Silva: Interactively Assessing Machine Learning Fairness Using Causality

Jing Nathan Yan, Ziwei Gu, Hubert Lin, Jeffrey M. Rzeszotarski
Cornell Univesity
Ithaca, United States
{jy858, zg48, hl2247, jmr395}@cornell.edu

ABSTRACT
Machine learning models risk encoding unfairness on the part of their developers or data sources. However, assessing fairness is challenging as analysts might misidentify sources of bias, fail to notice them, or misapply metrics. In this paper we introduce Silva, a system for exploring potential sources of unfairness in datasets or machine learning models interactively. Silva directs user attention to relationships between attributes through a global causal view, provides interactive recommendations, presents intermediate results, and visualizes metrics. We describe the implementation of Silva, identify salient design and technical challenges, and provide an evaluation of the tool in comparison to an existing fairness optimization tool.

Author Keywords
Machine learning fairness; bias; interactive systems

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI);** *Haptic devices;* User studies;

INTRODUCTION
Machine learning has been introduced into domains such as health-care[10, 22], internet search[35], market pricing[16, 27], and policy [21] with the goal of reducing costs and improving accuracy in decision-making. However, these data-driven applications risk silently introducing societal biases into the decision-making process. For example, a recent analysis of a recruiting system at Amazon [17], trained on hiring data collected during a 10 year window, found that gender biases encoded in the model were inadvertently incorporated in the hiring process as a whole. Unable to convincingly resolve all potential biases, Amazon abandoned the system. Similar examples make evident the need to study fairness in data-driven systems [2], and it is now a crucial component in many workflows. Central to this is machine learning system practitioners' ability to accurately and efficiently assess fairness.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376447

The machine learning community has focused on statistical definitions to quantify fairness[28, 20, 15]. Given the complexity of bias, many metrics [6] have been proposed. For example, disparate impact is used to evaluate positive outcomes for a privileged group against an unprivileged group. Recent research has exposed usability flaws in metric-driven approaches [54] – given the large number of metrics, practitioners tended to over-calibrate to intuitive metrics regardless of suitability of others. Automatic toolkits have been proposed to resolve this "metric burden" in assessing fairness. However, juggling metrics can be very challenging. There is often a catch-22: existing proposed metrics can be mutually exclusive, which means conclusions drawn with one metric could be contradictory with those drawn from another. [37, 41]. The choice and application of fairness metrics alone may be insufficient without a deeper understanding of the data and problem.

One avenue for improving how practitioners make sense of fairness metrics and their data is to employ causality to help triangulate on sources/causes of bias. Recent research has used causal relationships (e.g. the influence of height on weight) to help individuals reason about sources of bias [36, 69, 46]. By looking into causal relationships, one might track how hypothetical attributes influence one another and potentially convey bias in a dataset. Further, relationships might exist between unexpected attributes that ought to be considered. Yet, as with metrics, deciding on whether a particular influence path is socially acceptable or fair requires deeper investigation. As social conventions evolve over time, automatic results without carefully encoded social awareness risk reaching incorrect or biased conclusions (as was the case with Amazon's hiring system). As a result, though comprehensive causal information may help to inform an analysis, it also may require burdensome training and manual analysis time to properly evaluate.

In this paper we present Silva, an interactive system that uses causality to help individuals assess machine learning fairness effectively and efficiently. Silva allows users to interactively diagnose sources of bias to improve fairness in data by helping users to integrate their own social awareness and domain expertise when making fairness decisions using metrics. Silva helps provide additional context for users, assisting them in delineating the impact of bias by connecting bias sources and existing popular metrics through causal relationships between data attributes. Causality not only provides additional context for users when employing metrics, but also helps to expose

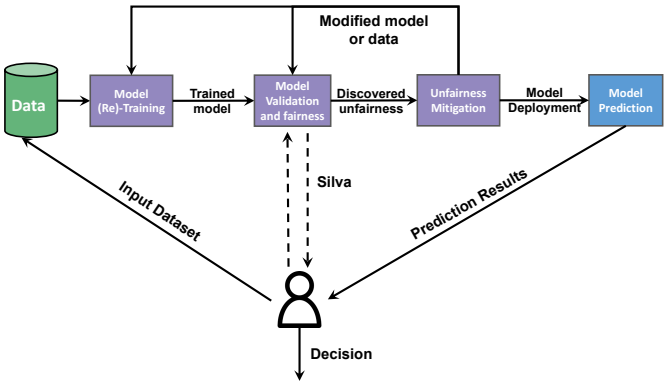


Figure 1. Example Working Pipeline of ML system Practitioner

hidden or unexpected relationships that may be meaningful in evaluating fairness holistically. Within a broader machine learning pipeline (Figure 1), we view Silva as promoting interactive, high-feedback investigations during the model validation phase. Much as in traditional sensemaking processes [48], tightening the assessment loop by exposing more information and reducing costs of investigation might help users better integrate their own social awareness and more deeply investigate sources of unfairness.

We have several aims in this work: First, we examine how tools can help users understand complex causal relationships which lead to social bias in data through visualizations. Causality discovery algorithms have long been studied, and probabilistic graphical models [38] visualize causal relationships through network diagrams. However, as complexity increases, traditional node-link diagrams can become difficult to interpret. We explore the feasibility of causal visualizations for fairness evaluations and identify ways in which automatic node highlighting and hiding may improve their utility. Second, we consider how allowing users to explore "what if" questions enhances their ability to draw useful conclusions. In contrast to existing tools which limit users to predefined definitions of unfairness, our design allows users to explore potential sources of bias in a sandbox. Key here is understanding how users track their progress and use affordances for storing, referring to, and comparing between scenarios found while exploring. Finally, we consider how tools like Silva may be integrated into a broader pipeline. Silva's affordances for machine learning model training, causal graph view of the data, group comparisons of user-selected subsets of data, and grouped metric visualizations might be incorporated into a larger workflow.

Our work offers three core research contributions:

- We present Silva, an interactive sandbox environment that uses causality linked with quantitative metrics to help individuals assesses machine learning fairness.
- We develop and study an interactive user interface and causal graph visualization to help users ask hypothetical "what if" questions as they examine causal paths.
- We present results of user studies which demonstrate the effectiveness of Silva over comparable systems. Silva users efficiently detected sources of social bias in datasets.

