Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment

Jonathan Dodge
dodgej@eecs.oregonstate.edu
Oregon State University

Q. Vera Liao
Yunfeng Zhang
vera.liao@ibm.com
zhangyun@us.ibm.com
IBM Research AI

Rachel K. E. Bellamy
Casey Dugan
rachel@us.ibm.com
cadugan@us.ibm.com
IBM Research AI

ABSTRACT

Ensuring fairness of machine learning systems is a human-in-the-loop process. It relies on developers, users, and the general public to identify fairness problems and make improvements. To facilitate the process we need effective, unbiased, and user-friendly explanations that people can confidently rely on. Towards that end, we conducted an empirical study with four types of programmatically generated explanations to understand how they impact people's fairness judgments of ML systems. With an experiment involving more than 160 Mechanical Turk workers, we show that: 1) Certain explanations are considered inherently less fair, while others can enhance people's confidence in the fairness of the algorithm; 2) Different fairness problems–such as model-wide fairness issues versus case-specific fairness discrepancies–may be more effectively exposed through different styles of explanation; 3) Individual differences, including prior positions and judgment criteria of algorithmic fairness, impact how people react to different styles of explanation. We conclude with a discussion on providing personalized and adaptive explanations to support fairness judgments of ML systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Fairness, Machine Learning, Explanation, Empirical Studies

1 INTRODUCTION

Increasingly, important decisions that impact human lives and societal progress are supported by machine learning (ML) systems. Examples where ML systems are used to make decisions include hiring, marketing, medical diagnosis, and criminal justice. This

trend gives rise to concerns about algorithm fairness—or possible discriminatory consequences for certain groups of individuals. Machine learning algorithms are trained based on data from past decisions, decisions which may have themselves been biased and discriminatory. Research shows that by optimizing for the unitary goal of accuracy, ML algorithms trained on historical data not only replicate, but may amplify existing biases or discrimination [40]. The possibility of spiraling discriminatory consequences is driving a distrust and "fear of AI" in public discussions (e.g., [1, 2]).

There is a growing body of work on developing non-discriminatory ML algorithms (e.g., [16, 17, 38]), equal attention has not been paid to the human scrutiny necessary to identify and remedy fairness issues. The need for such research is highlighted by recent studies, which show that algorithmic fairness often may not be prescriptively defined, but is multi-dimensional and context-dependent [14]. Public scrutiny of the usage of risk assessment algorithms in the criminal justice system [3, 21] brings attention to the need to progress the accountability and fairness of such algorithms.

Accurately identifying fairness issues in ML systems is extremely challenging, however. Most ML algorithms aim to produce only prediction or decision outcomes, while humans tend to rely on information about decision-making *processes* to justify the decisions made. ML algorithms are often seen as "black boxes", where one can only see the output and make a best guess about the underlying mechanisms. This problem is further exacerbated by the popularity of deep learning algorithms, which are often unintelligible even for experts. This lack of transparency drives a sweeping call for *explainable artificial intelligence* (XAI) in industry, academia, and public regulation. For example, the EU General Data Protection Regulation (GDPR) requires organizations deploying ML systems to provide affected individuals with meaningful information about the logic behind their outputs.

Critically, explanations are not just for people to understand the ML system, they also provide a more effective interface for the human in-the-loop, enabling people to identify and address fairness and other issues. When people trust the explanation, it follows that they would be more likely to trust the underlying ML systems.

Much recent research is dedicated to the generation of explanations in various styles, including model-agnostic approaches [26, 30] applicable to any ML algorithm. However, this growing body of research is criticized for "approaching this [XAI] challenge in a vacuum considering only the computational problems" [27] without the quintessential understanding of how *people* perceive and use the explanations.

In this paper, we conduct an empirical study on how people make fairness judgments of ML systems and how explanation impacts that judgment. We aim to highlight the nuances of such judgments, where there are different types of fairness issues, different styles of explanation, and individual differences, to encourage future research in this area to take more user-centric and personalized approaches.

Specifically, we identify four styles of explanation based on prior XAI work and automatically generate them for a ML model trained on a real-world data set. In the experiment, we explore the effectiveness of explanations in exposing two types of fairness issues–model-wide unfairness produced by biased data, and fairness discrepancies in cases from different regions of the feature space. Our user study demonstrates that judging fairness is not only influenced by explanation design, but also an individual's prior position on algorithmic fairness, including both the general trust of ML systems for decision support and one's position on using a particular feature. We also present user feedback for the four styles of explanation. Our results provide insights on the mechanisms of people's fairness judgment of ML systems, and design guidelines for explanations to facilitate fairness judgment making. We first review relevant work, then present the study overview and research questions.

2 BACKGROUND

2.1 Fairness of Machine Learning Systems

One of several definitions for algorithmic fairness is: "*...discrimination is considered to be present if for two individuals that have the same characteristic relevant to the decision making and differ only in the sensitive attribute (e.g., gender/race) a model results in different decisions*" [8]. The consequence of deploying unfair ML systems could be *disparate impact*, practices which adversely affect people of one protected characteristic more than another in a comparable situation [8, 14].

Despite the "statistical rationality" of ML techniques, it has been widely recognized that they can lead to discrimination. Many reasons can contribute to this, including biased sampling, incorrect labeling (especially with subjective labeling), biased representation (e.g., incomplete or correlated features), suboptimal or insensitive optimization algorithm, shift of population or data distribution, and failure to consider domain-specific, legal, or ethical constraints [8, 15]. Various techniques have been proposed to address these causes of "unfair algorithms" [15–17, 38, 39]. For example Calders and Žliobaitė suggested techniques to de-bias data [8], including modifying labels of the training data, duplicating or deleting instances, adding synthetic instances, and transforming data into a new representation space.

We use a recently proposed data de-biasing method that applies a preprocessor to transform the data [9]. The result is a new dataset which is "fairer"–while also limiting local deformations from the data transformation. This is because the preprocessor optimizes data transformations with respect to penalties that rise with the magnitude of a feature change (e.g. changing a persons age from 5 to 60 will result in a higher penalty than from 5 to 8). Simply put, if raw data contains biases that lead to an unfair model with a discriminatory feature (e.g., certain racial category is weighed more negatively than others), the data preprocessing mitigates the bias introduced by that feature. This method has the benefit of retaining all features (as opposed to removing the discriminatory feature), which, among other benefits, would also allow exploration of correlations among them [8].

