions each participant took (e.g. forming a group, checking metric results, reading recommendations). For any activities where the two researchers did not agree, they negotiated until they arrived at one tag. In the second pass, researchers re-examined the videos and tagged activities. At this level of investigation, they examined the *sensemaking activities* of the participant using a schema derived from the notional model [54]. This process was akin to the multiple layers of process analysis commonly performed in contextual design in human-computer interaction [72].

The second phase analysis tracked several different elements in a participant session. First, the researchers listened for signals that the participant had formulated a hypothesis (or shifted back to a prior one). As participants were primed to mention their goals and strategies, signals for these shifts were present in the logs. Researchers then identified intermediate steps where participants were following up on hypotheses (either gathering evidence, testing, or drawing conclusions). Finally, researchers tagged the end of the session.

In a final pass, the researchers went through the activities and tagged higher level events and sketched a state model for each participant session. Participants all began at a starting node. As they used system features, they created new nodes for those activities. Groups of activities that were directed towards a specific hypothesis were grouped together in boxed regions. Transitions between nodes were directional, and looping behaviors as a result of iteration were reflected as cycles in the model. The models were then revised with improved layouts to emphasize the serial or parallel development of hypotheses during the session by considering timestamps (i.e. early events at the top, later events at the bottom).

Once the sketches were generated, the two researchers reconvened. Examining their participant process models, they developed a refined model together that reflected both of their sketches. Having build these models, they then extrapolated, pulling several models that were representative of patterns they observed for specific tools and datasets. Using this group of activity patterns, the researchers began identifying higher level themes that came through across participants. They referred back to the tagged activity logs and transcripts for further evidence of their themes, bringing in quotes, quantitative findings from analysis of the surveys/logs, and segments of activity. Refer to Figure 1 for one example of a log and Figure 2 for a finished summary process model. In the following section we identify the salient themes that emerged across tools and datasets from this process modeling exercise.

5 FINDINGS

Several patterns emerged as we examined the process diagrams and data from our comparison of AIF, Silva, and What-if. In the following subsections, we identify key themes from our data.

5.1 Exploratory Tools Invite Iteration

In Section 3, we identified how, despite having more interface actions and interactive components, an exploratory tool such as Silva still resulted in better outcomes compared to ground truth. We hypothesized that one reason for the higher level of performance was that the exploratory tool promoted iteration. While time consuming, perhaps this cycle of iterative improvement ultimately reached a higher optimum than the recommender system did. In examining the behavior of users during their think-alouds, we found evidence of iterative behaviors across the tools. However, the structure of the iteration was greatly shaped by the tool itself.

Figure 2 shows prototypical process models created as a result of our think-aloud studies. These figures aggregate common patterns shared across participants. Note the amount of cycles in the Silva model on the left. This is an indication that participants engaged in a variety of exploration behaviors. Rather than being routed towards one conclusion by a system, participants explored several lines of inquiry. While the size of the model is a good indicator also of the amount of effort the user had to put forth, the behaviors connect well with parts of theoretical models of sensemaking (e.g. [54]). Early in the process participants engage in foraging, rapidly developing simple hypotheses and testing them (H1-3). As time goes on their understanding of the data becomes more sophisticated. Towards the end of the exploration, the participants tended to branch out (H5-8) from a central view as they investigated ancillary lines of inquiry. This reflects the "explore vs. exploit" trade-off highlighted in sensemaking literature. While initially the user explores, as their understanding develops they switch to exploitation.

Contrasting this with the other two tools on the right of Figure 2, we find that they have a far more linear structure. We argue that this is a result of the automated affordances shaping the process. In the most extreme case, AIF pushes the participants towards a very linear workflow (upper left). The middle case, What-if, demonstrated some parallel hypotheses, but ultimately the flow was largely linear. While participants explored a number of hypotheses, we did not observe the same iterative development process - instead the participants were guided to different features by the system.