RELATED WORK

Machine learning fairness has drawn attention from a variety of fields including machine learning, HCI, databases, and statistics. In general, the machine learning community has focused on novel statistical definitions, metrics to quantitatively measure the fairness of algorithms and datasets, tools for optimizing these metrics, and reasoning about sources of (un)fairness. At the same time, the HCI community has examined the causes, sources, and consequences of fairness as it relates to socio-technical systems, policy, and psychology in the real world. An increasing interest has developed among both of these communities towards investigating approaches that make algorithmic ML tools more usable or robust. Connecting to this broader investigatory area, Silva combines metrics and approaches from the ML community with traditional usable interface development from the data visualization and HCI communities. In this section we will explore related work within these various communities.

Understanding Fairness

Emerging applications of machine learning systems for decision-making across a wide range of domains [2] (e.g., marketing [16, 27], policy [21] and search engine results [35]) have drawn much attention towards the implications of their judgments and dependence on potentially biased training data. As systems become increasingly integrated into domains not traditionally associated with machine learning, researchers have identified cases where models have marginalized groups or otherwise unfairly influenced decisions. Researchers have explored patterns underpinning cases of under-representation [3, 24, 44], scrutinized existing systems to assess how they handle unfairness, and explored the challenges of managing unfairness [2, 12, 24]. For example, researchers identified how image search results amplified stereotypes towards race [35]. Credit scoring systems have been examined to expose implicit discrimination [57]. With the rise of data privacy legislation and policy interests in data storage [25], attention has also been drawn to how populations are affected by unfairness [63, 49], and the difficulties of resolving unfairness [29, 60].

Machine Learning Fairness

The machine learning community has developed many statistical definitions of fairness for both data and models [28, 20, 15, 6]. These measures of fairness quantify biases in decisions (such as hiring or salary assignment) with respect to different groups. Minimizing unfairness in data or in learned models ought to reduce the impact of unfair biases in (semi-)automated decision making. In general, fairness is achieved through (conditional) independence between sensitive attributes S, prediction O, and some target variables Y. However, these metrics can be mutually exclusive [41, 37], causing confusion to users if contradictory results are shown. Further, machine learning system practitioners report that that existing statistical definitions fail to meet their expectations [8, 42] in terms of relating the results to fairness. Additionally, [54] examined user attitudes towards unfairness and concluded that, in reality, calibrated fairness is more preferred compared to multiple statistic definitions which might lead to misapplication or misinterpretation.

The lack of ubiquity and generalization of metrics has motivated investigations of machine learning fairness through the lens of causal reasoning [36, 69, 43, 40, 46]. Causal reasoning attempts to relate how attributes influence other attributes (e.g. height influences weight – taller people tend to weigh more). Causal reasoning can reveal sources of bias that arise from such relationships between attributes. Nabi et al. [46] presented causality to users through a graphical model, and proposed path-specific fairness in which paths between sensitive attributes and output attributes are blocked. This approach proved to be intuitive for users. Unfortunately, path-specific fairness requires strong assumptions to compute automatically which are rarely feasible in practice [52].

Systems for Improving Fairness

There are two general patterns in systems intended to improve fairness. On one hand, some systems try to optimize for metrics and automatically deliver improved results. On the other hand, some systems try to enable interactive exploration.

Optimization

The machine learning community has focused on mitigation of unfairness at different stages of the machine learning pipeline. There are two general threads of research. The first examines ways to improve fairness by optimizing machine learning algorithms [23, 32, 32, 33, 34, 67]. These are model-specific or algorithm-specific approaches. The second thread [13, 28, 64, 50, 14, 55, 68, 1] applies optimizations during the pre-processing or post-processing stage of the machine learning pipeline. These methods are not tied to specific models, but may be over-tailored to specific datasets. [64] considered specific machine learning methods and incorporated fairness metrics for a fair prediction which may not be sustainable with other machine learning algorithms and existing metrics. [13] proposed a convex optimization to transform the dataset to remove bias and treats learning algorithms as block boxes. However these methods fail to discover bias sources, do not integrate up-to-date social awareness (which informs which biases are unacceptable), and can be hard to balance.

Automated Systems

A number of hybrid automated and interactive systems exist. IBM AI Fairness 360 (AIF) [4] is an automatic system that identifies model or dataset biases based on existing fairness metrics and employs bias-reducing algorithms (see above) to reduce unwanted model bias. Google's What-If tool [26] incorporates human interaction by providing visualization of data features and hooks for programmatic mitigation of bias. However, [19] highlight that many mechanisms employed by automated systems encode assumptions which may not hold true in all data and model contexts. Further, [11, 60, 62] suggest that end-users may have misconceptions of the techniques at play, and as the result the underestimate of the effect of unfairness on underrepresented group or the implications of using an automated tool to correct their data.