The above debiasing techniques are *normative* by nature, i.e., they rely on prescriptively defining the criteria of fairness in order to optimize for that criteria. A recent paper pursued a complementary *descriptive* approach by empirically studying how people judge the fairness of features used by a decision support system in the criminal justice system [14]. Their study uncovered the underlying dimensions in people's reasoning of algorithmic fairness, and demonstrated individuals' variations on these dimensions.

We adopt the same descriptive view, empirically studying how people judge fairness of an ML system and considering individual differences in their prior position on algorithmic fairness. However, we also fill a gap in prior work by investigating how normative fairness (via the use of the preprocessor) is *perceived* by people, and what factors impact such perception.

2.2 Explanation of Machine Learning

Explainable AI (XAI) is a field broadly concerned with making AI systems more transparent so people can confidently trust an AI system and accurately troubleshoot it — fairness issues included. Work on model explanations can be traced to early work on expert systems [11, 33], which often explicitly revealed reasoning rules to end-users. There has been a recent resurgence of XAI work driven by the challenge to interpret increasingly complex ML models, such as multi-layered neural networks, and by the evidence that ethical concerns and lack of trust hampers adoption of AI applications [15, 22].

A large volume of XAI work is on producing more interpretable models while maintaining high-level performance (e.g. [10, 23]), or on methods to automatically generate explanations. Given the complexity of current ML models, explanations are often *pedagogical* [34], meaning that they reveal information about how the model works without faithfully representing the algorithms. Many methods rely on some kind of sensitivity analysis to illustrate how a feature contributes to the model prediction [26, 30], so they can be model-agnostic, thus applicable to complex models. For example, LIME explains feature contribution by what happens to the prediction when the feature-values change (perturbing data) [30]. Another common category is *case-based explanations*, which use instances in the dataset to explain the behavior of ML models. Examples include using counter-examples [36] and similar prototypes from the training data [18]. Case-based explanations are considered easy to consume and effective in *justifying* the decision, but may be insufficient to explain how the model works.

Work on how people perceive explanations of ML systems is a growing area [4, 20, 25, 32] which aims to inform the choices and design of explanations for particular systems or tasks. Recent work calls for taxonomic organizations of explanations to enable design guidelines [25]. In earlier work on explaining expert systems, researchers argued the difference between *description v.s. justification*–by making not only the *how* visible to users, but also the *why* [33]. Accordingly, Wick and Thompson discussed the taxonomy of *global-local* explanations [37]. During initial practice, users

may need global explanations that describe "how the system works." During actual use, users tend to rely on justifications of why the system did what it did on particular cases.

Another useful taxonomy is proposed by Kulesza et al. by considering two dimensions of explanation fidelity: *soundness* (how truthful each element in an explanation is with respect to the underlying system) and *completeness* (the extent to which an explanation describes all of the underlying system) [20]. They empirically showed that the best mental models arose from explanations with both high completeness and high soundness. However, crafting highly complete explanations comes with a tradeoff, as completeness usually requires increasing the length and complexity of the explanation, which was shown to be detrimental to task performance and user experience in previous studies [28].

While researchers have explored user preferences in explanation styles, they have paid little attention to individual differences in such preferences. Meanwhile, psychological research has long been interested in individual differences in explanatory reasoning. For example, research shows that some prefer simple, superficial explanations and others are more deliberative and reflective in their reasoning [13, 19]. Such individual differences can be predicted by cognitive style (e.g., cognitive reflection, need for cognition) [13] and culture [19]. It is therefore possible that individuals differ in preferences for completeness and soundness of explanations.

Our work is concerned with how explanations impact fairness judgments of ML systems. We build on a recent study by Binns et al., which examined human perception of a classifier's fairness in the insurance domain [5]. They provided four different explanation types applied to *fictional scenarios* to elicit fairness judgment. While the study provided rich qualitative insights on the heuristics people use to make fairness judgments, the authors acknowledge a lack of ecological validity as the explanations were not drawn from real ML model output. Moreover, the explanations were not produced for the same data points, so they were incommensurate, which could possibly explain the absence of conclusive preference in their quantitative results.

Our work set out to overcome limitations on prior work by *automatically generating* four types of explanations on a real ML model, and quantitatively examining how they impact people's fairness judgments. Combining this advancement with the use of the data preprocessor allowed us to perform more carefully controlled experiments for ML fairness perception than prior work.

3   STUDY OVERVIEW

Related work informed four main considerations of our study: the use case, choices of explanation styles, fairness issues we focus on, and the individual differences we explore. Through both quantitative and qualitative results, we aim to answer the following research questions:

RQ1 How do different styles of explanation impact fairness judgment of a ML system?

RQ1a Are some explanations judged to be fairer?

RQ1b Are some explanations more effective in surfacing unfairness in the model?

RQ1c Are some explanations more effective in surfacing fairness discrepancies in different cases?

RQ2 How do individual factors in cognitive style and prior position on algorithmic fairness impact the fairness judgment with regard to different explanations?

RQ3 What are the benefits and drawbacks of different explanations in supporting fairness judgment of ML systems?

3.1   Use Case: COMPAS recidivism data

We conducted an empirical study with a ML model trained on a real data set. Similar to [14], we chose a publicly available data set for predicting risk of recidivism (reoffending) with known racial bias[1]. The data set was collected in Broward County, Florida over a two year span. It is used by COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a commercial algorithm to help judges score criminal defendants' likelihood of reoffending. However, ProPublica has reported on troubling issues with the COMPAS system [3, 21]. First, the classifier may have low overall accuracy ([3] reported 63.6%). Second, the model is reported to exhibit racial discrimination, with African American defendants' risk frequently overestimated.

We chose a criminal justice use case because it carries weight to elicit reaction on fairness, even for the general population. Note our goal is not to study the actual users of COMPAS. Rather, we are using the use case as a "probe" to empirically study fairness judgments. The same use case was used in previous studies to understand how people perceive algorithmic fairness with regard to features used [14].

