Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation

Clarice Wang
The Harker School
San Jose, CA, USA
22claricew@students.harker.org

Kathryn Wang
Carnegie Mellon University
Pittsburgh, PA, USA
kathryn1wang@gmail.com

Andrew Y. Bian
River Hill High School
Clarksville, MD, USA
abian1960@inst.hcpss.org

Rashidul Islam
University Of Maryland, Baltimore County
Baltimore, MD, USA
islam.rashidul@umbc.edu

Kamrun Naher Keya
University Of Maryland, Baltimore County
Baltimore, MD, USA
kkeya1@umbc.edu

James Foulds
University Of Maryland, Baltimore County
Baltimore, MD, USA
jfoulds@umbc.edu

Shimei Pan
University Of Maryland, Baltimore County
Baltimore, MD, USA
shimei@umbc.edu

ABSTRACT

Currently, there is a surge of interest in *fair Artificial Intelligence (AI) and Machine Learning (ML)* research which aims to mitigate discriminatory bias in AI algorithms, e.g. along lines of gender, age, and race. While most research in this domain focuses on developing fair AI algorithms, in this work, we examine the challenges which arise when human- fair-AI interact. Our results show that due to an apparent conflict between human preferences and fairness, a fair AI algorithm on its own may be insufficient to achieve its intended results in the real world. Using college major recommendation as a case study, we build a fair AI recommender by employing gender debiasing machine learning techniques. Our offline evaluation showed that the debiased recommender makes fairer and more accurate college major recommendations. Nevertheless, an online user study of more than 200 college students revealed that participants on average prefer the original biased system over the debiased system. Specifically, we found that the perceived gender disparity associated with a college major is a determining factor for the acceptance of a recommendation. In other words, our results demonstrate we cannot fully address the gender bias issue in AI recommendations without addressing the gender bias in humans. They also highlight the urgent need to extend the current scope of fair AI research from narrowly focusing on debiasing AI algorithms to including new persuasion and bias explanation technologies in order to achieve intended societal impacts.

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03…$15.00
https://doi.org/10.1145/3490099.3511108

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

ACM Reference Format:
Clarice Wang, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2022. Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3490099.3511108

1 INTRODUCTION

Artificial Intelligence (AI) is increasingly used in consequential decision making, but many recent discoveries have shown that AI systems often exhibit discriminatory bias in their behavior, particularly along gender, age, and racial lines [3, 12, 20, 61]. For example, an AI tool that helps judges assess the risk of an incarcerated individual committing a crime in the future was found to be biased against African Americans [3]. In other domains including personalized search, ads and recommendation, AI systems lead to skewed outcomes [2]. AI models trained on text data have been found to encode gender stereotypes such as associating computer programming with men and homemaking with women [10], which could potentially impact AI-based career counseling and automated hiring decisions. Indeed, Amazon had to scrap its AI recruiting tool because it was found to be biased against women [20].

Since biased AI systems can be discriminatory against vulnerable populations in our society and/or reinforce harmful stereotypes, there are strong motivations to develop AI debiasing interventions [14]. We argue that there are two aspects which must be considered when it comes to mitigating AI bias: the algorithmic aspect and the human- fair-AI interaction aspect. Most existing approaches in the AI community focus on the algorithmic aspect. The mission of the field has primarily been to develop *(a)* new

metrics that quantify fairness [24, 48], and *(b)* new machine learning techniques that remove bias from AI models [21, 27, 85]. On the other hand, the human- fair-AI interaction aspect as well as its broader social context are equally important and significantly understudied. For example, in many contexts such as targeting ads on search and social network platforms it is well understood that there is a tension between building a fair system and achieving the platform's own revenue goals [60], and this tension cannot be resolved by algorithms alone. As bias in AI may arise from the human side of the socio-technical system via systemic bias and/or human prejudice encoded in data [7], will bias mitigation be effective if we focus only on removing the bias in AI without addressing the bias in humans?

In this research, we systematically study the interplay between AI debiasing techniques and human bias, using AI career recommendation as a case study. Here, we define an AI career recommender as a recommender system that is capable of automatically recommending career-related items such as college majors, job openings and job candidates although the system we built in this study focused specifically on recommending college majors. The conclusions we draw from the study are primarily applicable to recommending college majors to college students.

Career choice is a major part of human life, as it often defines one's economic success, social standing, and quality of life. Humans have a tendency to associate masculine and feminine traits with specific careers, which results in the perception that certain genders are better suited for certain occupations [78]. Machines that learn from career decisions made by humans are thus expected to be influenced by the resulting gender gaps in career choices unless bias mitigation techniques are performed [20].

In this study, we first dive into the algorithmic aspect of the problem, using machine learning to systematically mitigate bias in AI systems so that they do not reinforce harmful stereotypes. Second, we examine the human- fair-AI interaction aspect of the problem. We perform a user study to investigate whether users will typically prefer a fair AI system with gender bias removed over a biased one. Our results show that users' acceptance of a debiased AI system can be influenced by their own biases. The following are the main contributions of the paper.

- While the fact that human biases may impact one's career choice is not entirely new, it has largely been neglected by the fair AI/ML research community, and to the best of our knowledge, this is the first systematic and rigorous study on the role of human bias plays in human- fair-AI interaction. We conducted a user study to illustrate how human bias interferes with the effectiveness of a fully implemented fair AI college major recommender.
- As a case study, we develop a debiased recommender system which mitigates gender bias in college major recommendations [42]. Our offline evaluation shows that the debiased recommender is both fairer and more accurate than the gender-aware biased recommender.
- We conduct an online user study with over 200 college students to understand their acceptance of the debiased system. The results indicate that participants in general prefer the gender-aware (biased) career recommender over the gender-debiased (fair) one. We analyzed the role a participant's own bias plays in his or her acceptance of the recommendations. Our results indicate that the perceived gender disparity associated with a recommended career is significantly correlated with its acceptance.
- Based on the study results, we recommend a few new areas of fair AI research to address the issues we uncovered in this study including AI-based human bias assessment, AI bias/fairness explanation and behavior "nudge" via novel persuasion technologies.

In the rest of the paper, we describe related literature, the implementation of a debiased machine learning algorithm for fair college major recommendation, an offline evaluation of the system, an online user study for understanding the relationship between human bias and the acceptance of a fair AI system, and a discussion on new fair AI technologies that are needed in developing an effective fair AI system.

2 RELATED WORK

In this section, we summarize recent work on fair AI and ML, review social science studies on the relationship between gender bias and career decisions, and briefly discuss the work in the HCI community on AI bias/fairness.

2.1 Fair AI and ML Research

Recently, there has been a sharp focus in the AI community on how to prevent AI from perpetuating or, worse, exacerbating social unfairness. Most efforts concentrate on (1) developing metrics to quantify the bias in data as well as in ML algorithms, and (2) developing fair ML algorithms that mitigate these biases.

It is difficult to develop a universal definition of fairness because fairness/bias is a complex, multifaceted concept whose definition heavily depends on the social, culture and application context. Consequently, many definitions have been proposed. In fact, AI Fairness 360, the IBM open source platform for fair machine learning, has 70+ metrics for fairness/bias [8]. Among them, some are about *individual fairness* and others are about *group fairness*.

Individual fairness seeks to ensure similar individuals get similar outcomes. Widely used individual fairness measures include *Fairness Through Awareness* ("An algorithm is fair if it gives similar predictions to similar individuals") [24], *Fairness Through Unawareness* ("An algorithm is fair as long as any protected attributes such as race, age, gender are not explicitly used in the decision-making process") [29], and *Counterfactual Fairness* ("a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group") [48].