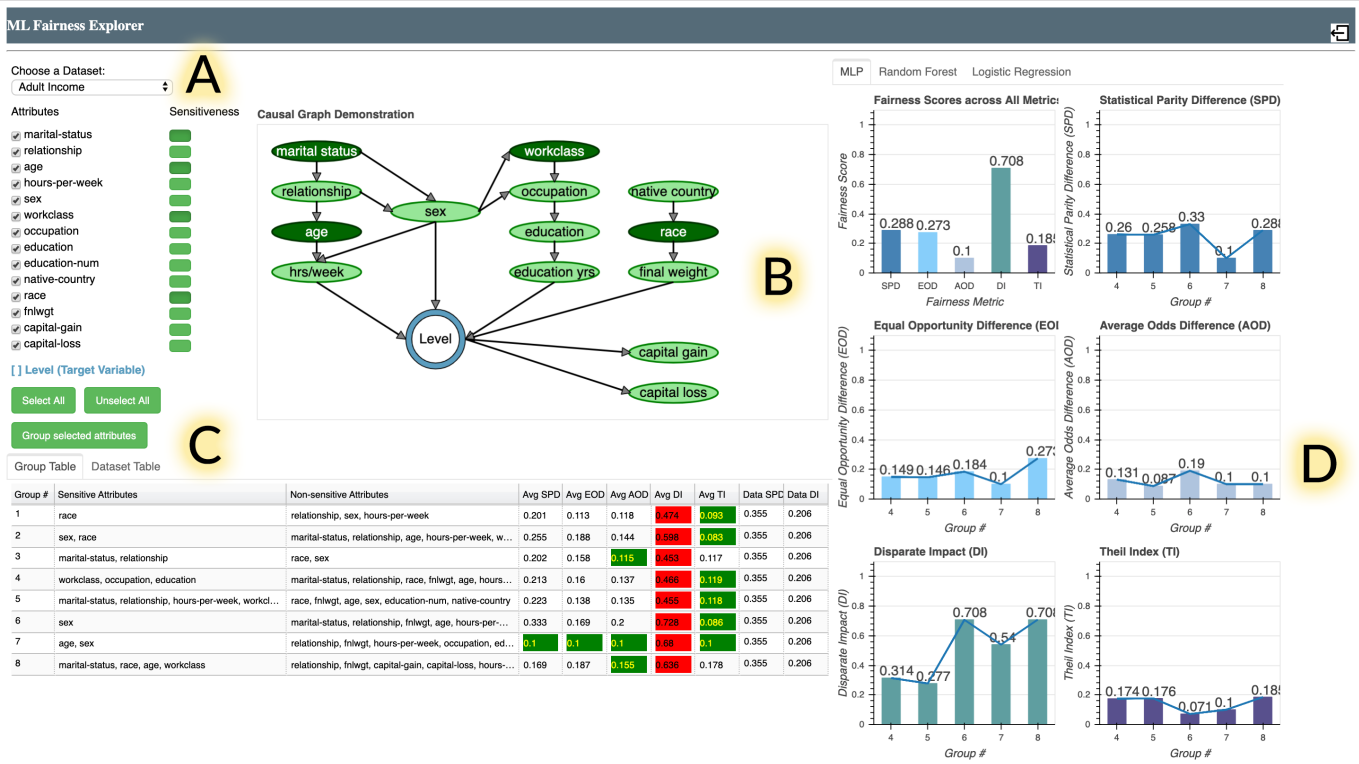
Interactive Systems

Interactive systems provide real-time feedback in response to human input, and in a data science context are often employed improve the sensemaking process [48] of analysts. HypDB[51] is designed to help users understand causality. In particular it can help users understand and resolve the Simpson paradox[61]. Northstar[39], another interactive data analysis system, protects users from false discoveries and makes advanced analytics and model building more accessible by allowing users to focus on contributing their domain expertise without having to take care of technically involved tasks. AnchorViz[56], an interactive visualization that integrates human knowledge about the target class with semantic data exploration, supports discovering, labeling, and refining of concepts. Those recent systems as well as sensemaking theory suggest that interactive interfaces have the potential to close the gap between automated optimization-focused approaches and human understanding when designed effectively.

Causal Fairness

Throughout this paper we refer to causality. In this subsection, we provide a brief outline of causality concepts and literature. Causality refers to causal relationships between variables (i.e. one variable *causes* another). For example, a patient choosing to smoke has a causal relationship with the chance that they later will be diagnosed with cancer [47]. Causal relationships are sometimes self-evident to analysts, but in many cases they can be counter-intuitive or reflect biases in a dataset that ought to be examined (e.g. a causal relationship between gender identity and salary would often be considered socially undesirable, and in the case of a hiring system would be important to notice). There are many approaches to inferring the causality[38, 31]. One common way to express causality is based on graphical model which presents the causal relationship as a directed acyclic graph (DAG). The causal DAG represents each variable as a node and leverages the direct edges between nodes to model the interaction[45]. The AI community has long worked to infer DAGs from raw data[58, 30, 30]. Causal fairness is achieved if the given protected/sensitive attributes have no causal relationship to the final outcome, and could be indicated by a lack of paths from sensitive attributes to outcome variables in the DAG structure.

Another advantage of employing the causality is to enforce some *do operations* which change the state of portions of a causal model while keeping the rest of model the same (e.g. probing a salary model by adjusting past work experience). Hence, the effects on a particular outcome[47] can be observed, akin to what-if analysis. This has motivated the idea of counterfactual fairness: actively modifying the value of sensitive attributes' *do*-operation to search for conditions in which sensitive attributes will be not a cause of the outcome (e.g. under what conditions does gender identity no longer have an influence over salary in this dataset). However, [53] shows that counterfactual fairness is obtained with strong assumptions which are not always grounded in practice. As mentioned before, path-based approaches face similar tractability issues. Recent work [65] introduces a new idea for causality fairness by taking the biased paths between of one DAG as the input to generate more paths. However, the effect of this approach is largely based on the configuration of the input paths. For the purposes of Silva, we encourage users to consider counterfactuals and the influence of attributes through visual representations of causality.



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	0.483	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

Figure 2. Silva's main interface. (a) Dataset Panel allows users to select attributes for training classifiers and toggle the sensitivity of selected attributes. (b) Causal graph generated by structural learning algorithms depicts the causal relationships among attributes. (c) Table Group displays information on user-defined groups and the training dataset in the Group Table and Dataset Table, respectively. (d) Fairness Dashboard contains groups of bar charts showing fairness values across models, metrics, and groups. The four components work together to aid users in the bias exploration process.

SILVA

Before we outline our core design rationale, we illustrate one intended real-world use case for Silva:

Alex is a government-employed data scientist who is analyzing a crime dataset with 5 attributes: biological sex, age, race, prior counts, and charge degree. Alex has a series of machine learning models trained on a subset of the 5 attributes to predict whether a person will reoffend (commit another crime). Before applying the classifier in real-world scenarios, Alex needs to make sure the model is fair. More specifically, they need to find out which attributes may introduce significant bias to the classifiers and reason about the source of these biases.

Alex loads up Silva and imports their dataset (Figure 2). They immediately see their data reflected in the interface. In the Causal Graph visualization, a few recommended initial attributes to explore are shown (not depicted). Alex hypothesizes that "sex" may be a sensitive attribute, and that age is an important attribute for classifier accuracy. They select "sex" and "age" by clicking on the checkboxes in the Dataset Panel and saving these as a group. The fairness scores are displayed on the bottom in the Table Group and in the plots on the right in the Fairness Dashboard.

Alex notices that the Causal Graph has been updated with recommendations for two previously unselected nodes – "prior counts" and "charge degree". The arrow from "sex" to "prior counts" shows that the attribute "sex" influences the number of prior counts which in turn determines whether a person will reoffend. This is surprising to Alex, who initially believed that "sex" directly determined whether a person will reoffend. As a result, they mark the attribute "sex" as sensitive by clicking on the toggle. Immediately, the node representing "sex" turns dark green as a way to draw attention. Alex forms a group again and notices a decrease in 4 of the 5 fairness values.