Examining the data quantitatively, we note a significant difference (t(14)=2.4532, $\rho < 0.0279$) between the average hypothesis that per participant made while using Silva (M: 5.4, Std: 4.92), AIF (M: 2, Std: 0) and Google What-if (M: 3.9, Std: 2.81). This is further evidence that exploratory tools promote hypothesis generation over iterations. In Silva, we found that one hypothesis almost always began from an open-ended visual element such as the causal graph and ended with results checking using metric views. This transition from identifying candidates to verifying metric changes lines up with previous observations about participants' iterative behavior. In examining event logs, we also note that many of the hypotheses towards the end of the exploration return to the same root attributes, providing more evidence for "exploit" behaviors.

From the perspective of the effectiveness of each system, participants using Silva (M: 0.54; Std: 0.28) had a significantly lower false discovery rate compared to those using AIF (M: 0.75; Std: 0.15) and Google What-if (M:0.89; Std: 0.32) (t(14)=2.2902, $\rho < 0.0320$). In particular, one participant generated 14 hypotheses when using Silva to analyze the Adult dataset and reached a nuanced final conclusion

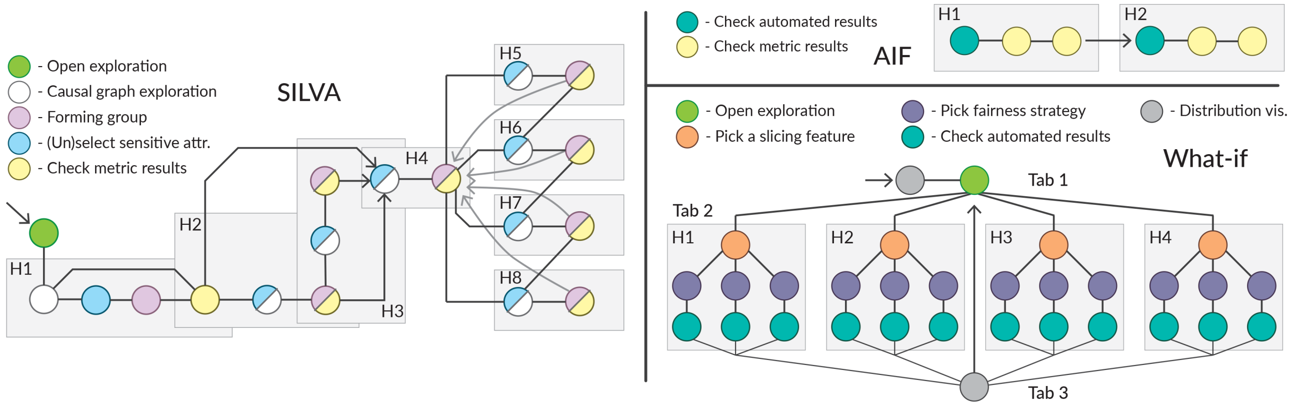


Figure 2: Common patterns in sensemaking (a) Sensemaking loops for Silva (left) tend to be hierarchical with iterative structures and overlapping hypotheses. (b) Sensemaking loops for AIF (top right) are linear, implying sequential information processing. (c) Typical sensemaking loops for Google What-If (bottom right) contain parallel branches: participants explore a certain hypothesis and then move on to the next, with little connection in between.

that was well matched to the ground truth. The participant found their end results surprising, as it did not reflect their initial assumptions about the dataset. This is consistent with the prior work that open exploration tools like Silva lead to better performance [75].

5.2 Recommendations Require Less Investment

Although open-exploration tools seemed to lead to better performance on the tasks, users' experiences with open-exploration tools were mixed in our think-aloud study. 3 out of 8 participants indicated that the exploratory tools were confusing and preferred the alternative recommendation tool they used to complete the other task. In other words, while exploration might lead to better outcomes, the cost of extra effort outweighs potential perceived benefits for users. We hypothesize that this is as much as case of exploration tools' effortfulness being a perceived disadvantage as it is a case of participants' not being able to accurately judge the potential benefits of a specific tool. This is a common issue in human-computer interaction, where system developers' understanding of the benefits and costs of a system does not align with the real-world perceptions of the tool.