Group fairness partitions a population into groups defined by protected attributes (or intersections of protected attributes) and seeks to ensure that statistical measures of outcomes are equal across groups/subgroups. Widely used group fairness measures include *Demographic Parity* ("The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group") [24], *Equalized Odds* ("the probability of a person in the positive class being correctly assigned a positive outcome and

the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members") [35], *Equal Opportunity* ("the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected group members") [35] and *Differential Fairness* ("the probabilities of the outcomes will be similar with respect to different subgroups defined by the intersections of multiple protected attributes such as race, gender and race") [28]. Although most fairness measures are defined for classification tasks, there are a few fairness measures for other AI tasks. For example, *Non-parity Unfairness* [82] is designed to evaluate the fairness of recommender systems.

In terms of mitigating bias in AI systems, one focus is to remove the bias from the data that is used to train AI models. For example, vector projection-based bias attenuation method is used to remove bias from word embeddings [21]. A convex optimization approach is used to pre-process and transform data to control discrimination, limit distortion, and preserve utility [13]. There is also a large body of work that optimizes ML models under both traditional accuracy-based and new fairness-based objectives jointly [1, 27, 80, 84]. Furthermore, adversarial learning has been used to improve model accuracy and at the same time minimize an adversary's chance of finding out protected attributes (e.g., gender and race) [85].

2.2 Social Science Research on Gender Bias and Careers

According to Glick et al. [30], gender, or the cultural construction of sex differences, is the "most automatic, pervasive and earliest learned" categorization that shapes social relations and identities. Social science research on gender bias and stereotypes in career choices consistently finds that gender-based differences in career selection exist even when controlling for measured competency and ability [16]. Career-related gender bias exists across country, culture and age. For example, even as kindergartners, girls select mostly traditional female careers such as teaching and nursing [73]. Scottish pupils were found to perceive Truck Driver, Engineer, Plumber/Electrician, Laborer, Armed Forces as "male" jobs while Nurse and Care Assistant "female" jobs. Boys, but especially girls, have strong preferences against working in sectors and industries that are traditionally the domain of the opposite gender [56].

Many theories have been developed to explain why gender bias exists in career selection. Some of them focus on psychological constructs (i.e., variables at the level of individuals), while others focus on socioeconomic conditions and cultural understandings of gender roles.

Social Cognitive Theory [4] and *Social Cognitive Career Theory* [50] are the most influential social cognitive frameworks for understanding individual human behavior as well as career decisions. They posit that human behavior is primarily explained through self-efficacy beliefs, outcome expectations, and goal representations. Self-efficacy beliefs refer to "people's judgment of their capabilities to organize and execute courses of action required to attain designated types of performances." Outcome expectations concern a "person's estimate that a given behaviour will lead to a certain outcome." Goal representations are defined as "determinations of individuals to engage in a particular activity." Of the three determinants, self-efficacy has the strongest influence on behavior. Gender difference in self-efficacy beliefs may explain observed gender bias in career choice. For example, it was found that women possess lower levels of mathematics confidence than men because women had fewer learning possibilities and role models to stimulate them [5, 50].

In addition to psychological constructs, social and cultural beliefs about gender may also influence the career choice of men and women. For example, gender beliefs are cultural schemas for interpreting or making sense of the social world. They represent what we think "most people" believe or accept as true about the categories of "men" and "women." Substantial evidence indicates that certain careers (e.g., mathematics) are often stereotyped as "masculine" [41, 57]. This cultural belief about gender channels men and women in substantially different career directions since it impacts the self-efficacy of individuals (e.g., belief about their own mathematical competence) [16].

In terms of overcoming gender bias in career decisions, the most commonly cited interventions include the availability of role models in the same social circle, especially same sex role models for women [38, 51, 56]. Encouragement from friends and family [49] also improves self-efficacy.

2.3 HCI and AI Fairness/Bias

Recent work on AI fairness/bias in the HCI community mostly focused on identifying and analyzing biases in AI systems such as image search [45, 65], social media analysis [44], image and persona generation [66, 67], sentiment analysis [22], text mining [19] and natural language generation [72]. In addition, Nourani et al. [63] investigated the impact of cognitive biases such as anchoring bias in Explainable AI (XAI) systems.

The HCI community also worked on fairness perception and definition. For instance van Berkel et al. [75] evaluated the effect of information presentation on fairness perceptions of Machine Learning predictors. Htun et al. [40] studied the relationship between personality and fairness perception in group music recommendation. Wang et al. [77] conducted an online experiment to better understand the perception of fairness, focusing on three sets of factors: algorithm outcomes, algorithm development and deployment procedures, and individual differences. McDonald et al. [55] studied students' perception of ethics and fairness in information systems. Hou et al. [39] and Woodruff et al. [79] explored how intended users, especially those marginalized by race or class, feel about algorithmic fairness. Dodge et al. [23] conducted an empirical study on how people judge the fairness of ML systems and how explanations impact that judgment. Chen et al. [15] tried to quantify fairness/biases by measuring the difference of its data distribution with a reference dataset using Maximum Mean Discrepancy.

There is a rich body of HCI work on mitigating bias in AI systems through the use of better system designs. One principled approach is known as equitable and inclusive co-design, which is about engaging diverse stakeholders especially underrepresented minorities directly in the design process [53, 59, 70, 76]. Furthermore, better tooling [17, 81] and better algorithms [6, 72] also help address the problem.

So far, there is little existing work focusing on what happens *after* biases in AI systems are identified and systematically removed. Does it automatically achieve its intended societal impact? Our work explores this domain.

3 DEBIASING AI CAREER RECOMMENDATIONS

We use an AI college major recommendation system as a case study to illustrate how a career recommender, which uses state-of-the-art debiasing techniques to systematically remove gender stereotypes from its recommendation, may not produce intended outcomes.

Recommender systems have gained widespread acceptance in the era of the internet, social network, and e-commerce. The basic idea of recommender systems is to infer user interest or behavior based on user-generated data. For example, collaborative filtering, a key method used in recommender systems, is based on the assumption that similar users have similar preferences of items [68]. Thus, an e-commerce recommender will recommend a product to a customer if it was purchased by customers who gave similar ratings to other products in the past. Following the same idea, an AI career counseling system recommends similar college majors (or jobs) to people who share similar interests.

A major source of bias/unfairness in machine learning outcomes arises from biases in the data [7, 58, 74]. A machine learning model trained on biased data may lead to unfair/biased predictions. For example, user preference data may encode real-world human biases (e.g., gender or racial biases). As a result, the AI model may inherit the bias into its recommendations, for example, by suggesting, as jobs, physicians for boys and nurses for girls. In the following, we describe the methodology used to implement a debiased algorithm for fair college major recommendation.

3.1 Backend Algorithm Implementation

The system is designed to make college major recommendations based on an individual's interests. The input to the system is a user's interests indicated on Facebook. The output contains the top college majors recommended by our system.

We employed an existing Facebook dataset widely used in social media research [46, 54, 69, 83]. The data was collected with an explicit opt-in consent for reuse for research purposes.

The dataset contains the following information of a large number of anonymized Facebook users: (a) the demographic information (e.g., gender, age), (b) the Facebook pages (a.k.a items) they "like," and (c) their declared college majors (a.k.a academic concentrations) such as Computer Science, Psychology, and Mechanical Engineering.

To a certain extent, a person's "likes" of Facebook pages, which cover a wide range of topics (e.g., books, music, movies, brands, sports, hobbies, famous people and proverbs/statements.), is a good indication of his/her interests and personality [83].