Alex hovers on the node for "prior counts" and "charge degree" to view the median and variance of the two attributes, picking the former because of its higher variance (which can indicate that it encodes impactful information). When they mark it as sensitive, its two parent nodes "sex" and "age" are highlighted. A summary message under the graph reminds Alex that those two attributes might have caused "prior counts". After looking at the data table, they also include "race" in the dataset. Three other groups are formed along the way, and Alex is now ready to dig into the factors influencing the fairness scores.

Having chosen 5 different groups, Alex now looks at the bar chart grid which plots fairness values for all of their groups. Alex's random forest models perform best in this dataset, so they mainly focus on results in the second panel. Alex hovers on the overview charts to review the definitions of the 5 metrics, and decides that Theil Index, a measure of segregation and inequality, is the most relevant for their particular use case. They sort the groups based on Theil Index by clicking on the column header, and click on the first row to see bars

corresponding to the highest-value group more clearly. They also note that a bar in the Equal Opportunity Difference (EOD) charts is particularly low. Curious about the reason behind it, they hover on that bar to view sensitive and non-sensitive attributes included in that specific group. When they click on it, all the bars that correspond to that group in the dashboard area are highlighted to facilitate comparison of fairness values. The corresponding row in the group table is also highlighted, making it easier for Alex to locate that group in the table and read more about its composition, other fairness values, and its ranking in terms of Theil Index. Now with a deeper understanding of their dataset, Alex feels prepared to report their findings about potential biases that may make this dataset systematically unsuitable for their organization.

Core Design Rationale

Through Silva, we aim to help individuals assess machine learning fairness effectively and efficiently. At its core, Silva encourages users to openly explore their data and experiment. Causality acts both as another data channel and a means to promote reflection on part of the analyst. Quantitative measures help to ground the investigation and provide comparison points. We had 6 goals in mind when designing Silva:

Connect Causality and Statistical Metrics

We view causality and fairness metrics as providing overview and detail for a user. Fairness metrics give attribute-specific feedback, but may lack holistic context for accurately interpreting results and reasoning about sources of bias. The causal graph provides this overview for inter- and intra-attribute issues. The causal graph ought to help the user track sources one-by-one to their root causes if metrics contradict each other. Additionally, Silva could help to trace and exclude unavoidable sources of unfairness through the its recommendation mechanic. By showing unexpected causal relationships through recommendations, it may provoke analysts to think introspectively about societal or implicit biases. This connection between metrics and causality must be fluid and implicit for the user, and is emphasized through shared elements and redundancy in the Silva interface.

Explore Freely

Compared to existing tools like AIF and Google What-If, Silva offers users more freedom to investigate each attribute in a dataset through customized groups over multiple iterations. Instead of looking at the protected attributes provided by a black-box, users have an opportunity to identify and reason about sensitive attributes themselves. In addition to providing a overview of causal and attribute data, we preserve local structures in the causal graph so that users can examine details that may be important in their analysis. We believe that active engagement in the model validation process might help to bridge the gap between users and the bias mitigation process.

Link Views

The four distinct components of Silva (see Figure 2) work together to support users as they search for bias and unfairness. When possible, we design tools so that they link to one another (much as in Attribute Explorer [59] and other dynamic querying systems). For example, nodes in the causal graph correspond to attributes in the dataset and in the dataset panel. Changes or highlights in one view are reflected in the others. In addition, the same set of operations can be performed both within multiple views in the tool, adding redundancy. Similarly, the group table and fairness dashboard are also connected to help users integrate data in both parts: fairness values corresponding to the same group are highlighted across the tool as a means to facilitate comparisons.

Encourage Connections and Comparisons

One of the major goals of Silva is to bring different factors and measures of machine learning fairness together into one interface so that connections among them are more obvious to users. We allow arbitrary grouping of attributes in a dataset to help users to track bias among different sets of attributes over the course of their exploration. For each attribute group that users generate, we calculate 5 fairness metrics across 3 different machine learning models. In other words, we compute 15 different fairness values for every single group and map them to plots. Users are free to look at a specific plot that interests them the most, but they are also given the power to compare fairness scores along different dimensions. Further, they can recover past groups if they wish to compare to prior moments in their analysis.

Guide the Extraction of Insights

Besides helping users to explore connections, we also provide informative hints and annotations consistent with users' current phases of exploration to keep them on the right track. We keep in mind that users of Silva might come from vastly different backgrounds and thus we add numerous tools to help people of all levels succeed in their tasks, including pop-up definitions for fairness metrics and summaries of causal relationships. Further, in an attempt to reduce excessive time spent on trivial attributes or unproductive elements, we include recommended nodes, path, and next steps in the causal graph to help users focus their effort on the attributes that matter more to the results. We also show detailed messages when users make an illegal move such as forming a group without any sensitive attributes to correct their mistakes.

Extend, Not Replace

Silva can be easily integrated into existing machine learning pipelines and coupled with existing tools. Silva might provide reliable input suggestions to power users' applications of "What-If" and AIF. Both tools offer excellent visualization and bias mitigation solutions. However, the steps that lead to their choices of "protected attributes" seem to be hidden from users. If users are able to identify the attributes through Silva before using "What-if" or AIF, then they will likely gain more insights from these tools. In addition, Silva can also be extended into automated machine learning or bias mitigation pipelines.

Implementation

Silva was implemented as a back-end web application using the Bokeh [9] visualization library for chart elements and page templating. A Python 3 Flask server supports the Bokeh instance and provides user account and logging capabilities. Bokeh simplified the process of implementing interactive

client-server calls, hastening interface development. To make Silva easily extensible for future upstream and downstream data science applications, we created all high-level model objects in Python and represented data using Pandas DataFrames. Client side interface elements and callback events were implemented in plain JavaScript.

Silva's final design is the result of several iterations. To guide our development during the early stages of the project, we conducted pilot studies, inviting participants to use Silva to analyze a large dataset of their choice. We found that (1) participants tended to work with the causal graph directly and spent a lot of time switching between the dataset panel and the graph to form new groups; (2) individuals expressed strong preferences for charts to compare fairness values; (3) non-experts needed clarifications on definitions of fairness concepts. These findings inspired us to add more linkage between Dataset Panel (A) and Causal Graph (B), as well as between Table Group (C) and Fairness Dashboard (D). We reorganized Silva's workflow, making it possible to create and modify groups on the causal graph. We also added a animations and hover highlights to emphasize the linkages/connections between different components of Silva and facilitate comparison across groups, models, and fairness metrics. This augmented interactivity, along with higher data density, allowed us to bring the 4 distinct components of Silva together.