3.2   Explanation Styles

We chose to programmatically generate the four types of explanations introduced by Binns et al. [5] (details to be discussed in System Overview section) because they represent a set of common approaches in recent XAI work. They embody the categorization of *global v.s. local* explanations. Specifically, influence and demographic-based explanations are global styles as they describe how the model works; sensitivity and case-based explanations are local styles as they attempt to justify the decision for a specific case. These explanation styles vary along the taxonomy introduced by previous work in other ways: e.g. sensitivity based explanation is similar to a "what if" [25] and the case-based explanation is the least "sound" of the explanation types discussed [20].

3.3   Fairness Issues

Given our use case, we consider fairness issues in terms of racial discrimination. While there are other controversial features in the dataset [14], race is generally considered inappropriate to use in predicting criminal risk (termed *protected variable*). We focus on two types of fairness issues.

3.3.1   Model Unfairness. As discussed, the COMPAS data set is known to be racially biased, but we mitigate that bias problem by using the data processing method in [9]. In the experiment, we introduce the use of the **data processing** technique as a between-subject variable. By comparing participants' fairness judgment for a model trained on the *raw* data to that of *processed* data, we aim to understand whether participants could identify the model-wide

---

[1]https://www.kaggle.com/danofer/compass

fairness issue, and whether certain explanations expose the problem better.

3.3.2 Case-specific disparate impact. Predictions from an ML algorithm are not uniformly fair–consider *disparate impact* from a protected variable. For example, if two individuals with identical profile features but different racial categories receive different predictions, it should be considered unfair [8, 14]. Statistically, these cases are on the decision boundary of the feature space given the relatively small weight of the race factor, meaning they have low-confidence predictions that may be unfair. In the experiment, we introduce **disparate impact** by race factor as a within-subject variable (i.e., each participant would be asked to judge some cases with disparate impact and some not). We adopt a factorial design with *disparate impact* and *data processing*. For subjects given models after data processing [9], disparate impact is reduced. We aim to discover how well participants identify the case-specific fairness issues using different explanations.

Our hypothesis is that given local explanations focus on justifying a particular case, they should more effectively surface fairness discrepancies between cases. In contrast, global explanations may require additional effort to reason about the position of the case with respect to the decision boundary (e.g., "*This person's features all have no impact in the model, except race*"). Note that local explanations may expose the case-specific fairness issue differently. Case-based explanation exposes the boundary position with a low percentage of cases justifying the decision. Sensitivity-based explanation explicitly describes disparate impact–"*Changing this person's race changes the prediction*".

3.4 Individual Difference factors

Based on prior work, we focus on two areas of individual factors: cognitive style and prior position on algorithmic fairness. For cognitive style, we measure individual's *need for cognition* [7]. For prior positions, we consider two levels: one's *general position on the fairness of using ML systems* for decision support, and one's *position on the fairness of using a particular feature*–here we focus on the race factor.

4 SYSTEM OVERVIEW

4.1 Re-offending Prediction Classifier

The model is a binary classifier predicting whether an individual in the COMPAS data set is likely to re-offend or not, implemented by Scikit-learn's logistic regression. The use of a regression model is ecologically valid—many current decision support systems use such simple and interpretable models [35]. However, the explanation styles we study are not limited to regression models.

We built the model using a subset of features in the COMPAS dataset[2], including *Race* as the feature with fairness issues. For simplicity, we focus on two racial groups (Caucasian and African-American), and filtered others. Other features included: *Age* (18-29/30-39/40-49/50-59/>59), *Charge Degree* (Felony/Misdemeanor), *Number of Prior Convictions* (0/1-3/4-6/7-10/>10), and *Had Juvenile Convictions* (True/False). According to Grgic-Hlaca et al. [14], charge degrees and criminal history were deemed fair in a similar

use case. Age is also important in assessing re-offense risk. Following statistical convention, all categorical features are dummy coded using the median category as the reference level, where possible.

The accuracy of the model is 67.1% on raw data and 67.6% on processed data, comparable to reports on the accuracy of COMPAS system [6, 21]. Note that logistic regression also produces a confidence level implicitly in class probabilities.

4.1.1 Data processing and cases with disparate impact. We used the method introduced in [9] to perform data processing then re-trained the model. The resulting model reduced bias against the African American group, as evidenced by the feature co-efficient being reduced from 0.177 to -0.036 (A feature co-efficient of 0 corresponds to the feature having no effect in the decision).

To identify cases with unfair treatment of disparate impact, we follow the definition "*treating one person less favorably on a forbidden ground than another...in a comparable situation*" [8]. That is, if perturbing a test example's protected variable (race) changes the algorithm's prediction, we consider it to have disparate impact. We found 23 cases in the raw dataset–all very near to the decision boundary[3].

4.1.2 Sampling cases for the user study. Due to user study time constraints, we could only show each user a small sample of the explanations. Since we intended to study fairness discrepancies between disparately impacted and non-impacted cases, we over-sampled the former category. Among the 23 disparately impacted instances, we sampled all 8 unique cases (i.e. the rest had the same feature-values as one of the 8 we sampled). From the non-impacted group of 992 instances, we sampled 16 unique cases.

4.2 Explanation Generation

As discussed, we patterned our explanations, shown in Figure 1 (truncated version, see supplementary materials for the full version), after the templates presented by Binns et al. [5]. While Binns et al. manually created examples of these explanation, we developed programs to automatically generate them to obtain comparable explanation versions for the same data point, controlling for differences in representation and presentation. These generation methods can also be broadly applied to ML prediction models using relational features.

4.2.1 Input Influence-based Explanation. describes the decision boundary itself. Because the feature coefficients of the logistic regression model encode the relative importance of each feature, we present them as strings of '+' and '-' in our explanations, as shown in Figure 1. To do this, we discretized them into 11 buckets, based on the range of the maximum and minimum coefficient. This type of explanation is *global* in that the decision boundary is a property of the classifier, and thus will be described in the same way for all samples.

4.2.2 Demographic-based Explanation. describes the structure of the training data and how it is distributed with respect to the decision boundary. We simply summarize, for training data matching

---

[3]The raw data classifier's confidence on the disparately impacted sample group had an average of 52% and a max of 54%. The processed data classifier had average and max confidence both at 50%.

[2]We split the data into training set (4222 samples) and testing set (1056).