Since the college majors in our data are declared by Facebook users, they are quite noisy (e.g., "Defense Against the Dark Arts" is a declared major). To prepare the Facebook data to train our career recommender, we filtered out college majors that occurred less then 3 times in the data. The final data used in training and testing the backend system contains a total of 16,619 users (of which 60% are female, 40% are male and no gender non-binary individuals), 1,380 unique college majors, 143,303 unique items that a user can like on Facebook and 3.5 million+ user-item pairs.

Figure 1 shows the architecture of our backend system. To develop such a system, first, we train a neural collaborative filtering (NCF) [37] model for predicting the items a user "likes," encoded as 1, or 0 if otherwise. A user's gender is not taken into account during the training of the NCF model. We also included 10% negative instances to train the system from those user-item pairs marked as "0." In the input layer, the users and items are one hot-encoded (they are represented as vectors, each with a single "1" and all the others "0"). They are mapped into two separate embedding layers with embedding size of 100 (user and item embedding). Since NCF adopts two pathways to model users and items, the user and item embeddings are concatenated. One hidden layer with 10 linear units is added on the concatenated vector along with dropout regularization of probability 0.1, followed by a linear output layer. Finally, we train the model by optimizing MSE loss using *Adam* in batch mode with a learning rate of 0.001 for 20 epochs. Note that "relu" activation is used for the hidden and output layers. L2 regularization with tuning parameter 0.0001 is also used to optimize the loss for the NCF model.

We then study the use of the learned user embeddings to suggest academic concentrations by training a logistic regression classifier. We train the multi-class logistic regression model by minimizing a multinomial loss that fits across the entire probability distribution using stochastic average gradient descent in batch mode with a learning rate of 0.001 for 500 iterations. We further use L2 regularization with tuning parameter 0.0001 in the logistic regression model.

To gender-debias the recommendation, we add an extra debiasing step prior to applying logistic regression. Our debiasing approach adapts a recent work on attenuating bias in word embeddings [21]. Since traditional word embeddings are usually trained on massive text data, they inherit some of the human racial and gender biases from the data, as demonstrated by this well-known example, in which vector arithmetic on the embeddings solves an analogy task [11]:

$$\texttt{doctor} - \texttt{man} + \texttt{woman} = \texttt{nurse}\ .$$

The user embeddings we have trained experience a similar problem. Let p_u denote the embedding of a user, and let v_B, which is a unit vector in the same embedding space, denote the global gender bias in our system. We then debias p_u by removing p_u's projection on the gender bias vector v_B:

$$p'_u = p_u - (p_u \cdot v_B)v_B\ . \tag{1}$$

The question is how to find v_B. We consider v_{female}, given below, is the representation of an average female user:

$$v_{female} = \frac{1}{n_f}(f_1 + f_2 + \cdots + f_{n_f})$$

where $f_1, f_2, \cdots, f_{|n_f|}$ are the embeddings of female users. We define v_{male} in the same way. This allows us to derive the overall gender bias vector as:

$$v_B = \frac{v_{female} - v_{male}}{\|v_{female} - v_{male}\|}\ .$$

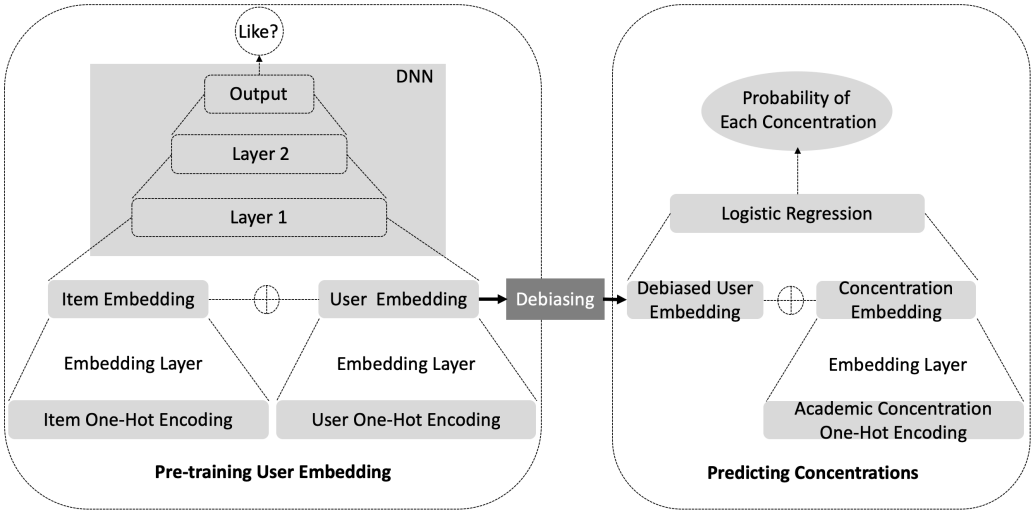

Figure 1: The Architecture of a Gender-Debiased Career Recommender

Please note that our vector projection-based bias attenuation method is not a simple "fairness through unawareness" method. It can systematically remove bias related to not only the sensitive variable (e.g., gender) but also all the proxy variables (e.g., like a particular brand such as "Victoria Secret" may be highly correlated with gender). In the college major recommendation phase, the objective is to suggest top-N academic concentrations to a new user based on the user's preference/interests indicated on Facebook. First, we construct the user embedding of the new user by analyzing the liked items of the user using the pre-trained NCF model. Then the embedding of the new user is used as the input features to the pre-trained logistic regression classifier. For the gender-debiased system, we dropped the intercept terms during prediction to further remove popularity bias [71]. Finally, the logistic regression model recommends top-N academic concentrations by sorting the probabilities of the 1380 majors for a given user.

4 OFFLINE EVALUATION

To compare the performance of the gender-debiased recommender with a gender-aware recommender, we implemented two variations of the same system. The **gender-aware system** makes career recommendations based on the choices by the people of the same gender (e.g., recommending to girls based on the career choices of other girls) while the **gender-debiased system** employs the linear projection-based gender de-biasing strategy to systematically remove gender stereotypes from user embeddings.

4.1 Evaluation Metrics

We employ the following performance measures to evaluate the accuracy and fairness of each system.

Normalized discounted cumulative gain at K (NDCG@K). This is a well-known metric for assessing the quality of a *ranked* list of results (e.g., recommendations) [36].

$$NDCG@K = \frac{\sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)}}{\sum_{i=1}^{|REL_K|} \frac{rel_i}{\log_2(i+1)}}$$

where

$$rel_i = \begin{cases} 1 & \text{if the recommendation at position } i \text{ is accepted,} \\ 0 & \text{otherwise.} \end{cases}$$

and REL_K is the ideally ranked list of career recommendations (ordered by rel_i) up to position K. In general, the higher the NDCG score is, the higher the prediction accuracy is.

Non-parity unfairness (U_{PAR}). This metric is designed to evaluate the fairness of recommender systems [82]. It computes the absolute difference of the average ratings between two groups of users:

$$U_{PAR} = |E_g[y] - E_{\neg g}[y]|$$

where $E_g[y]$ is the average predicted score from one group of users, and $E_{\neg g}[y]$ is the average predicted score for the other group of users. In our case, we consider scores for N academic concentrations for male and female subjects.

$$U_{PAR} = \frac{1}{N} \sum_{n=1}^{N} |E_{female}[y_n] - E_{male}[y_n]|$$

In general, the lower the U_{PAR} value is, the fairer the system is.

4.2 Experimental Settings and Results

To train an AI model to predict college majors, we need negative examples as well, that is, a college major that is not a good fit for a user. We generate random pairs (u, c) as negative training instances, where c is any academic concentrations not explicitly declared by u. Furthermore, we split the data and use 70% of it for training and the remaining 30% for evaluation.