Causality Computation and Visualization

One key element on Silva is identifying causal relationships between data attributes. A number of methods (and corresponding toolkits) exist for causality computation. Probabilistic models are one of the most prominent approaches [38, 47], and structure learning algorithms are often used to extract attribute relationships. For Silva, we opted to employ off-the-shelf techniques. We adapted the dependency model illustrated in [38] and used the library Tetrad, integrated into Silva's back-end, to extract the underlying causal structure. One potential issue in causal models is redundancy (different underlying graphical structures which express the same causality). For the use cases explored in this paper we did not notice this issue, but it might emerge in a practical setting. In this case, there are a number of approaches for mitigating redundancy[31]. Scalability is also a concern here, which we revisit in the Discussion. Causal data is visualized for users in the form of a node-link diagram via the Bokeh framework. We use different colors and styles of nodes to imply the different types of attributes, and attempt to hide or merge isolated or low-signal nodes when there are many attributes on screen. In particular, long chains are compressed into summary nodes. Silva also recommends potential sensitive attributes to users by showing suggested attributes as dashed nodes on the causal graph once certain nodes are selected.

Model Training and Testing

Silva provides users three different types of models, Multilayer Perceptron (MLP), Random Forest (RF) and Logistic Regression (LGF). We opted to include these models as they are widely deployed in practice. Data are automatically split into training, validation and testing sets. We take a threshold for the classification which achieves the best accuracy on validation datasets. For the purposes of this investigation, we did not include parameter tuning. We argue that state-of-the-art auto-tuning approaches could be extended to improve model accuracy, making use of Silva's back-end portability.

Metric Calculation and Visualization

After model training, Silva calculates metrics based on the results of the model on the testing dataset. For our initial investigation, we followed the pattern of both AIF 360 and Google What-if[5, 26], including five metrics: Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD), Disparate Impact (DI), and Theil Index (TI). These metrics are all commonly used in practice and offer different insight into fairness. There are additional metrics which might be included, and there are numerous ways to improve metric calculation [54]. For the purpose of this investigation, we used standard approaches and added hooks in the back-end for additional metrics (or potentially even user-defined ones). Metrics are displayed through a dashboard, visualizing individual and summary results for comparison.

EVALUATION

In order to understand how Silva might help both data scientists and inexperienced users efficiently assess machine learning fairness, reason about sources of bias, and correctly identify bias, we conducted a controlled user study. Through this study we sought to identify promising application scenarios for Silva and potential shortcomings for future development. Silva's central features include: visualizations of causal interactions, interactive exploration of sensitive attributes, and comparisons of user-identified attribute groups. Our study assesses each of these components in isolation and together. In terms of the effectiveness of Silva, our study also evaluated whether individuals correctly located sources of unfairness in a model or dataset.

As no comparable causal investigation tools existed at the time of Silva's development, we aimed in our evaluation to contrast Silva against state-of-the-art tools that a practitioner might plausibly use for similar use cases. IBM AI Fairness 360 (AIF) [4] is a widely distributed open-source toolkit for bias debugging and has recently been extended as an automatic system for social bias detection and mitigation. Its performance in unfairness assessments is well established. We chose AIF as a comparison case for Silva. While affordances are not a 1-to-1 match and AIF is a more mature software product (which might offer it an unfair advantage), its mix of manual and automated tools acts as a beneficial counterpoint for the interactive sandbox approach of Silva. In particular, by choosing AIF we hoped to expose trade-offs between the immediacy of automated systems (AIF) and the understanding gained over exploration (Silva).

Methodology

During our user study, participants used Silva and AIF to complete two different tasks in a 50 minute session. Afterwards, users assessed both systems through surveys. Our study employs two datasets: Adult Census Income (Adult)[1] [67] and

[1] https://archive.ics.uci.edu/ml/datasets/adult

Berkeley Dataset 1973 (Berkeley) [7]. These datasets are open-source and have been widely employed by bias evaluation and explanation researchers. As these datasets have been well studied, there exists reliable ground truth data on attribute sensitivity and bias.

Participants' first task was to explore whether there is bias in salary predictions for an individual if predictions are above $\$50,000$ per year in the Adult dataset. Their second task was to investigate whether there is bias in Berkeley's graduate school admissions (a well-known case study for Simpson's paradox). These two datasets have been widely applied in machine learning fairness research. It is well established that salaries in the Adult dataset reflect biases with respect to race and gender, but admission outcomes in the Berkeley dataset do not encode biases with respect to gender [7, 67]. As Berkeley is a relatively small dataset, we anticipate that both skilled and unskilled participants will perform well in the second task, however Silva participants (should the tool prove effective) ought to perform better. On the other hand, the Adult dataset has higher complexity, which might expose gaps between novice and expert participants, as well as potentially emphasize the benefits of Silva.

The two user study tasks are representative of common patterns encountered by data scientists [18]. Given a prediction task, a data scientist might first identify relevant data attributes based on their existing experience and knowledge. Then, they may use tools to explore the given dataset in more detail. To mimic this process, we first asked participants to identify attributes relevant to the prediction task, and to identify any potential bias in the dataset. Participants make use of their own knowledge without the aid of any tool. Then, they explore the dataset with the assigned tool (Silva or AIF). After using one tool, users are asked to re-identify relevant attributes and sources of bias. They use the other tool during the second task. We counterbalance both dataset and tool order so that there is even exposure to experimental conditions. During the study, participants were asked to evaluate tool components after they finished using them. At the end of the study, we asked users to complete a post-survey, reflecting on their answers on the pre-survey, and providing qualitative feedback.

As Silva and AIF may be relatively complicated, we provided participants with short tutorial videos explaining the tools used in the study. The tutorials showcased a separate dataset not used in the user study. We used the same examples to develop the tutorial video, and both videos had comparable length. As AIF also has debiasing components that are not present in Silva, our protocol stopped participants at the end of the bias detection phase of the tool. We also provided participants with cheat sheets and plain English definitions of statistical fairness metrics in case they forgot instructional content. After training, participants were given a few minutes to use the tool and ask the experimenter questions.

Participants were recruited through a university research pool as well as through social media. Participants were screened based on prior exposure and experience with data analysis, the study tasks, machine learning background, and algorithmic fairness. The pre-screen was employed to select two groups of participants in roughly equal proportions: 1) Novices who do not have experience with fairness analysis nor solid knowledge of the two datasets of the studies, and 2) Skilled participants who have at least some working knowledge of machine learning and machine learning fairness. We make use of pre-screen responses as a comparison point in our post-survey analysis.