Though Silva participants favored the causal graph and found it useful (6 out of 8 participants rated causal graph highly in the post-survey), those same participants reported a preference for AIF or Google What-if based on which one they used in their session. Yet, at the same time only 3 out of 12 participants were satisfied with the automatic recommendation results provided by AIF or Google What-if. Examining qualitative responses, participants were generally unsatisfied with the recommendations they received, though it is unclear if other confounding factors such as expertise may be at play here. Even though participants were not entirely satisfied with individual components of AIF or Google What-If, they seemed to find the tools valuable as a whole. Indeed, one participant went as far as to say that "I like that [AIF] wasn't trial-and-error; it just told me the results and gave an interpretable outcome". The survey data hint at a general pattern where straightforward demonstration was more favorable than "trial and error" free exploration, especially for those who weren't familiar with ML fairness or the datasets in question.

On a high level, this brings into focus the issue of satisficing [63]. In literature on sensemaking, satisficing is commonly invoked to highlight a central trade-off between effort and achieving "good enough" results. While in an ideal world an analyst would find the absolute optimum, in practice they satisfice for the best outcome given real-world constraints. We see some evidence of this in our think-aloud data. While participants benefited from the exploratory tool, the amount of effort it required went past the threshold for reasonable effort by the participants. The cost structure for recommendations was perceived as much more favorable, leading to the observed preference differences. As we move towards developing further tools in this domain, it merits consideration how best to convey the potential costs and benefits of a tool to future users so that they can make more informed judgments.

This issue is heightened when we consider skill and background knowledge. We hypothesized that inexperienced participants might have a harder time using open exploration tools, especially with regards to using complex tools such as the causal graph in Silva. We observed that participants using Silva spent on average 26% of their time on the causal graph. Still, 2 participants misunderstood the relationships represented by the causal graph. They did poorly on the survey questions and expressed frustration. Overall, participants who interpreted the causal graph correctly were able to reason about the dataset, provide more accurate answers, and also rated Silva higher than those who didn't. Though we do not have data for or against skill issues in the recommendation systems in our study, it is reasonable to suggest that interpreting recommendations accurately so as to minimize the risk of misinterpretation also requires a degree of expertise. In both cases, this indicates that adequate training to eliminate misunderstandings is also key to maximizing the benefit of ML de-biasing tools. This, however, requires a fair amount of initial investment on the part of the user in training and studying, which is also a factor in their perceived risk/reward considerations.

5.3 Information Overload When Balancing Exploration and Recommendation

While the open explorations tools led to better results, they were less favored by participants. On the other hand, participants favored recommendations but they proved less effective in bias detection tasks. In our preliminary investigation we observed initial signs of this dichotomy, and so we included Google What-if, a rough combination of the recommendation and exploration, in our think-aloud study. However, What-if ultimately failed to read a happy medium between the two poles.

Through event logs and post-study surveys, we observed that participants who used What-if spent an average of 34 seconds to analyze a pair of attributes (interpreting the plot and the distribution of datapoints, etc.). On average, they spent 8 minutes selecting features that might be helpful, but only explored 8.2 out of 14 attributes in the Adult dataset, and 3.9 out of 6 attributes in the Compas dataset. That is, participants covered only a small portion of all attributes while spending quite a bit of time in parts of the tool that exposed attributes. Such inefficiencies may have help participants back from gaining valuable insights. We observed that participants often found themselves overwhelmed by the huge number of potential next steps in the tool at any given moment. For instance, one participant reported, “I don’t think I can study all the attributes so I will start with the ones I thought was [sic] biased". However, they weren’t able to follow their plan as they spent the rest of the session only looking for relationships between age and other attributes. Besides exploring just a couple of features, some participants were on the other end of the extreme, trying exploit as many features as they could without adequate contemplation of what they were observing. For example, among the 4 participants who used What-If to analyze Adult dataset, 2 looked at over 10 attributes and the others examined no more than 4. This implies that there might be a two-sided issue at play here: while exploration tools can lead to choice overload, ready access to recommendations may actually exacerbate the overload by supplying more avenues for interaction and hooks for exploration.

The interface of Google What-if segmented bias detection into three separate stages where participants first edited data points, then moved to change the parameters of models, and finally detected unfairness. This interface design cut off the connection between each step and significantly enlarged the search space of users. Part of the low input-output ratio with What-If may be due to an explosion of pairwise slicing and plotting combinations (e.g. a search space of size $2^{15} = 32768$ for the Adult dataset). By contrast, the input-output ratio is higher for Silva, a free-exploration system. On average, participants using Silva formed 5.4 groups for the Adult dataset, covering 12.0 attributes out of 14 attributes (85.7%), and 2.0 groups for the Berkeley dataset, covering all the attributes in that dataset.