Table 1 shows the evaluation results. Since the NDCG scores at position 3, 10 and 20 for the gender-debiased system are consistently

	NDCG@3 ↑	NDCG@10 ↑	NDCG@20 ↑	U_{PAR} ↓
GenderAware	0.0009	0.005	0.007	1.1445
GenderDebiased	**0.0050**	**0.010**	**0.013**	**1.1188**

Table 1: Offline Evaluation Results on the *Facebook* dataset. Higher Values are better for NDCG; lower values are better for U_{par}.

higher than those for the gender-aware system, the gender-debiased system is considered more accurate than the gender-aware system. In addition, since the gender-debiased system also has lower U_{PAR} score, it is considered fairer than the gender-aware system.

During the offline evaluation with well-established evaluation metrics for recommender system accuracy and fairness, we have demonstrated that the gender-debiased recommender is fairer without any loss of prediction accuracy, a highly desirable "fairness for free" situation [43] in bias mitigation. In summary, the gender debiased career recommender is considered a success based on typical measures of machine learning accuracy and fairness.

In the following, we describe a user study to investigate whether the gender-debiased career recommender can achieve the expected outcome with intended users, especially when gender-bias still exists in our society and even the most fairness-conscious individuals may still hold unconscious bias.

5 ONLINE USER STUDY

The goal of the user study was to investigate *(a)* whether users prefer a gender-debiased career recommender over a gender-aware recommender, and *(b)* whether their own biases play a role in their preferences.

We adopted a between-subject design. We randomly assigned participants to use either a gender-debiased or a gender-aware career recommender (except for those who declare themselves gender non-binary or decline to disclose their gender, in which cases the gender-debiased recommender was always used). For participants assigned to the gender-aware recommender, we further assigned them to interact with either a "female" model or a "male" model, based on whether they were identified with the female or male gender. Here, the "female (or male)" model means a gender-aware career recommender trained on female (or male) data only.

We invited students from all majors at a mid-size university in the Mid-Atlantic region of the U.S. to participate in the online user study. Prior to the study, the survey protocol received the IRB approval, and participants confirmed that they were 18 years or older and agreed to an informed consent before they could proceed. After taking the survey, each participant was entered in a raffle for multiple $50 Amazon Gift Cards. Overall, we received responses from 202 participants. The entire online survey took 5-10 minutes to complete.

In the following, we describe the questionnaire used in the study. The questionnaire included six sections: demographics, user interests, personal beliefs about gender-bias in career choice, career-specific gender disparity, recommendation acceptance and general usability.

5.1 Demographics

We collected minimal personal information (such as gender) needed for the study.

(1) What gender do you identify as (female, male, non-binary, or do not want to disclose)?
(2) What is your class standing (freshman, sophomore, junior, senior, or graduate students)?
(3) How "set" is your choice of major / concentration (still open to suggestions, or already determined, unlikely to Change)?

As shown in Figure 2, we have roughly the same number of male and female participants (48% females and 49.5% males). Unlike the Facebook dataset used to train the backend algorithm, our online user study includes 2% participants who were gender non-binary and 0.5% who did not want to disclose their genders. In terms of academic standing, 14.4% were freshmen, 16.8% were sophomore, 23.3% were juniors, 19.8% were seniors and 25.7% were graduate students. In addition, only 18.3% were open to career suggestions. The rest were set on their chosen majors/concentrations.

5.2 User Interests and Preferences

Users have varied interests and preferences. Accurately capturing their interests and preferences is critical to build a good recommender. In our work, we used a Facebook dataset to model user interests and preferences. The dataset stored 16K Facebook users' *likes* of 140K items. Given the large variety of the items, the set of items *liked* by a user could be a good representation of his or her interests and preferences.

Unfortunately, our participants were not among the 16K users in the Facebook dataset used to train our recommender system. In fact, many of them did not even use Facebook. It was certainly impossible to ask our participants to indicate whether they like each of the 140K items. In order to model our participants' interests and preferences, we first used a dimension reduction technique to group the 140K items into a small number of categories/topics. We then selected representative items from each category, and finally we asked our participants how they liked those representative items. The proposed user preference elicitation method was motivated by two observations: (1) among the large number of items on Facebook, users are more likely to rate popular items due to awareness, (2) items on the same topic (e.g., "The Lord of the Rings" and "The Hobbit") are highly correlated, so we only need to rate one of them (e.g., only rate "The Lord of the Rings").

Specifically, the Facebook dataset we used can be considered as a (sparse) user-item matrix with 16, 619 rows (users) and 143, 303 columns (items). An entry is 1 if a person liked the item and 0 otherwise. We considered each item as a "word," and for each person, all the items he or she liked form a "document." We then performed

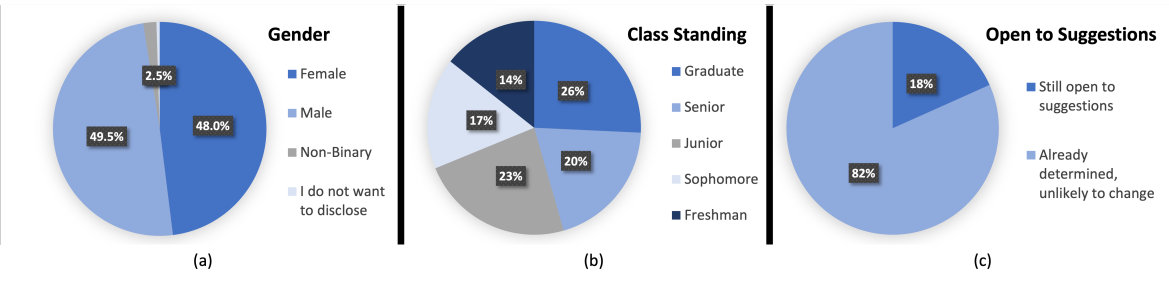

Figure 2: Participant Demographics: *(a)* Gender, *(b)* Academic Standings and *(c)* Open to Suggestion.

a 100-topic Latent Dirichlet Allocation (LDA) [9] analysis to automatically identify 100 latent topics in all the "documents." Each of the latent topic was represented by a bag of "words," or in our case, a set of items (e.g., a latent topic related to high fantasy novels may contain representative items such as *The Lord of the Rings*, *The Hobbit*, *J.R.R. Tolkien*, *The Well at the World's End*, and *The Chronicles of Prydain*).

From each of the 100 topics, we asked three volunteers to individually select one representative item from the top 10 items identified by LDA. We then picked the common items selected by the three volunteers to ensure most participants of our study are familiar with the items. We finally decided on 48 well-known items, which are not too many to elicit a participant's interests in them during the user study.

5.3 Personal Belief in Career Choice

We asked two questions to help us understand a participant's beliefs about gender roles in career selection.

(4) **(Q-Stereotype)** Please indicate whether you agree with the following statement or not: "A gender stereotype in career selection is undesirable since it limits women's and men's capacity to develop their personal abilities."
(5) **(Q-DisparityPersonal)** Please indicate whether you agree with the following statement: "If I am a female, I do not want to choose a career that is male-dominated" *(for female participants)*, or "If I am a male, I do not want to choose a career that is female-dominated" *(for male participants)*.

Both questions were rated on a 5-point Likert scale from *strongly disagree* to *strongly agree*.