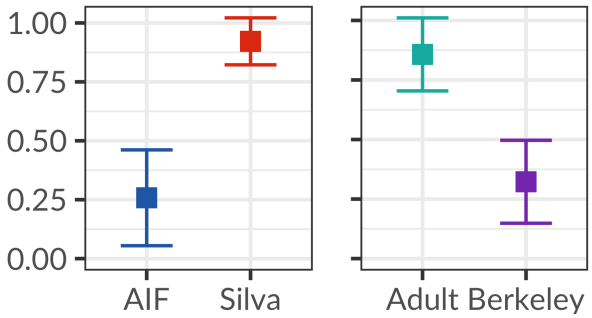


Figure 3. Mean and standard error for self-reported usefulness feedback (0 being neutral, 1 being useful) split by tool and dataset.

33 individuals participated in our study. Of those participants, 30 completed the entire protocol and submitted usable survey responses. 3 participants either did not use both tools or did not submit post-surveys and left the session early. 10 participants identified as male and 20 as female. 14 were university graduate students, and the other 16 were university undergraduate students. 15 participants ultimately fit into our Skilled category and another 15 fit our Novice category. Participants were grouped evenly (7 or 8 per Latin square cell) into tool and dataset conditions, counterbalanced for order effects.

Results

Self-reported Usability

Participants reported their experiences with each component of Silva and AIF on a 5-point Likert scale ranging from *not useful at all* (-2) to *very useful* (2). Participants had the option to state that they did not use a component and did not feel comfortable rating it. These were counted as missing data in our analysis.

In general, users of Silva rated the causal graph highly (M:1.1, SD:0.85), indicating that they found this central feature to be very useful in helping them finish their tasks. Participants also reported that saving groups (M:0.75 , SD:0.79), metric visualization (M: 0.93, SD: 0.98) and toggling sensitive attributes (M:0.79, SD:0.89) were useful as well. For AIF, participants' results show moderate usefulness ratings for the automatic and efficient analytic result (M: 0.27, SD:1.2) and the metric visualization (M:0.29, SD: 1.13).

We averaged survey responses for each tool into a single factor for comparison (factor analysis confirmed item-level agreement). Overall, participants reported significantly higher responses for Silva ((M: 0.90, SD: 0.56) compared to AIF (M: 0.29, SD: 1.13). This suggests that Silva indeed provided value to participants in completing the tasks, and that it may outperform AIF in terms of overall usability.

In order to understand how task, tool, and participant experience relate to each other with respect to these self-reported utility measures, we constructed a mixed-effect linear model

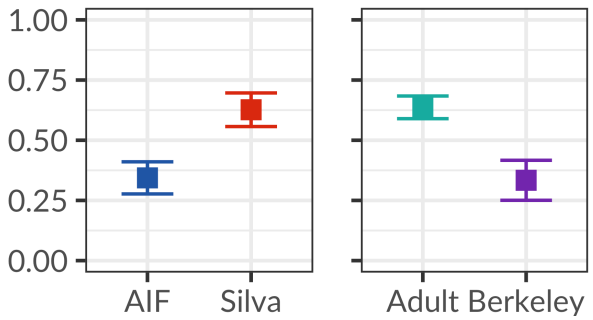


Figure 4. Mean and standard error for discovery F-score (higher is more accurate) split by tool and dataset.

testing interactions between all three independent measures and predicting for averaged self-reported utility. We employ a mixed-effect model to account for repeated measures in our within-subjects study design. The model detected two significant main effects: task ($F(1,24) = 7.78$, $p = .01$) and tool ($F(1,24) = 15.71$, $p = .0005$) as depicted in Figure 3. In general, individuals reported more positive responses to Silva and after completing the Adult task. It is possible that the high complexity of the Adult dataset allowed individuals to more fully explore and make use of tool capabilities, exposing more potential benefits of the tool. We did not detect a main effect for novice/skilled and did not find any significant interaction effects. This is also encouraging, as it suggests that experience did not ultimately play an observable role in tool satisfaction.

Examining Participant Discoveries

In addition to evaluating the usefulness of Silva from the perspective of self-reported utility, we also considered effectiveness in terms of true positive discoveries vs. false positive discoveries made by participants during their investigation.

In our post-survey, participants were required to identify and explain whether there was social unfairness in the Adult prediction task and whether there exists gender bias in admissions in the Berkeley task. As these tasks have ground truth answers, we can compare whether the answer participants provide (and the evidence they cite to justify their response) is valid or not. We employ an F-score to measure the truth discovery rate. F-score is the harmonic mean of precision (how many identified biases are indeed biases?) and recall (how many biases are identified?) in the evidence given by the participants. A high F-score indicates that users are able to correctly identify many biased attributes without mistakenly selecting many unbiased attributes. A low F-score indicates that users make mistakes when identifying biased attributes, either by missing many biased attributes or by incorrectly selecting many unbiased attributes.

In general, Silva achieved a higher F-score (M:0.63, SD:0.38) in helping identify unfairness in existing datasets compared to AIF (M: 0.35, SD: 0.37). A two-tailed t-test indicates that these differences are significant ($t(58) = 2.93$, $p < .0049$). To further validate these claims, we constructed another mixed-effect linear model examining potential interactions between task, skill, and tool in predicting the overall F-score of participant assessments. Our results mirror our earlier model predictions for self-reported utility. While there were no interaction effects and skill did not play an observable role in discovery f-score, we again observed a main effect between task ($F(1,24) = 11.13$, $p = .0028$) and tool ($F(1,24) = 11.19$, $p = .0027$) as depicted in Figure 4. Generally, skilled practitioners achieve a marginally higher F-score in discovering the unfairness compared to novices for both datasets. Notably, participants tended to achieve a higher accuracy with a more complex dataset (Adult). This implies that complexity provides users with more room to explore (and potentially make mistakes). Experience may not be a necessity in our scenarios. This is especially encouraging, as it provides further evidence that Silva may help both novice and expert users to obtain useful results.

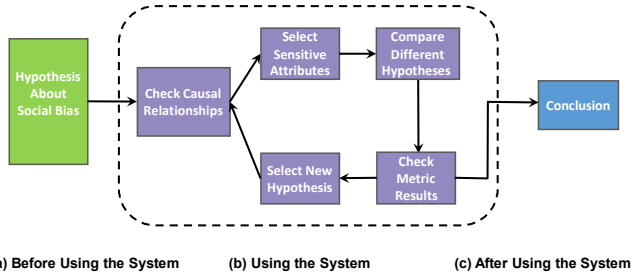


Figure 5. Reasoning path of practitioners.