Sensitivity
- Iliana's race is **African American**.
If it had been **Caucasian**, she would have been predicted as NOT likely to reoffend
- Iliana's age is **18-29**.
If it had been **older than 39**, she would have been predicted as NOT likely to reoffend

Input-Influence
The more **+**s/**-**s means a person with that attribute is more/less likely to re-offend.
* Appears next to Iliana's attributes
Race
- Caucasian (0)
- *** African-American (+)**
Age
- *** 18-29 (++++)**
- 30-39(**+**)
- …
Charge degree:
- …
Number of prior convictions
Has juvenile priors:

Defendant: Iliana
- Race: African-American
- Age: 18-29
- Charge degree: Misdemeanor
- Prior convictions: 0
- Has juvenile priors: Yes

Prediction:
  Likely to reoffend

Case
The training set contained 10 individuals identical to Iliana

6 of them reoffend (60%)

Demographic
The prediction is based on the likelihood of previous cases with different attributes re-offended or not.
A * appears next to Iliana's features.
Race
- 40% in Caucasian race group re-offended
- *** 55% in African-American race group re-offended**
Age
- *** 58% in 18-29 age group re-offended**
- 49% in 30-39 age group re-offended
- …
Charge degree:
- …
Number of prior convictions
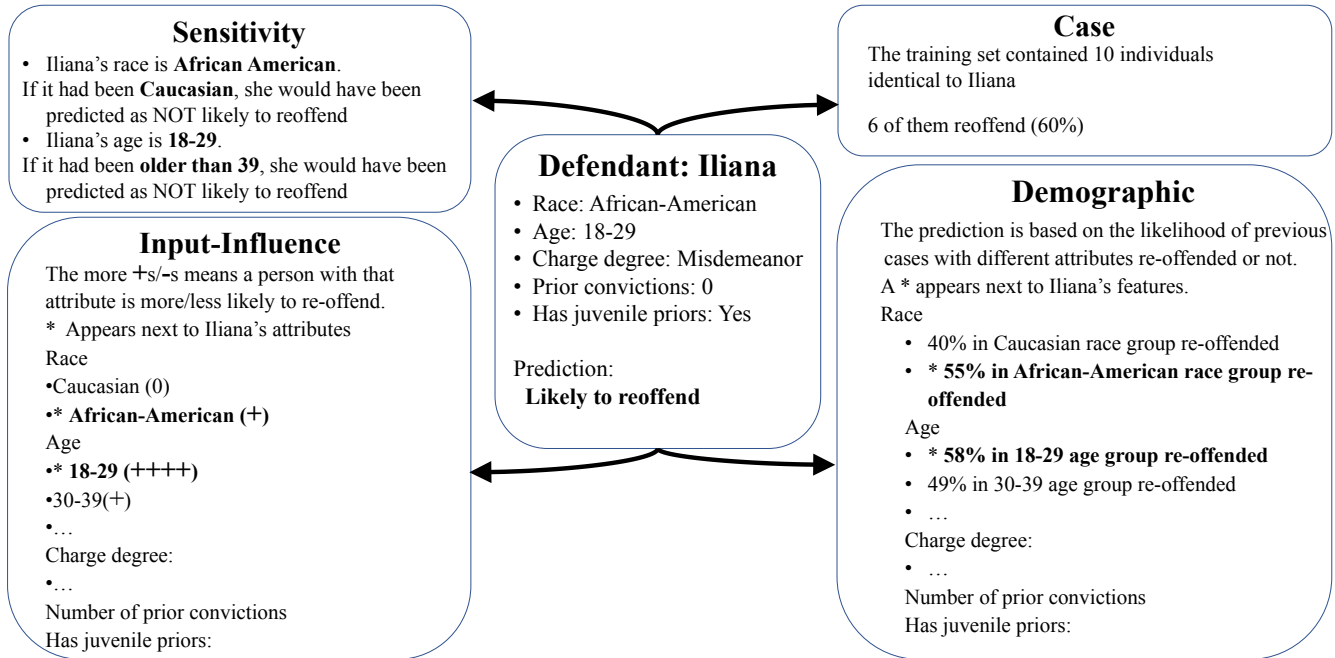Has juvenile priors:

Figure 1: Examples of explanations generated using the raw data classifier, adjusted and truncated for brevity. Consult our supplementary materials for full explanation output from both classifiers – as seen by participants (including * and highlights).

each feature category, the percentage with the same label as predicted for the presented example. This type of explanation is *global* and generates the same description for all samples on each side of the decision boundary.

4.2.3 Sensitivity-based Explanation. seeks to modify the presented sample along each feature until the prediction changes. When the prediction does change, we report back to the user the necessary feature change to produce the change in output. This type of explanation is *local* as it is specific to each presented example, and justifies the decision by indicating changes needed to produce a different output.

4.2.4 Case-based Explanation. does a nearest neighbor search in the training data to find similar cases. Since our study has a large data set with respect to the feature space, we frequently find neighbors occupying the same feature space location as the sample presented for explanation. When this is the case, we show the % of those neighbors with the same label as the prediction. When no exact matches are found, we simply show the features and label for the nearest neighbor in the training data. This is a modification to the design in Binns et al., which describes only a single identical or similar case. This explanation is *local*, and attempts to justify the decision by indicating similar examples with similar outputs.

5 METHODOLOGY
Our study adopted a mixed design by having data processing (raw/processed) and explanation styles (4 styles) as between-subject variables, and disparate impact as a within-subject variable. Each participant completed 6 fairness judgment trials, where each trial consisted of judging a single case. 2 cases were randomly selected from the 8 disparately impacted cases, and we sampled 4

from the 16 non-impacted cases in the test data. The order of the trials was randomized.

In September 2018 we recruited 160 Amazon Mechanical Turk workers, with the criteria that the worker must live in the US and have completed more than 1000 tasks with at least a 98% approval rate. They were randomly assigned to the 8 conditions (2 data processing treatments × 4 explanations). Among them, 62.5% are male, 78.8% are self-identified as Caucasian, 29.4% are under 30, and 13.3% are above 50. In our quantitative analysis, we included participants' race and gender as co-variance but observed no significant effect.