5.4 Career-specific Perceived Gender Disparity

For each of the top-3 career recommendations, we asked participants whether they perceive it to be a female- or male-dominated career or I do not know **(Q-DisparityCareer)**. The main difference between **(Q-DisparityCareer)** and **(Q-DisparityPersonal)** is that **(Q-DisparityPersonal)** expresses a general personal belief, while **(Q-DisparityCareer)** is a perception specific to a career.

5.5 Recommendation Acceptance

For each of the top-3 career recommendations, we asked a participant whether he/she will consider it as a possible future career choice (yes, no, I don't know) **(Q-Acceptance)**.

5.6 General Usability

We also asked about a participants' agreement with two general system usability-related statements: *(a)* **(Q-UseAgain)** "I would like to use a career recommendation system like this in the future," and *(b)* **(Q-RecommendToOthers)** "I would like to recommend the system to my friends if it is available." Both are rated on a 5-point Likert scale.

6 RESULT ANALYSIS

We summarize the main findings of the user study, with a focus on *(a)* user acceptance of gender-debiased versus gender-aware career recommendations, and *(b)* whether a user's own belief/bias plays a role in the acceptance of a recommended career.

6.1 Summary of Career Recommendations by Each System

The AI recommender gave each participant a recommendation of three college majors (606 recommendations in total for 202 participants). Figure 3 summarizes the recommendations by the gender-aware and the gender-debiased systems. It shows the top 13 most recommended college majors by both systems and the probabilities (on the x-axis) they are recommended for male and female participants respectively. As the left side of the chart shows, among all the academic concentrations recommended by the gender-aware system, Psychology had 10% chance of being recommended to females and 7.7% chance of being recommended to males. The most frequently recommended careers by the gender-aware system for males were Psychology, Mechanical Engineering, History, Computer Science, and Criminal Justice; while the most frequently recommended careers for females were Psychology, English, Nursing, Biology and Accounting. Moreover, Computer Science and Mechanical Engineering were exclusively recommended to males by the gender-aware system. In contrast, the recommendations made by the gender-debiased system showed less gender stereotypes. As shown in the right chart of Figure 3, Computer Science was recommended to both males and females with similar probability (4.8% versus 4.1%). Based on this analysis, it seems the gender-debiased system is capable of mitigating some existing gender biases in career recommendation.

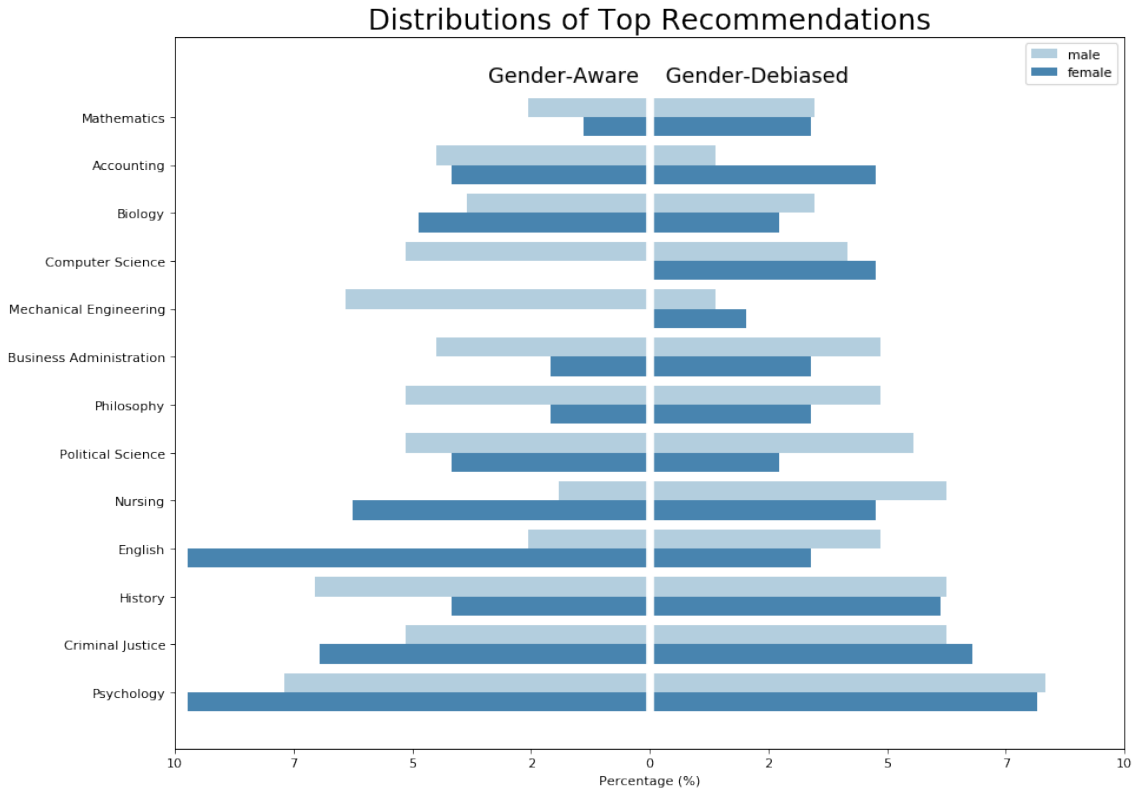

Figure 3: Distributions of Career Recommendations by Gender from the *Gender-Aware* and *Gender-Debiased* Systems.

6.2 Do People Prefer a Gender-Debiased Recommender?

To test this, for each of the top 3 recommended college majors, if a user indicated that they would consider it as a possible future career choice, the system received 1 point. The system received 0 points if the user said "no" and 0.5 if the user said "I don't know." Based on an independent sample t-test, the mean acceptance score for the gender-debiased system was 0.279 while that for the gender-aware system was 0.372. The difference is statistically significant with $p < 0.01$. Despite the results from the offline evaluation that showed the gender-debiased recommender was more fair while maintaining the same level of recommendation accuracy, users in general did not seem to prefer the recommendations made by the gender-debiased system more than those by the gender-aware system. In fact, the acceptance score for the gender-aware system was significantly higher than that for the gender-debiased system.

In the following, we try to explore whether a participant's own belief/bias plays a role in explaining the finding.

6.3 Self-reported Belief and Recommendation Acceptance

Here we focus on a participant's responses to Q-Stereotype and Q-DisparityPersonal. Both responses were rated on a 5 point Likert scale, 5 being the most biased (strongly disagree with Q-Stereotype or strongly agree with Q-DisparityPersonal), 1 being the least (strongly agree with Q-Stereotype and strongly disagree with Q-DisparityPersonal) and 3 being neutral. Figure 4 shows the distribution of the responses. The majority of the participants received a score of either "1" or "2." Very few people scored more than 3. In fact, only 4% of the participants scored 4 and 1.5% scored 5 in Q-Stereotype. Only 2% scored 4 and 0% scored 5 in Q-DisparityPersonal. In summary, based on self-reported user responses to Q-Stereotype and Q-DisparityPersonal, only a small number of the participants exhibited some degree of gender bias in career selection.

To test whether a participant's self-reported belief impacts his/her acceptance of a recommendation, we employed a Generalized Linear Model (GLM) where the dependent variable was his/her acceptance score regarding a recommended career and the independent variables were his/her responses to Q-Stereotype or Q-DisparityPersonal. We also controlled the variation of demographics such as age, gender and academic standings as they could be confounders.

The GLM results indicate that the main effect for Q-Stereotype on Q-Acceptance is not statistically significant($p < 0.667$). In contrast, the main effect for Q-DisparityPersonal on Q-Acceptance is significant ($p < 0.050$).