Qualitative Feedback

We collected qualitative responses concerning how participants' working process and their user-experience. Here we briefly outline some themes we noted:

Reasoning on the sources of unfairness: We invited users to briefly describe why they believe certain attributes lead to potential social bias. Although some participants did not justify their reasoning, fifteen responses explained how they arrived at their conclusion. We identified a general pattern participants followed to come to their conclusion in Figure 5. We note that practitioners' high-level descriptions suggest a loop of creating and validating hypotheses. The central causal graph in Silva played a role in helping participants compare among different groups and enable them to develop alternative explanations (as evidenced by the discovery metric results). The way that participants leveraged Silva is consistent with sensemaking theory [48].

Causal graph proved helpful: In their responses, participants expressed their appreciation of the causal graph. One claimed, "I can see the relationship between different attributes in Silva"; and another expressed, "the causal graph shows the influence of sex in Berkeley". Participant mentioned the ability of the causal graph to expose dependencies ("...causal graph was the most helpful as it differentiated dependencies") and attribute-level relationships ("Silva allows a lot of explanations to help us determine why it is ok or not by looking at direct and indirect relationships").

Interactivity is valued: Participants pointed out that Silva interactions provided valuable information for making sense of unfairness, especially in comparison to AIF. One commented on AIF that, "I want to know why it is biased, not machine tell me why," compared to a Silva participants' claim, "Silva has more components to help me explore the data." The lack of interactivity in AIF was a broader concern among participants.

One expressed "I don't know any details and reasons of their results." Another claimed "AI 360 could have more analytic options," and "It would be better if AI 360 incorporates the features of Silva". While AIF's automated efficiency might speed investigations, the lack of interactivity could ultimately have a negative impact on overall user experience. Combined with the quantitative results, there is evidence that increased interactivity lead to an improved fairness analysis process.

DISCUSSION AND LIMITATIONS

In this section, we discuss the results of evaluation, and identify some potential limitations for Silva, and highlight areas for future investigation.

Conclusion of Evaluation

The results from the evaluation suggest that Silva's interactivity helps practitioners effectively identify bias in machine learning algorithms and datasets. Encouragingly, we also noticed that participant skill level did not play a role in our outcome measures. In addition to being robust to user skill level, Silva offered efficient exploration over both datasets in the study, suggesting that it can be generally applied, even in competition to a mature software platform. User feedback indicates that Silva enhanced their sensemaking process, but further studies are necessary to explore this fully.

Potential Limitations

Scalability: As with any interactive data-driven system, scalability is a major potential limitation. There are three problem areas where scalability issues might emerge:
(1) Training, modeling, and metric calculations: The training time of models largely depends on the model complexity, scale of the datasets and hardware constraints. This is an ongoing area of study in the machine learning research community. While we endeavored to use recent approaches available, new research advances may assist in making training/causality computations more efficient at scale.
(2) Communication and latency: Due to pre-computation, the interactive visualizations themselves remained performant with large-scale or complex data. That may not remain true as complexity increases. Data might reach scales that cannot reliably be transmitted over a web connection or stored in browser memory. This might necessitate additional load balancing between the front- and back-end. Likewise, computations for interactions (e.g. hiding nodes in SVG) might reach a point where latency occurs. Both cases are known issues for web tools, and there are numerous approaches for mitigating them.
(3) Human factors: In addition to computational complexity, graphs may be overly complicated if there are many attributes or relationships (i.e. graph spaghetti), leading users to make mistakes or experience overload. While we introduce some fixes in Silva like bundling similar nodes, this is a concern. Affordances such as attribute selection and navigation may be challenging to use, especially at high attribute counts. There are some potential fixes: One might employ scalable widgets (e.g. fisheye menus) and limit detail through clustering/hierarchies. We leave this for future study. While we did not notice divisions based on user skill, training was still a significant component in our experimental protocol. In a production context, training and providing adequate information on metrics could also pose scalability concerns.

Representing causality: Though probabilistic graphical models are one of the most useful unifying approaches for connecting both graph theory and causality, it can be difficult for novices to understand the causal graph. Even for skilled users, two participants expressed that they wanted to learn the strength of the dependency through the edges that connects attributes in the user study. At the moment, Silva provides a summary table with short explanations of deterministic relationships. There is an opportunity for providing greater detail to help explain the quantitative strength of the deterministic relationship in addition to the summary view, at the risk of additional complexity.

As the graph is automatically learned by default structure learning algorithms, sometimes the causal graph structure might be counter-intuitive [38]. Consistency of the causality structure has been long studied in the artificial intelligence community. Multiple independence testing could be a window to handle this issue. Many efficient independence testing methods have been proposed to interactively construct causal views with the help of users and may prove to be helpful. As mentioned earlier, redundancy might also pose an issue, especially as complexity increases.

Offering paths for mitigation: Two participants commented that discoveries made with Silva are an important step for resolving unfairness; however, it would be beneficial for Silva to provide some heuristics for selection of down-stream mitigation. Silva outputs the privileged group and under-privileged group with respect to biased attributes, which defines the groups that are unfairly represented by the algorithm. Many optimal policy algorithms use these group definitions as input to achieve fairness through the counterfactual settings. Silva can couple well with those down-stream approaches, though at present it has not been integrated into a pipeline with them.

Potential Benefits and Future Work

Efficiency gains over time: Although it took a few minutes for users to familiarize themselves with Silva's components and to ramp up their understanding during our lab study, users were quick to analyze the dataset afterwards. We noted that participants working with more complex data tended to ramp up more quickly. We posit that the high level of interactivity might play a role here. By inviting users to explore, Silva might soften the initial barrier to entry and encourage users to experiment with new features. This could be beneficial for novice data scientist adoption.

Enhanced reasoning: As mentioned in the evaluation, we noticed that users tended to follow a sensemaking loop of creating and validating hypotheses in Silva. This feedback loop has proved difficult to achieve with existing optimization tools (which resonates with participants' negative reactions to "black-box" recommendations by AIF). In addition, users noted that the causal graph enhanced their understanding of the test dataset and model. We infer that Silva might help users enhance their sensemaking process of machine learning

unfairness assessment through their interactions. There is an opportunity for additional investigation of the mental model of practitioners as they evaluate data fairness, building on current design research on machine learning user experiences [66].