5.1 Study Procedure
We conducted an online survey-style study. Participants first provided informed consent, then reviewed a scenario about a ML system developed to assess defendants' re-offense risk, to help a judge make bail decisions. Each participant was given 6 trials, and within each trial, the participant would first review information about an individual (Figure 1, center region), then make his or her own prediction about whether that individual would re-offend. This step was to ensure they carefully consider the features used in the prediction. Subsequently, they would be presented with the prediction from the machine learning classifier with one of the four styles of explanations. They were asked to rate agreement with: "*How the software made the prediction was fair*" based on a 1 (Strongly Disagree) to 7 (Strongly Agree) Likert Scale, together with an open question justifying their rating.

We checked attention after the third trial, by asking participants to find the feature not used by the software from a list of choices. All participants passed the check. After completing all tasks, participants would answer a survey measuring their individual differences,
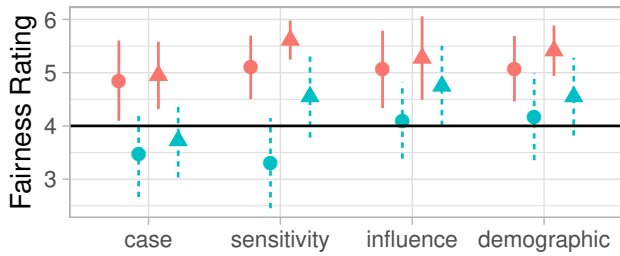
Figure 2: Overall mean ratings of fairness, per explanation type, data process treatment *(raw=●, processed=▲)*, and sample group *(impacted=blue dashed lines, non-impacted=red solid lines)*. The lines indicate the 95% confidence intervals.

and collecting demographic data. On average the study took 18 min to complete, and each participant was compensated with $3.

5.2 Individual Differences

We measured two types of individual differences: prior positions on algorithmic fairness, and cognitive style. For prior positions, we further differentiate between one's general position on using ML systems to assist decision-making, and position on the fairness of using race as a feature. We use a semantic differential scale that is often used to measure attitude on controversial topics [24]. Specifically, we ask participants to rate the statements "*Considering race as a factor in criminal risk prediction*" and "*Using machine learning software to replace or augment human decision making*", on 5 pairs of bipolar adjectives such as "*harmful-beneficial*" and "*unethical-ethical*" with a 7-point Likert scale. The ratings of the 5 pairs are averaged to represent the individual's prior position. The Cronbach Alpha for the scale on general ML position is 0.95, and for the race feature is 0.98, showing excellent consistency. For cognitive style, we chose to measure one's Need for Cognition using only 4 of the items from the standard scale [7], due to the time constraints.

6 RESULTS: QUANTITATIVE

We start with examining the effect of explanation style, data processing (*raw/processed*) and disparate impact (*true/false*) on participants' fairness judgment (RQ1). We then explore how individual differences, including prior position on fairness of ML, prior position on fairness of using the race feature, and need for cognition, make a difference on the judgment (RQ2). All statistical analyses were done in R. The lmerTest package was used to run mixed-model regressions.

6.1 Explanation, data processing, and disparate impact

Given the complexity of the statistical model, we first describe the trends with the descriptive data, then report statistical testing results. In Figure 2, we plot the mean and the 95% confidence interval of the mean of fairness ratings in all experiment treatments, showing several trends:

(1) Predictions made on the processed data (triangles) were rated fairer than those on the raw data (circles). It suggests that participants perceived fairness issues for the model trained on the raw data, and the processing technique mitigated the problem.

(2) Predictions made on cases with disparate impact (blue dashed lines) were rated less fair than those without it (red solid lines). This shows participants' fairness perceptions align with the presence of a fairness discrepancy between groups.

(3) Explanation styles made nuanced differences. As expected, the two local explanations led to higher discrepancy of fairness ratings between disparately impacted cases and non-impacted cases (difference between the dashed and solid lines) than the two global explanations. Thus, the former are more effective in exposing case-specific fairness issues. Moreover, this difference is most prominent for sensitivity-based explanations applied to raw data. This could be caused by sensitivity-based explanation being the most explicit in exposing disparate impact, while data processing mitigated the problem.

We now report the statistical significance of these observed trends. In particular, to validate that sensitivity-based explanation is most effective in exposing the disparate impact issue in the raw data, we expect to see a three-way interaction between explanation style, data processing, and disparate impact. We construct a mixed-effect regression model with the three-way interaction (and all the lower order interactions) as fixed effects, and participant as a random effect. We control for gender and race as covariances and neither has significant effect. The three-way interaction we expected is *not* significant, $F(3, 152) = 0.54$, $p = 0.66$. There is a marginally significant[4] two-way interaction between explanation style and disparate impact, $F(3, 152) = 2.35$, $p = 0.07$, and significant main effect of disparate impact, $F(1, 152) = 103.25$, $p < 0.001$, and data processing, $F(1, 152) = 4.65$, $p = 0.03$.

The main effect of data processing and disparate impact prove statistical significance for the first two observed trends. The two-way interaction indicates that explanation styles had differential impact on exposing the disparate impact issue. We conduct pairwise comparison for this interactive effect to identify between which explanation styles this perceived fairness discrepancy significantly differ. We found that if we use sensitivity-based explanation as the reference level, influence-based explanation is significantly different, $F(1, 156) = 5.14$, $p = 0.02$, and demographic-based explanation is marginally significant, $F(1, 156) = 3.29$, $p = 0.07$; If we use case-based explanation as the reference, influence-based explanation is marginally different, $F(1, 156) = 3.36$, $p = 0.07$. These results validate the observation that local explanations are more effective than global ones in exposing fairness discrepancies in different cases.

While we did not find statistical significance of the three-way interaction that validates the effectiveness of sensitivity-based explanation, a possibility is that there are individual differences for which the model did not account. In the next section, we explore that possibility.

---

[4]We consider $p < 0.05$ as significant, and $0.05 \leq p < 0.10$ as marginally significant, following statistical convention [12]
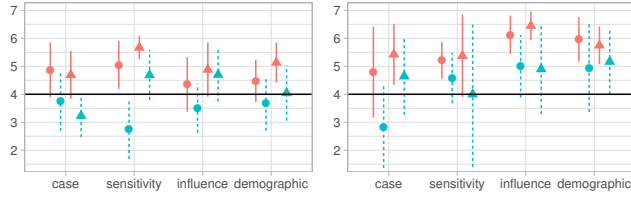
Figure 3: Same data as Figure 2, split by prior position on the fairness of using the race feature. Left: Participants that consider using race "Unfair" *(race_pos < 4)*. Right: Participants that consider using race "Fair" or neutral *(race_pos >= 4)*.