In addition, participants had to switch between different tabs in the hybrid system. A typical Silva session was iterative, leading to a hierarchical sensemaking loop. By contrast, when exploration was parallelized, as we observed in What-If (see Figure 2), sensemaking is forced into parallel branches that don’t lend themselves to iterative improvement. Our think-aloud studies provide evidence that hierarchical patterns of exploration could potentially reduce information overload in complex de-biasing tasks, albeit with potentially an issue in bootstrapping at the very start of a task.

6 DISCUSSION AND RECOMMENDATIONS

Account for expertise in exploration and recommendation: We noticed a conflict between user preferences and system effectiveness, finding that participants who didn’t fully understand the causal graph in Silva reported that the tool was not useful. Exploration requires practitioners to have a fair amount of basic knowledge, otherwise the barrier to entry is to great. Yet, initial training investments may pay dividends. On the other hand, recommendations, while superficially easier to read, may be subject to misinterpretation or bias if training is lacking. In both of these cases, expertise is a critical design criterion. While the consequences manifest differently in both cases (direct dislike in exploration, subtle biases in recommendations), the impact is potentially great. For this reason, it is crucial that designers of de-biasing systems adequately test and account for the skill level of their users and choose affordances that best reflect that balance.

Think-aloud as a method for evaluating de-biasing systems: While the preliminary analysis showed that a great deal of performance information could be derived from traditional satisfiable user studies, our think-aloud demonstrated the kinds of nuances that qualitative investigation can expose in terms of usability and effectiveness of interactive systems. While this is certainly not an unknown in human-computer interaction, we emphasize that understanding the mental model and process of participants using a de-biasing tool is critical, lest biases go missed or re-incorporated into the pipeline.

Motivating efficient exploration through hybridization: We noticed that combining exploration and recommendation risks blowing up the number of options in a tool, leading to choice overload among users. A simple combination of both approaches when dealing with complex, multidimensional datasets may not be sufficient. However, there is a possible resolution: connecting individual interface components closely, reducing the costs of switching between subtasks during iteration, and strategically incorporating recommendation when bootstrapping or exploitation are involved.

Balancing tuning and broader contexts: While the Silva tool abstracted away the details of model parameter tuning, essentially making the classifier a black box to users, Google What-If exposed the technical details of the model including editable classification thresholds, applicable optimization/regularization strategies, and a datapoint editor. While this design gets users more involved in model development, especially for parameter tuning , it directs users’ attention away reasoning about bias. Given the concrete way in which parameters can be tuned and optimized, this may be a tantalizing distraction versus de-biasing which is, as our think-alouds showed, a complex and inter-dependent task which requires effort. There is certainly a place in ML fairness systems for the nuts and bolts of ML development, but we emphasize that these affordances risk acting as a distraction from the harder sensemaking tasks which expose biases (both implicit and explicit) and help users develop knowledge about how their data work that might be transferred to future tasks. In the future we hope to investigate how to structure ML fairness tools so that they can strike the right balance of surfacing model mechanics while promoting higher level reasoning about fairness.

6.1 Limitations

Our study is subject to a number of limitations. Foremost, our conclusions might have been influenced by the limited sample size and randomized assignment. Participant-level noise may have confounded differences we observed. Additionally, the extent to which participants' skill level affected their preference and performance remains unclear. In our preliminary log analysis, we noted that skill level was not a factor when comparing participants' understanding of the bias in the datasets. However, we did note that more skilled participants were more likely to interpret causal relationships correctly and react to interactive graphs more quickly. The current sensemaking patterns we characterized will be informed by future studies that capture a wider variety of skill levels with greater number of participants. In terms of methodology, our think-alouds were necessarily limited by the remote study structure and our selection of qualitative analysis techniques. It is likely that different strategies for processing the variety of data coming out of the think-alouds might emphasize different features in the data (for example, exposing subtle effects of skill versus our focus on process-level details). This is something we endeavor to explore as we work with more participants and develop interactive de-biasing tools.