Deeper causality: One central design goal for Silva was to help users explore the connection between causality and metric results as a means to accelerate fairness evaluation. We found that when there are direct causal relationships from sensitive attributes (, Silva performs well. For example, if two sensitive attributes both affect fairness, existing metrics might provide two different values for their impact. The causal graph can help the user trace why the influence from each attribute is different. We hypothesize that the direct causal relationship permits users to explore and isolate sources of unfairness in their data. However, it might be hard for the user to identify specific influences if data are very complex or intra-correlated. Visualizing causality in these scenarios remains an active area of study. Further, there is a potential for providing additional signals to users about causality. For example, in complex datasets, large causal chains might be hard for users to parse. Improved visual metaphors and interactive tools might assist users in untangling these relationships.

Pipeline integration: In Figure 1, we identified Silva's major target area. In practice, this represents a slice of a much larger data science pipeline. Thinking holistically, there are a number of opportunities for greater integration of Silva into data science workflows, which might provide benefits for users both in terms of understanding and efficiency. Silva users appreciated the explanatory ability of the tool, but expressed a desire for pathways to mitigate bias. Including some of the automated features for mitigation (such as those in AIF) could help to close this loop. Further, additional interactivity for data exploration might remove some of the "black-box" concerns users expressed about recommendations. Leading in to Silva, there is also an opportunity to connect the tool to existing exploratory data analytics platforms, supporting users from hypothesis generation to final fairness mitigation/decision-making. Along these lines, we hope to conduct a larger, long-term deployment of Silva by integrating it into a data science pipeline in an institutional context.

CONCLUSION

This paper introduced Silva, an interactive system that helps data scientists to reason effectively about unfairness in machine learning applications. Silva couples well with existing machine learning pipelines. It integrates a causality viewer to assist users in identifying the influence of potential bias, multi-group comparisons to help users compare subsets of data, and a visualization of metrics to quantify potential bias. In a user study we demonstrated that Silva was favored by both skilled and novice participants. Silva achieved a higher F-score accuracy in assisting participants in locating socially unfair biases in benchmark datasets. The user study also indicates that the usability and effectiveness of Silva is not dependent on practitioners' skills, which means that Silva might be more widely applicable. As a whole, we have provided some initial signs that integrating causal reasoning in interactive fairness assessment tools can provide benefits for analysts.

ACKNOWLEDGEMENTS

This work was supported by NSF grant IIS-1850195. We would like to thank the associate chairs and anonymous reviewers for their invaluable feedback.

REFERENCES

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

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

[3] David Beer. 2009. Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11, 6 (2009), 985–1002.

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

[5] 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. 2018b. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. (Oct. 2018). `https://arxiv.org/abs/1810.01943`

[6] 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. *arXiv preprint arXiv:1901.04562* (2019).

[7] 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.

[8] 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.

[9] Bokeh Development Team. 2019. *Bokeh: Python library for interactive visualization*. `https://bokeh.org/`

[10] 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.

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

[12] Jenna Burrell. 2016. How the machine âĂŸthinksâĂŹ: Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[14] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 319–328.

[15] 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.

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

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

[18] Thomas H Davenport and DJ Patil. 2012. Data scientist. *Harvard business review* 90, 5 (2012), 70–76.

[19] 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.

[20] 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.

[21] Benjamin G Edelman and Michael Luca. 2014. Digital discrimination: The case of Airbnb. com. *Harvard Business School NOM Unit Working Paper* 14-054 (2014).

[22] 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 (2017), 115.

[23] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[24] Tarleton Gillespie, Pablo J Boczkowski, and Kirsten A Foot. 2014. *Media technologies: Essays on communication, materiality, and society*. MIT Press.

[25] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a âĂIJright to explanationâĂİ. *AI Magazine* 38, 3 (2017), 50–57.

[26] Google. 2017. What-if Tool. (2017). `https://pair-code.github.io/what-if-tool/`

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

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

[29] 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.

[30] Tommi Jaakkola, David Sontag, Amir Globerson, and Marina Meila. 2010. Learning Bayesian network structure using LP relaxations. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 358–365.

[31] Michael I Jordan. 2003. An introduction to probabilistic graphical models. (2003).

[32] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[33] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

[34] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[35] 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.

[36] 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.

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

[38] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

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

[40] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

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

[42] 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.

[43] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[44] Caitlin Lustig and Bonnie Nardi. 2015. Algorithmic authority: The case of Bitcoin. In *2015 48th Hawaii International Conference on System Sciences*. IEEE, 743–752.

[45] James Massey. 1990. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*. Citeseer, 303–305.

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

[47] Judea Pearl and others. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.

[48] 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.

[49] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *26th USENIX Security Symposium (USENIX Security 17)*. 935–951.

[50] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[51] Babak Salimi, Corey Cole, Peter Li, Johannes Gehrke, and Dan Suciu. 2018a. HypDB: a demonstration of detecting, explaining and resolving bias in OLAP queries. *Proceedings of the VLDB Endowment* 11, 12 (2018), 2062–2065.

[52] Babak Salimi, Johannes Gehrke, and Dan Suciu. 2018b. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, 1021–1035.

[53] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *Proceedings of the 2019 International Conference on Management of Data*. ACM, 793–810.

[54] 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.

[55] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2239–2248.

[56] Jina Suh, Soroush Ghorashi, Gonzalo Ramos, Nan-Chen Chen, Steven Drucker, Johan Verwey, and Patrice Simard. 2019. AnchorViz: Facilitating Semantic Data Exploration and Concept Discovery for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 1 (2019), 7.

[57] Astra Taylor and Jathan Sadowski. 2015. How companies turn your Facebook activity into a credit score. *The Nation* 27 (2015).

[58] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning* 65, 1 (2006), 31–78.

[59] Lisa Tweedie, Bob Spence, David Williams, and Ravinder Bhogal. 1994. The attribute explorer. In *Conference companion on Human factors in computing systems*. ACM, 435–436.

[60] 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.

[61] Clifford H Wagner. 1982. Simpson's paradox in real life. *The American Statistician* 36, 1 (1982), 46–48.

[62] 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)*. 271–285.

[63] 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.

[64] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081* (2017).

[65] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *Advances in Neural Information Processing Systems*. 3399–3409.

[66] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems (DIS) Conference*. ACM, 585–596.

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

[68] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 335–340.

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