Figure 4: Overall mean fairness ratings, broken down by prior position on "Trust in ML" *(high trust=×, low trust=◇)*.

6.2 Individual differences

We enter the following factors into the model: prior position on using machine learning to assist decision-making (*ML position*), prior position on fairness of using the race feature (*race position*), and *need for cognition*. We start from four level interactions of each of the individual difference factors with the three manipulated variables (explanation, data processing, disparate impact), and then iteratively reduce it to lower-level interactions if it is not significant. We eventually arrive in a model with the following terms: a four way interaction between race position and the three manipulated factors, $F(3, 144) = 2.59$, $p = 0.05$, and a marginally significant two-way interaction between ML position and explanation style, $F(3, 137) = 2.43$, $p = 0.07$. We did not find need for cognition to make a difference and removed it.

By including these individual difference factors in the model, we now find the three-way interaction between explanation style, data processing, and disparate impact to be significant, $F(3, 144) = 2.96$, $p = 0.03$ (its lower-level two-way interactions as well). In addition to the main effect of data processing ($F(1, 137) = 4.68$, $p = 0.03$) and disparate impact ($F(1, 144) = 28.86$, $p < 0.001$) as in the original model, we also find a main effect of ML position ($F(1, 137) = 17.31$, $p < 0.001$), race position ($F(1, 137) = 6.43$, $p = 0.01$), and a marginally significant main effect of explanation style, $F(3, 137) = 2.11$, $p = 0.10$.

The above significant three-way and four-way terms, after including race position in the analysis, demonstrate that the consideration of this individual factor "de-noised" the data. In other words, it is only when an individual considers using race to be unfair, that a sensitivity-based explanation like this–"*If Nolan had been 'Caucasian', he would have been predicted to be NOT likely to re-offend*"–heightens the concern and significantly lowers the perceived fairness. When an individual does not consider it problematic to use race as a decision factor, they would not perceive such an explanation negatively. This trend is illustrated in Figure 3, where we separate participants who considered the race factor unfair and those who considered it fair-to-neutral (33.1% of all participants). In fact, for those who consider race to be a fair or neutral feature to use (Figure 3, Right), they did not perceive predictions made on the raw data (circles) to be less fair than processed data (triangles), and generally rated fairness to be higher (thus the main effect of prior position on race).

The main effect of ML position and its interactive effect with explanation style indicates that a general positive position on algorithmic fairness enhanced perc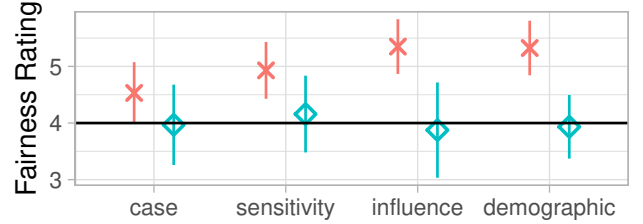eived fairness, and also led to different explanation preferences. We conducted pairwise comparison between styles of explanation, and found this interactive effect with ML position to be significant for influence-based explanation, $F(1, 148) = 6.25$, $p = 0.01$, and marginally significant for demographic-based explanation, $F(1, 148) = 2.77$, $p = 0.10$, if using case-based explanation as the reference. It is significant for influence-based explanation, $F(1, 148) = 3.73$, $p = 0.05$, if using case-based explanation as the reference. This implies that people who trust ML systems gain *even higher* confidence in the fairness of a prediction given global explanation (Figure 4).

It is worth noting that after controlling for these individual factors, we now see a marginally significant main effect of explanation style. Pairwise comparisons show that case-based explanation was rated marginally significantly less fair than influence-based ($F(1, 153) = 3.51$, $p = 0.06$) and demographic-based explanation ($F(1, 148) = 3.20$, $p = 0.08$). We consider it as evidence that case-based explanation is seen as generally less fair.

To summarize, in response to RQ1 and RQ2, we found evidence that: 1) Case-based explanation is seen as generally less fair; Global explanations further enhance perceived fairness for those who have general trust for machine learning systems to make fair decisions. 2) Local explanations are more effective than global explanations at exposing case-specific fairness issues, or fairness discrepancies between different cases. Sensitivity-based explanations are the most effective in exposing the fairness issue of disparate impact made by a particular feature—but only if the individual views using that feature as unfair. 3) In general, we show that individuals' prior position on ML trust and feature fairness have significant impact on how they react to explanations, and possibly more so than differences in cognitive styles.

7 RESULTS: QUALITATIVE

Along with collecting fairness ratings, we asked participants to justify their judgment. The authors reviewed this data and used open coding to extract themes in the answers. Here we discuss two groups of themes. One is to understand how participants made fairness judgments. Another is on participants' feedback for the four styles of explanations.

7.1 How is fairness judgment made?

In the open-ended answers, we investigated the criteria participants used to judge fairness. We see variations in reliance on the provided explanations, and depth of reasoning about the algorithm's

processes. These results provide further evidence of individual differences in the *criteria* used to make fairness judgments of ML systems.

7.1.1 General trust or distrust in ML systems. Some participants provided reasons not specific to a case or explanation, but that a general trust or distrust of ML systems dominated their judgment, as they tended to give consistent ratings across cases. Reasons for a general trust include "*based on objective data is better than subjective opinions*" (CR-31)[5], "*large data set*" (CR-37), "*uses statistics based on prior knowledge to make a judgment*" (IR-176). In contrast, some participants considered generalization by statistics is unfair to begin with–"*it might be unfair to group everybody together - makes more sense for the judge to have individual judgment.*" (IP-184), while others think that "*there needs to be a human element to the decision*" (CR-62). These observations corroborate Binns et al. [5] and further validate that participants varied on their general position on using ML system for criminal justice, and it influenced their fairness judgment.

7.1.2 Features used. Participants frequently cited features used by the algorithm as reasons for fairness or unfairness. Some explicitly differentiated between the process of the algorithm and the feature considered–"*The software makes it's decisions based on it's algorithm, so I believe it is fair and impartial on that account. However, some of the categories it is programmed to consider, such as age and race, are unfair*" (SP-71). It is interesting to note that we observe individual differences in the position on the fairness of race feature in the qualitative results as well. While many participants called out the problem of considering race, a few participants who saw the processed data commented that "*[if] race was not a predictor [it] may not accurately reflect the reality*" (DP-68). There is also some controversy on using age and juvenile priors as features. Participants' comments echo results from a previous study [14] showing that people consider multiple dimensions (e.g., relevance, disparate outcome, volitionality) in their judgments about the fairness of features used in decision-making algorithms, and individuals weigh these dimensions differently.