Note the self-reported bias measures such as Q-Stereotype and Q-DisparityPersonal may not accurately capture a person's true

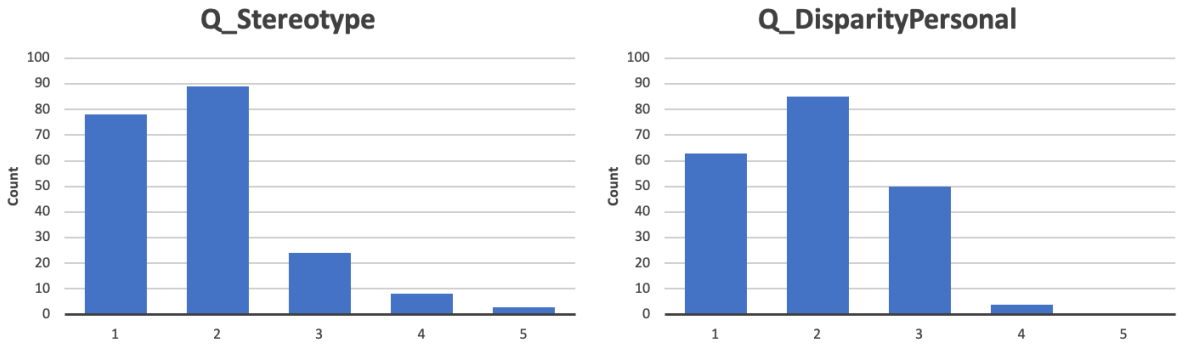

Figure 4: Self-reported belief about the Gender Role in Career Choices. Q-Stereotypes (1: strongly agree or least biased and 5 strongly disagree or most biased) and Q-DisparityPersonal (1: Strongly disagree or least biased and 5 Strongly Agree or most biased)

belief/bias. Prior research has demonstrated that social desirability bias is common in self-report surveys when the survey topics are sensitive (e.g., related to illegal acts such as drug use, income, ability and prejudice) [34, 47]. Due to social desirability concerns, there is a tendency for people to over-report socially desirable behaviors or attitudes and under-report socially undesirable behaviors or attitudes. Since gender bias is considered a sensitive topic, it is likely that our participants may have responded in a way that show less bias.

Mitigating social desirability in self-report surveys remains a challenging topic in social psychology as people differ in their tendency to engage in socially desirable responding [25, 26]. To overcome this problem, in the following, we propose a new measure to assess implicit human bias based on perceived gender-disparity of a college major. Since perceived gender-disparity of a college major (e.g., to ask a person whether Computer Science is a male or female-dominated career) is a less personal and less sensitive topic, it may not suffer from the same degree of social desirability bias as in Q-Stereotype and Q-DisparityPersonal.

6.4 Perceived Gender Conformity and Recommendation Acceptance

We define the perceived gender conformity (PGC) associated with a career, an implicit bias measure based on (a) a participant's own gender and (b) his/her responses to Q-DisparityCareer. PGC is equal to "1" or "conform" if the perceived dominant gender of a career is consistent with the gender of the participant (e.g., a career is perceived to be male-dominated and the participant is a male or a career is perceived to be female-dominated and the participant is a female). PGC will be "0" or "conflict" if the perceived dominant gender of a recommended career conflicts with the gender of the participant (e.g., the career is male-dominated and the participant is a female or the career is female-dominated and the participant is a male). For all the other cases (where participants answered "I don't know" to Q-DisparityCareer or the gender of the participant is non-binary or non-disclose), the value of PGC is assigned to "0.5" or "neutral."

We built a GLM model to study the relation between a user's PGC and his/her acceptance of a recommended career. The dependent variable was Q-Acceptance and the independent variable was PGC. We controlled demographic factors such as age, gender and academic standing. Our analysis results show a positive correlation between PGC and user acceptance ($p < 0.050$). Figure 5*(a)* shows the average acceptance scores grouped by PGC. The mean acceptance score was the lowest (0.23) when the perceived "gender" of the recommended career differed from the gender of the participant (PGC=0). In contrast, when they were the same (PGC=1), the mean acceptance score was the highest (0.364). When there was no perceived gender-disparity or the participant was gender non-binary or non-disclose (PGC=0.5), the mean acceptance score was in between (0.342). Since the acceptance gap between *PGC=0* and *PGC=0.5* was much larger than between *PGC=0.5* and *PGC=1*, the observed correlation between PGC and recommendation acceptance seems mainly due to the avoidance of careers that were perceived to be dominated by the opposite gender. This may partially explain why our participants preferred the gender-debiased system less since it tried to overcome some of the gender stereotypes and was more likely to recommend careers dominated by the opposite gender.

6.5 Interaction Between Personal Belief and PGC

We also studied whether there was any significant interaction effect between a user's personal belief (Q-Stereotypes and Q-DisparityPersonal) and the perceived gender conformity of a career (PGC) on user acceptance. We performed two new GLM analyses where the dependent variable was Q-Acceptance and the independent variables were Q-Stereotype*PGC or Q-DisparityPersonal*PGC respectively. We also controlled for demographics such as age, gender, and academic standing. Our results show that the interaction effect between a user's response to Q-Stereotype and PGC on Q-Acceptance was not significant ($p < 0.9223$). But there was a marginally significant interaction effect between a user's responses to Q-DisparityPersonal and PGC on user acceptance ($p < 0.052$).

To understand the interaction effect between Q-DisparityPersonal and PGC on user acceptance, we grouped

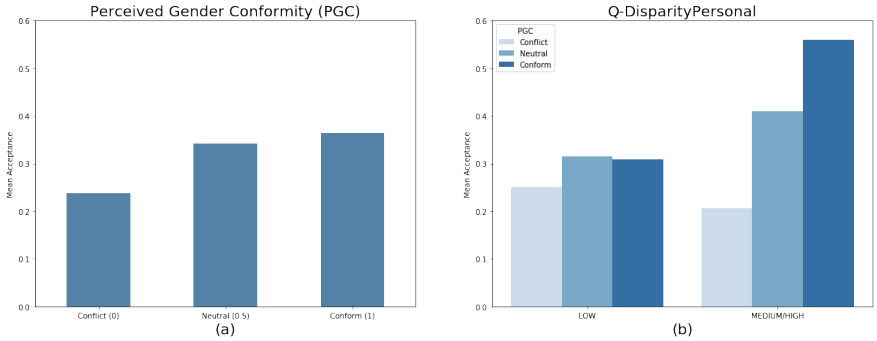

Figure 5: *(a)* The Relationship between User Acceptance (y-axis) and Perceived Gender Conformity (PGC) *(b)* The interaction between Q-DisparityPersonal and PGC.

the responses to Q-DisparityPersonal into LOW (those with a score of 1 or 2) and MEDIUM/HIGH (those with a score of 3, 4, and 5). Since there were very few people with a score of 4 or 5 (only 2% of the participants), most of the people in the Medium/High group had a score of 3. Figure 5(b) shows the comparison between these two groups of people. When people did not mind selecting a career dominated by the opposite gender (in the Q-DisparityPersonal=LOW group), there wasn't much difference in their acceptance of careers conflicting or conforming to their genders. In contrast, people in the MEDIUM/HIGH group showed high preference to careers that conform to their genders (PGC=1) while avoiding careers that conflicting with their genders (PGC=0).

6.6 User Acceptance and Other Factors

Among the demographics of a participant, gender was found to be significantly correlated with recommendation acceptance ($P < 0.017$); the correlation between age and recommendation acceptance was marginally significant ($p < 0.063$). Academic standing (e.g., a freshman or a senior) was not significantly correlated with the acceptance ($p < 0.911$).