7 CONCLUSION

This paper aimed to answer a fundamental question: *How does the design of de-biasing systems affect how ML practitioners reason about biases?* Through a preliminary study, we identified how exploratory and recommendation tools might evoke different responses. Through the online think-aloud study, multiple surveys and semi-structured interviews, we synthesized process-level details about three different de-biasing tools, identifying key distinguishing features among their interface affordances and connecting them to differences in participant sensemaking. We find that exploratory tools tend to invite iteration, while recommendation tools requires less investment. Tools that attempt to balance both exploration and recommendation risk overloading users. Finally, we provided design recommendations for future interactive de-biasing systems.

REFERENCES

[1] 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[4] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[5] Morten Sieker Andreasen, Henrik Villemann Nielsen, Simon Ormholt Schrøder, and Jan Stage. 2007. What happened to remote usability testing? An empirical study of three methods. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1405–1414.

[6] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.

[7] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[8] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[10] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[11] Peter J Bickel, Eugene A Hammel, and J William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404.

[12] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. 149–159.

[13] 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. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.

[14] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. [n.d.]. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Technical Report. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://www ....

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[16] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting student emotions in computer-enabled classrooms.. In *IJCAI*. 4125–4129.

[17] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (2017), 30–44.

[18] Stuart Card, JD Mackinlay, and B Shneiderman. 2009. Information visualization. *Human-computer interaction: Design issues, solutions, and applications* 181 (2009).

[19] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating classifier error discovery through interactive semantic data exploration. In *23rd International Conference on Intelligent User Interfaces*. 269–280.

[20] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[21] Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times* 25 (2016).

[22] Kate Crawford. 2017. Artificial intelligence with very real biases. *The Wall Street Journal. Retrieved from https://www. wsj. com/articles/artificial-intelligencewith-very-real-biases-1508252717* (2017).

[23] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[24] Jeffrey Dastin. 2018. Rpt-insight-amazon scraps secret ai recruiting tool that showed bias against women. Reuters, 2018.

[25] Michael A DeVito, Jeremy Birnholtz, and Jeffery T Hancock. 2017. Platforms, people, and perception: Using affordances to understand self-presentation on social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. ACM, 740–754.

[26] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[27] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.

[28] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[29] Alex Endert, William Ribarsky, Cagatay Turkay, BL William Wong, Ian Nabney, I Díaz Blanco, and Fabrice Rossi. 2017. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 458–486.

[30] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*. 160–171.

[31] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "

I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on Human Factors in computing systems*. 153–162.

[32] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.

[33] Google. 2017. What-if Tool. https://pair-code.github.io/what-if-tool/

[34] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13.

[35] Bernard E Harcourt. 2008. *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

[36] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[37] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 600.

[38] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 259–268.

[39] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated machine learning: methods, systems, challenges*. Springer Nature.

[40] Robin Jeffries, James R Miller, Cathleen Wharton, and Kathy Uyeda. 1991. User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 119–124.

[41] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3819–3828.

[42] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[43] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent tradeoffs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[44] Tim Kraska. 2018. Northstar: An interactive data science system. *Proceedings of the VLDB Endowment* 11, 12 (2018), 2150–2164.

[45] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[46] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).

[47] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[48] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[49] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383* (2018).

[50] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[51] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[52] Besmira Nushi, Ece Kamar, Eric Horvitz, and Donald Kossmann. 2017. On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems. In *AAAI*.

[53] Helen Petrie, Fraser Hamilton, Neil King, and Pete Pavan. 2006. Remote usability evaluations with disabled people. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1133–1141.

[54] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[55] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[56] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.

[57] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. 2015. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 240–249.

[58] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[59] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 99–106.

[60] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[61] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*. 2503–2511.

[62] Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*. IEEE, 336–343.

[63] Herbert A Simon. 1990. Invariants of human behavior. *Annual review of psychology* 41, 1 (1990), 1–20.

[64] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.

[65] Aaron Springer, Jean Garcia-Gathright, and Henriette Cramer. 2018. Assessing and Addressing Algorithmic Bias-But Before We Get There.... In *AAAI Spring Symposia*.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[67] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. ACM, 440.

[68] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115.

[69] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. 2016. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the {US}. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*. 271–285.