7.1.3 Lacking features. As observed in [5], several participants criticized the limited features used in our simple model. Some suggested to have more detailed information on current features, such as "*frequency of priors or the interval of time since the last prior in order to get a more accurate assessment of what one's prior record means*" (DP-53). Others are less optimistic about the possible sufficiency of features to ensure fairness–"*software cannot fully take into account environmental factors that cause people to go down a bad path*" (CR-76).

7.1.4 Prediction process. Many participants based their fairness judgment on their understanding of the algorithms' process. Some, especially those presented with global explanations, closely examined explanation details, e.g. "*Software seems to be flawed in major areas... improper weighing of distant vs recent past, and a questionable choice of how to evaluate probabilities in each case*" (DR-119). Some also considered failure to account for external factors, e.g.

"*the number may be relatively accurate for the race and charge degree categories, but if the [past] laws were different they would probably be higher*" (DR-107). Moreover, multiple participants attributed their low fairness ratings to insufficient understanding of process, or "*'how' the data is used*" (DR-172).

7.1.5 Data issue. A few participants questioned the underlying data used. Almost all of them were in either the demographic- or case-based explanation conditions, as these two styles leverage information about distributions of similar cases to explain the decision. For example, "*'Not re-offend' rate for African Americans is a little low. I think the percentage may be higher in reality... data could have been biased*" (DR-107).

7.2 Explanation styles

Below we summarize codes that are prominent for each explanation style. These results could help us better understand the benefits and drawbacks of each explanation style, and inform future work on designs of ML explanations.

7.2.1 Influence based. It is a global explanation that faithfully describes how each feature contributes to the algorithm's decision-making process. We observed that this explanation prompted more comments on details of the process, such as the weights of different features, and the trends with regard to different categories of a feature, e.g. "*it is fair because it doesn't discriminate by race, but rather on age and prior convictions... if someone exhibits a behavior pattern it is likely to will continue, and I think people who are young are more apt to take risks*" (IP-208). On the one hand, the detailed description of the algorithm process adds to the confidence in participants' fairness judgments, which may help explain its enhancement of fairness perception among those trusting ML algorithms. On the other hand, it exposes more information for people to scrutinize, and thus subject to critiques from the heterogeneous standards of fairness.

7.2.2 Demographic based. This is a type of global explanation that does not expose the *process* of the algorithm, but justifies the decision with the data distributions. Sometimes, the distributions were seen as convincing, e.g. "*The high percentage of people with more than 10 prior convictions who end up reoffending was staggering, and justifies the prediction*" (DP-57). Other times, participants found its explanation of the process inadequate, as the percentages do not clearly connect to an outcome–"*The percentages aren't high enough. It could go either way*" (DR-157). Sometimes it also directed participants attention to the potential biases of underlying data.

7.2.3 Sensitivity based. The main benefit of sensitivity-based explanation seems to be its conciseness and explicitly directing attention to features relevant to the particular decision. It appears to be convincing and easy to process when a decision is uncontroversial –"*The rationale is so basic (no prior offenses) that it has to be fair*" (SR-138). "*It's taking into consideration everything that we would and puts it into an easy to read manner*" (SP-220). Consistent with our quantitative results, for disparately impacted cases where the race factor is explicitly mentioned, sensitivity-based explanation heightens the concern and was perceived most negatively–"*It says that in the same situation, if the offender were African-American*

---

[5]Participant IDs give treatment info, explanation (**S**ensitivity, **C**ase, **I**nput-Influence, **D**emographic) followed by data processing (**R**aw, **P**rocessed).

rather than Caucasian, they would have been likely to offend. This is racial profiling and inaccurate in my opinion." (SR-124).

7.2.4 Case based. As we found in the quantitative results, case-based explanation was judged to be the least fair—and the qualitative results provided reasons. First, some found it to provide little information about *how* the algorithm arrives at a conclusion. Second, the number of identical cases and the percentage of cases supporting the decision are often considered too small to justify the decision–"*It was unfair for the defendant because she was compared to only 22 other identical individuals... not to mention that only a little over 50% reoffended.*" (CR-61). This observation is consistent with Binns et al. [5], however, our work is based on the actual output of a ML model trained on a real dataset – allowing us to empirically show a limitation of case-based explanation[6]. Lastly, we found variations in individuals' positions on the fairness of the "explained process" (as opposed to the actual algorithm process) to make decisions based on identical cases. While some people consider it to be fair to "*compare the actions of people with similar history and backgrounds*" (CP-200), others questioned the underlying rationale such as "*is anyone really identical if more things considered*" (CP-201).

8 DISCUSSION

8.1 Supporting different needs of fairness judgment

The most important take-away from our study is that there are multiple aspects and heterogeneous standards in making fairness judgments, beyond evaluating *features*, as studied in previous work [14]. Our experiment highlights two types of fairness issues: unfair models (e.g., learned from biased data), and fairness discrepancy of different cases (e.g., in different regions of the feature space). Our qualitative results further illustrate that algorithmic fairness is evaluated by various dimensions including data, features, process, statistical validity, as well as broader ethical and societal concerns.

Our results highlight the need to provide different styles of explanation tailored for exposing different fairness issues. For example, we show that local explanations are more effective in exposing fairness discrepancies between different cases, while global explanations seem to render more confidence in understanding the model and generally enhance the fairness perception. Hybridizing the two techniques reveals a possible human-in-the-loop workflow; using global explanations to understand and evaluate the model, and local explanations to scrutinize individual cases.

It is critical to note that different regions of the feature space may have varied levels of fairness and different types of fairness issues. This calls for development of fine-grained sampling methods and explanation designs to better support fairness judgment of ML systems. To that end, we envision an active-learning paradigm for fairness improvement, where the system interactively queries the human for fairness judgment of its predictions, together with explanation options, then optimizes the algorithm based on user feedback.