6.7 General System Usability

Finally, based on user responses to two general system usability questions (Q-UseAgain and Q-RecommendToOthers), our participants were generally positive about the systems. The mean score for Q-UseAgain was 3.34 and the mean for Q-RecommendToOthers was 3.40, both are better than neutral (3). Figure 6 shows the response distributions by different systems (5 being the best). Since the response distributions for Q-UseAgain and Q-RecommendToOthers are very similar, they show a consistency between these two usability measures. In addition, since the darker bars skewed more toward right, the mean usability scores for the gender-aware system are generally higher than those of the gender-debiased system (for Q-UseAgain, the mean is 3.20 for the gender-debiased system and 3.47 for the gender-aware system; for Q-RecommendToOthers the mean is 3.27 for the gender-debiased system and 3.52 for the gender-aware system). The differences are marginally significant (for Q-UseAgain, $p < 0.095$, for Q-RecommendToOthers: $p < 0.093$). This result is consistent with the Q-Acceptance-based evaluation measure, which confirms one of our main findings: the study participants preferred the gender-aware system more than the gender de-biased system.

7 DISCUSSION

In this section, we discuss the findings we have discovered in this work, its implications on fair AI system design, and the limitations of our current study which could be addressed in future work.

7.1 What Have We Discovered?

Much effort in the AI and HCI community has focused on identifying and removing AI bias using algorithmic or design-based solutions, including the debiased AI career recommender we developed. What we have discovered is that our participants *prefer biased recommendations over debiased ones*, despite the fair recommender achieved better fairness and better prediction accuracy ($p < 0.01$). For debiasing algorithms such as our fair career recommender to move the needle on equity in the real world, we may also need to find ways to "nudge" users to accept debiased recommendations (e.g., accepting careers dominated by the opposite gender).

To understand why participants did not prefer fair recommendations, *we found that their conscious or unconscious bias may play a role.* Our participants seemed to avoid careers that are dominated by the opposite gender (e.g. significant main effects for Q-DisparityPersonal and PGC on Q-Acceptance ($p < 0.05$) and a marginally significant interaction effect between a user's responses to Q-DisparityPersonal and PGC on Q-Acceptance ($p < 0.052$)). Similar results were found in previous research [73] where, even as kindergartners, girls would select mostly traditionally female careers and avoid traditionally male careers. In other words, the bias is so deeply ingrained that people consciously or subconsciously shun a career dominated by the opposite gender regardless whether it matches their interests, personality and skills. Societal bias may also play a role in the participants' preferences for gendered recommendation. Even supposing that a person is unbiased, they may still make a career choice that conforms to "social norms" if they believe that they might otherwise be disadvantaged in their career growth or subjected to discrimination on the job.

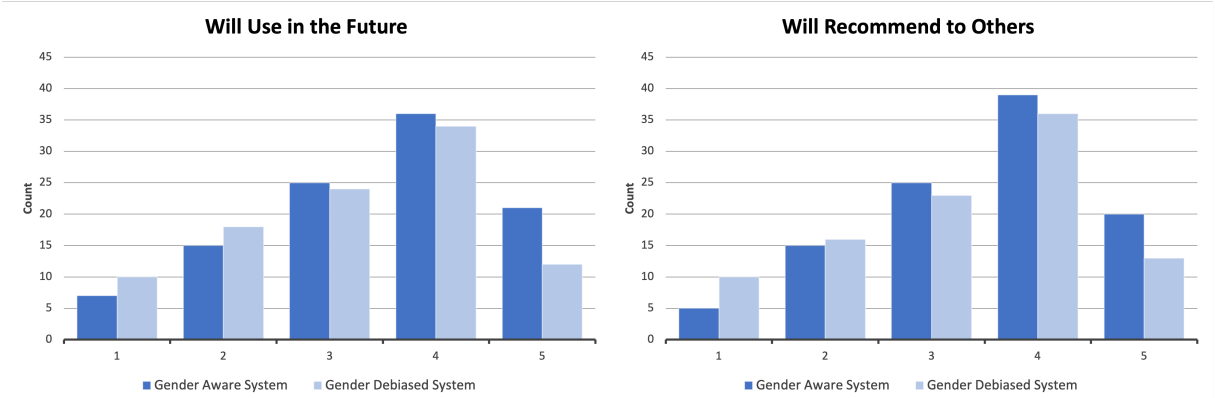

Figure 6: Distributions of *(a) Q-UseAgain (left)* and *(b) Q-RecommendToOthers (right).*

Since systemic bias and prejudice due to humans are root causes of inequities in our society [18] and hence in data, without addressing the human side of the issue, fair AI systems may not produce their intended societal impacts. Beyond addressing algorithmic issues, the AI and the HCI communities need to devote more attention to *developing technologies that can help humans identify and overcome their own biases.*

7.2 Implications on Fair AI System Design

Our study results have demonstrated the importance of addressing human bias in effective human- fair-AI interaction, which is an understudied area in fair AI research. Identifying, assessing, and removing human bias is clearly a challenging problem. Extensive social science research has been conducted to address these issues. It however has not gained much attention in fair AI/ML research. We recommend to focus on three new areas of fair AI/ML research that have the potential to improve the effectiveness of human- fair-AI interaction.

AI-based Human Bias Assessment. We could potentially use AI to help humans to detect and assess their own biases and raise their awareness. Social psychology research has developed bias measurement instruments to allow users to assess their own biases, especially unconscious bias [31, 32, 62]. But existing bias measures have some limitations. For instance, due to variance from external factors, implicit bias measures such as IAT [31, 33] are not stable at the individual level [52, 64]. We believe that a promising new research direction in fair AI system design could be the integration and enhancement of human bias assessment instruments via advanced AI modeling to help accurately assess human biases and raise bias awareness.

New AI-based Persuasion Technologies for Behavior Nudge. We believe new persuasion technologies aiming at "nudging" people to open more to new career choices may help users overcome their biases. One promising direction is to provide them with more accurate and actionable information about themselves as well as the recommended careers. For instance, a fair career recommender needs to convince a girl that despite its male dominance, a computer science career can still be woman-friendly and highly rewarding. Furthermore, if the system is able to help the girl to understand her personality traits and strength as well as why they make her a good candidate for a computer science career, then it may have a better chance to convince the girl to consider computer science as a future career.

AI Bias/Fairness Explanation. We believe AI bias/fairness explanation may play an important role in improving users' acceptance of a fair AI system. Since AI bias/fairness is an abstract concept unfamiliar to many people, we hypothesize that users may be more motivated to accept the recommendations of a debiased AI system if they understand how fairness is defined and how bias is removed from an AI system. Although explainable AI (XAI) is a hot research area aiming to help users to understand the decisions made by an AI system, so far, there has not been much XAI research on explaining the bias and fairness of an AI system to users.

Iterative Human-AI Bias Co-training. We believe humans and AI should work together to help each other to overcome their biases. On the one hand, an AI system that is capable of quantifying the biases in human behaviors/decisions, explaining why the biases are harmful can be used to raise human awareness and encourage behavior changes. On the other hand, since AI bias frequently originates from human bias and prejudice, which impacts AI training data, with less bias in the data from humans, there will be correspondingly less bias in AI. Humans can also periodically audit an AI system to ensure its fairness. Since some human biases are already deeply ingrained in a person's subconscious, it could be difficult to eliminate human biases. We envision an interactive and iterative human-AI bias co-training process where AI and humans work together iteratively and continuously to help correcting the biases in each other.