[70] Karl E Weick. 1995. *Sensemaking in organizations*. Vol. 3. Sage.

[71] Etienne Wenger. 1999. *Communities of practice: Learning, meaning, and identity*. Cambridge university press.

[72] Dennis Wixon, Karen Holtzblatt, and Stephen Knox. 1990. Contextual design: an emergent view of system design. In *Proceedings of the SIGCHI Conference on Human Factors in computing systems*. 329–336.

[73] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 656.

[74] Tongshuang Wu, Daniel S Weld, and Jeffrey Heer. 2019. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 4 (2019), 1–27.

[75] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[76] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32, 4 (2019), 403–414.

[77] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[78] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[79] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

APPENDIX

Details of Participant Allocation

To ensure each tool and dataset receives even exposure, we set up the experiment as shown in Table 1, counter-balanced for order effects. The two datasets Silva helps to analyze are Adult and Berkeley. The sample datasets we included for AIF demo are Adult and Compas. The datasets Google What-If uses are Adult and Compas. Each tool was used by exactly 8 participants in our think-aloud study. In total, there are 12 unique experiment conditions. Each participant was then randomly assigned an experiment condition.

Participant ID	Tool1 (dataset1)	Tool2 (dataset2)
P1	Silva (Adult)	What-If (Compas)
P2	What-If (Compas)	Silva (Adult)
P3	Silva (Berkeley)	What-If (Adult)
P4	What-If (Adult)	Silva (Berkeley)
P5	Silva (Adult)	AIF (Compas)
P6	AIF (Compas)	Silva (Adult)
P7	Silva (Berkeley)	AIF (Adult)
P8	AIF (Adult)	Silva (Berkeley)
P9	What-If (Adult)	AIF (Compas)
P10	AIF (Compas)	What-If (Adult)
P11	What-If (Compas)	AIF (Adult)
P12	AIF (Adult)	What-If (Compas)

Table 1: Details of Participant Allocation

Additional User Activity Log Analysis

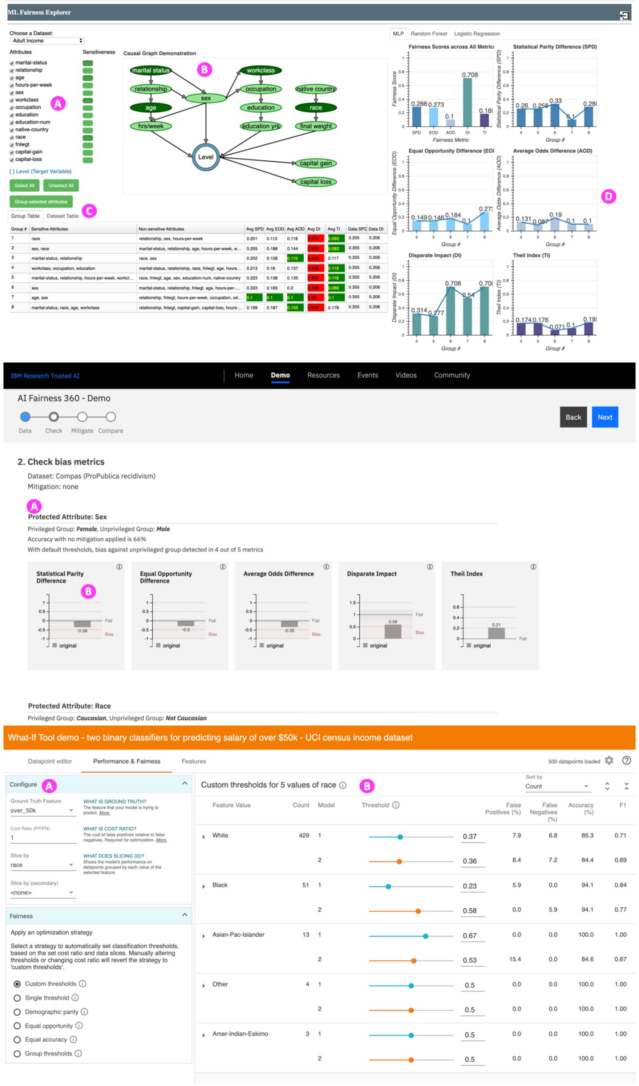
In the preliminary log analysis, we captured significant user activities in each session. In addition to time spent on the causal graph and total number of groups formed, we also looked at various metrics like inter-group operations and the structure of each group. Those statistics are summarized in Table 2.