Our qualitative results suggest another useful categorization of explanation styles: *process oriented v.s. data oriented explanation.* The case- and demographic-based explanations we studied leverage information on data distribution to justify its decision but reveal less on *how* the decision was made. Influence- and sensitivity-based explanations link each feature to the decision. We observe a general preference for process-oriented (*how*) explanations, although a focus on data has the potential benefit of directing attention to issues in the data and dilutes the "blame" on the algorithms.

8.2 Individual differences and descriptive fairness

Another contribution of our study is to empirically demonstrate how individuals' prior positions on algorithmic fairness impact their reaction to different explanations. We differentiate between a general position on algorithmic fairness, and position on fairness of a particular feature used.

The difference between normative (prescriptively defining what is fair) versus descriptive fairness and its implication for algorithmic fairness has been discussed in previous work [14]. Empirically, we show that even though race is considered a protected variable, individual positions on its fairness still vary (close to one third of participants considered it neutral or fair to use). This indicates a lack of agreement on the meaning of moral concepts, a result Binns et al. [5] hinted at qualitatively. In different contexts, an algorithm developer may have to choose between a normative or a descriptive position of fairness, and it is important to be aware of the variation of fairness position in the population. For example, if a ML system takes a normative position and aims to eliminate pre-defined biases based on people's feedback, it may need to account for their prior positions to weigh the feedback differently. It may be arguable whether explanation should always attempt *soundness* and *completeness* for all individuals. On the other hand, if a system aims to provide optimal decision support for individual needs, it would be useful to provide mechanisms for individuals to express their prior positions as direct input for the algorithm (similar to the idea of active-learning by tuning features [29, 31]).

8.3 Limitations

We performed our study with crowdworkers, rather than judges who would be the actual users of this type of tool. Additionally, there are many styles and elements of explanations not studied here. One important element not studied is confidence, which we declined to present to participants because we could not control for it.

9 CONCLUSION

Our work provides empirical insights on how different styles of explanation impact people's fairness judgment of ML systems, particularly the differences between a global explanation describing the model and a local explanation justifying a particular decision. We highlight that there is no one-size-fits-all solution for effective explanation, but depends on the kinds of fairness issues and user profiles. Providing hybrid explanations, allowing both overview of the model and scrutiny of individual cases, may be necessary for accurate fairness judgment. Furthermore, we show that individuals'

---

[6]We found that 16% of the test data exhibited the failure mode of *contradicting* the claim (< 50% of individuals with identical features share label). Meanwhile *insufficient justification* of the claim (between 45% and 55% label matches) was quite common, with 24% of the test data. The prevalence of these failure modes indicates inherent "unsoundness."

prior positions on algorithmic fairness influence how they react to different explanation types. The results call for a personalized approach to explaining ML systems. However, specific to fairness, ML systems may need to take a normative or descriptive position in different contexts, which may differentially require corrective or adaptive actions considering individual differences in their fairness positions.

ACKNOWLEDGMENTS

REFERENCES

[1] 2016. Artificial Intelligence's White Guy Problem. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html Accessed: 6/8/2018.

[2] 2017. Rise of the racist robots–how AI is learning all our worst impulses. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses Accessed: 6/8/2018.

[3] 2018. COMPAS Recidivism Risk Score Data and Analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis Accessed: 6/8/2018.

[4] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 377, 14 pages. https://doi.org/10.1145/3173574.3173951

[6] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior* 36, 1 (2009), 21–40. https://doi.org/10.1177/0093854808326545 arXiv:https://doi.org/10.1177/0093854808326545

[7] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307.

[8] Toon Calders and Indrė Žliobaitė. 2013. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*. Springer Berlin Heidelberg, Berlin, Heidelberg, 43–57. https://doi.org/10.1007/978-3-642-30487-3_3

[9] 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 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*. 2172–2180.

[11] William J Clancey. 1983. The epistemology of a rule-based expert system–a framework for explanation. *Artificial intelligence* 20, 3 (1983), 215–251.

[12] Duncan Cramer and Dennis Laurence Howitt. 2004. *The Sage dictionary of statistics: a practical resource for students in the social sciences*. Sage.

[13] Philip M Fernbach, Steven A Sloman, Robert St Louis, and Julia N Shube. 2012. Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research* 39, 5 (2012), 1115–1131.

[14] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 903–912. https://doi.org/10.1145/3178876.3186138

[15] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[16] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.

[17] 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.

[18] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*. 2280–2288.

[19] Gary Klein, Louise Rasmussen, Mei-Hua Lin, Robert R Hoffman, and Jason Case. 2014. Influencing preferences for different types of causal explanation of complex events. *Human factors* 56, 8 (2014), 1380–1400.

[20] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. K. Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[21] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm Accessed: 6/8/2018.

[22] 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. https://doi.org/10.1177/2053951718756684 arXiv:https://doi.org/10.1177/2053951718756684

[23] Xiaodan Liang, Liang Lin, Xiaohui Shen, Jiashi Feng, Shuicheng Yan, and Eric P Xing. 2017. Interpretable Structure-Evolving LSTM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1010–1019.

[24] Q Vera Liao and Wai-Tat Fu. 2014. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2745–2754.

[25] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128.

[26] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[27] Tim Miller. 2017. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269* (2017).

[28] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR* abs/1802.00682 (2018). arXiv:1802.00682 http://arxiv.org/abs/1802.00682

[29] Hema Raghavan, Omid Madani, and Rosie Jones. 2006. Active learning with feedback on features and instances. *Journal of Machine Learning Research* 7, Aug (2006), 1655–1686.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[31] Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1467–1478.

[32] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 82–91.

[33] William R Swartout. 1985. Explaining and justifying expert consulting programs. In *Computer-assisted medical decision making*. Springer, 254–271.

[34] Alan B Tickle, Robert Andrews, Mostefa Golea, and Joachim Diederich. 1998. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks* 9, 6 (1998), 1057–1068.

[35] 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 (CHI '18)*. ACM, New York, NY, USA, Article 440, 14 pages. https://doi.org/10.1145/3173574.3174014

[36] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. (2017).

[37] Michael R Wick and William B Thompson. 1992. Reconstructive expert system explanation. *Artificial Intelligence* 54, 1 (1992), 33–70.

[38] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web

Conferences Steering Committee, 1171–1180.

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

[40] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.