Metric	Mean (Berkeley)	Std (Berkeley)	Mean (Adult)	Std (Adult)
# of operations on the causal graph	11.00	5.69	34.45	35.23
# of groups formed	2.00	0.71	5.40	4.92
# of operations between groups	5.13	1.51	6.32	3.03
# of sensitive attributes in groups	1.00	0.00	1.58	0.88
# of non-sensitive attributes in groups	1.50	2.40	7.34	4.80
total # of operations	13.00	6.20	39.85	39.89
duration (s)	566.17	942.92	536.95	582.95

Table 2: Summaries Statistics of User Event Logs

Examples of Debiasing tools' interfaces

Figure 3 shows the example interfaces of three assessment systems that have been evaluated in this paper. The top one is Silva, with four major components: a Dataset Panel (A), a Causal Graph view (B), a Table Group (C), and a Fairness Dashboard (D). The middle one is AIF, which has two major components: a Results Explanation(A) and a Visualization Dashboard(B). The bottom one is Google What-if, which has two major components: a Fairness Configuration (A) and a Model Parameter Tunning (B).



Group #	Sensitive Attributes	Non-sensitive Attributes	Avg SPD	Avg EOD	Avg AOD	Avg DI	Avg TI	Data SPD	Data DI
1	race	relationship, sex, hours-per-week	0.201	0.113	0.118	0.474	0.093	0.355	0.206
2	sex, race	marital-status, relationship, age, hours-per-week, w...	0.255	0.188	0.144	0.598	0.083	0.355	0.206
3	marital-status, relationship	race, sex	0.202	0.158	0.115	0.453	0.117	0.355	0.206
4	workclass, occupation, education	marital-status, relationship, race, fnlwgt, age, hours...	0.213	0.16	0.137	0.466	0.119	0.355	0.206
5	marital-status, relationship, hours-per-week, workcl...	race, fnlwgt, age, sex, education-num, native-country	0.223	0.138	0.135	0.455	0.118	0.355	0.206
6	sex	marital-status, relationship, fnlwgt, age, hours-per-...	0.333	0.169	0.2	0.728	0.086	0.355	0.206
7	age, sex	relationship, fnlwgt, hours-per-week, occupation, ed...	0.1	0.1	0.1	0.68	0.1	0.355	0.206
8	marital-status, race, age, workclass	relationship, fnlwgt, capital-gain, capital-loss, hours-...	0.169	0.187	0.155	0.636	0.178	0.355	0.206

AI Fairness 360 - Demo

2. Check bias metrics

Dataset: Compas (ProPublica recidivism)
Mitigation: none

Protected Attribute: Sex

Privileged Group: ***Female***, Unprivileged Group: ***Male***
Accuracy with no mitigation applied is 66%
With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

Protected Attribute: Race

Privileged Group: ***Caucasian***, Unprivileged Group: ***Not Caucasian***

Ground Truth Feature: over_50k
WHAT IS GROUND TRUTH? The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1
WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: race
WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Slice by (secondary): <none>

Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

● Custom thresholds
○ Single threshold
○ Demographic parity
○ Equal opportunity
○ Equal accuracy
○ Group thresholds

Custom thresholds for 5 values of race

Feature Value	Count	Model	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
White	429	1	0.37	7.9	6.8	85.3	0.71
		2	0.36	8.4	7.2	84.4	0.69
Black	51	1	0.23	5.9	0.0	94.1	0.84
		2	0.58	0.0	5.9	94.1	0.77
Asian-Pac-Islander	13	1	0.67	0.0	0.0	100.0	1.00
		2	0.53	15.4	0.0	84.6	0.67
Other	4	1	0.5	0.0	0.0	100.0	1.00
		2	0.5	0.0	0.0	100.0	1.00
Amer-Indian-Eskimo	3	1	0.5	0.0	0.0	100.0	1.00
		2	0.5	0.0	0.0	100.0	1.00

Figure 3: Example Interfaces of Silva, AIF and What-if(From the top to the bottom)