7.3 Limitations and future directions

One limitation of the current study is that over 80% of the participants are not open to new career suggestions, which significantly limits the statistical power of our analysis. As most of the participants are university students and have decided an academic concentration, it would be better if most participants have not decided a college major, and thus could benefit more from the recommendations. A future possibility is to revise the current IRB protocol to recruit minors (e.g., high school students) for the study.

Gender bias is only one type of biases that may harmfully affect one's career choices. For instance, Title VII of the Civil Rights Act of 1964 prohibits employment discrimination based not only on sex/gender, but also race, color, religion, and national origin. In the future, we also need to address different types of biases (e.g., racial bias) to really make an AI career recommender trustworthy for all users.

Career-related biases exist not only in applicants who choose college majors or jobs but also in people who make college admission and job hiring decisions. Both an applicant and an admission officer/employer's decision will have significant impact on the inclusiveness and equality of the workforce. Thus, it is important that we study and address all the potential biases in a career pipeline. Unfortunately, it is not feasible to address all the issues related to a career pipeline in one study. We may consider the biases in college admission or job hiring a possible future research direction.

8 CONCLUSIONS

In this paper, we demonstrated that it is not sufficient to simply perform algorithm debiasing to achieve the desired societal outcomes that the field of fair AI aims to produce, at least in the context of mitigating gender bias in college major recommendations. In fact, we found that on average our participants did not prefer recommendations from a gender-debiased system, even though the system was fairer and more accurate than the gendered system. Our results suggest that participants' own biases are contributing factors to their acceptance of the AI recommendations (e.g., participants tended to avoid careers dominated by the opposite gender). To improve real-world equity in careers via a debiased AI system, it may be necessary to counter human bias as well as AI bias. We have discussed several promising research directions for fair AI system design which may help to address this issue, including integrating and enhancing human bias assessment instruments, providing fairness explanation, and implementing new "nudge" techniques to help people open to new career choices. Going forward, it would be valuable to repeat our experiments in other problem domains, and with other protected dimensions such as race and disability status. To ensure that the impacts of fair AI technologies fulfill their potential benefits to society, more research on human- fair-AI interaction, an understudied area, is urgently needed.

ACKNOWLEDGMENTS

This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology. This material is based upon work supported by the National Science Foundation under Grant No.'s IIS2046381; IIS1850023; IIS1927486. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* 23 (2016).

[4] Albert Bandura. 1977. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review* 84, 2 (1977), 191.

[5] Albert Bandura. 1978. Reflections on self-efficacy. *Advances in behaviour research and therapy* 1, 4 (1978), 237–269.

[6] Natã M Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[7] S. Barocas and A.D. Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[8] 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 T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* 63, 4/5 (2019), 4:1–4:15.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[10] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, and A.T. Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in NeurIPS*.

[11] T. Buonocore. 2019. Man is to doctor as woman is to nurse: the gender bias of word embeddings. https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17.

[12] Toon Calders and Indre Zliobaite. 2013. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*. Studies in Applied Philosophy, Epistemology and Rational Ethics, Vol. 3. Springer, International, 43–57. https://doi.org/10.1007/978-3-642-30487-3_3

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[14] A. Campolo, M. Sanfilippo, M. Whittaker, A. Selbst K. Crawford, and S. Barocas. 2017. *AI Now 2017 Symposium Report.* AI Now.

[15] Jiawei Chen, Anbang Xu, Zhe Liu, Yufan Guo, Xiaotong Liu, Yingbei Tong, Rama Akkiraju, and John M Carroll. 2020. A General Methodology to Quantify Biases in Natural Language Data. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[16] Shelley J Correll. 2001. Gender and the career choice process: The role of biased self-assessments. *American journal of Sociology* 106, 6 (2001), 1691–1730.

[17] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[18] K. Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.* (1989), 139–167.

[19] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2020. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[20] J. Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[21] Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word Vectors. arXiv:1901.07656 [cs.CL]

[22] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[23] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285.

[24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. Fairness through awareness. In *Proceedings of ITCS*. ACM, 214–226.

[25] Allen L Edwards. 1953. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology* 37, 2 (1953), 90.

[26] Wilbert E Fordyce. 1956. Social desirability in the MMPI. *Journal of Consulting Psychology* 20, 3 (1956), 171.

[27] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 424–432.

[28] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.

[29] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

[30] Peter Glick and Susan T Fiske. 1999. Gender, power dynamics, and social interaction. *Revisioning gender* 5 (1999), 365–398.

[31] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. 1998. Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74 (1998), 1464–1480.

[32] A. G. Greenwald, B. A. Nosek, and M. R. Banaji. 2003. Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology* 85 (2003), 197–216.

[33] A. G. Greenwald, T. A. Poehlman, E. L. Uhlmann, and M. R. Banaji. 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97 (2009), 17–41.

[34] Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing* (2010).

[35] M. Hardt, E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in NeurIPS*. 3315–3323.

[36] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1661–1670.

[37] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. *CoRR* abs/1708.05031 (2017). arXiv:1708.05031 http://arxiv.org/abs/1708.05031

[38] Elspeth JR Hill and James A Giles. 2014. Career decisions and gender: the illusion of choice? *Perspectives on medical education* 3, 3 (2014), 151–154.

[39] Youyang Hou, Cliff Lampe, Maximilian Bulinski, and James J Prescott. 2017. Factors in Fairness and Emotion in Online Case Resolution Systems. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2511–2522.

[40] Nyi Nyi Htun, Elisa Lecluse, and Katrien Verbert. 2021. Perception of Fairness in Group Music Recommender Systems. In *26th International Conference on Intelligent User Interfaces*. 302–306.

[41] Janet S Hyde, Elizabeth Fennema, and Susan J Lamon. 1990. Gender differences in mathematics performance: A meta-analysis. *Psychological bulletin* 107, 2 (1990), 139.

[42] Rashidul Islam, Kamrun Naher Keya, Shimei Pan, and James Foulds. 2019. Mitigating demographic biases in social media-based recommender systems. *KDD (Social Impact Track)* (2019).

[43] Rashidul Islam, Shimei Pan, and James R Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 586–596.

[44] Isaac Johnson, Connor McMahon, Johannes Schöning, and Brent Hecht. 2017. The Effect of Population and" Structural" Biases on Social Media-based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*. 1167–1178.

[45] 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*. 3819–3828.

[46] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70, 6 (2015), 543.

[47] Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* 47, 4 (2013), 2025–2047.

[48] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. 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.). 4066–4076.

[49] Campbell Leaper and Christine R Starr. 2019. Helping and hindering undergraduate women's STEM motivation: experiences With STEM encouragement, STEM-related gender bias, and sexual harassment. *Psychology of Women Quarterly* 43, 2 (2019), 165–183.

[50] Robert W Lent, Steven D Brown, and Gail Hackett. 1994. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of vocational behavior* 45, 1 (1994), 79–122.

[51] Penelope Lockwood. 2006. "Someone like me can be successful": Do college students need same-gender role models? *Psychology of Women Quarterly* 30, 1 (2006), 36–46.

[52] Edouard Machery. 2017. Do indirect measures of biases measure traits or situations? *Psychological Inquiry* 28, 4 (2017), 288–291.

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

[54] Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences* 114, 48 (2017), 12714–12719.

[55] Nora McDonald and Shimei Pan. 2020. Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–19.

[56] Ronald McQuaid and Sue Bond. 2004. Gender stereotyping in career choice. (2004).

[57] Judith L Meece, Jacquelynne E Parsons, Caroline M Kaczala, and Susan B Goff. 1982. Sex differences in math achievement: Toward a model of academic choice. *Psychological Bulletin* 91, 2 (1982), 324.

[58] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[59] Danaë Metaxa-Kakavouli, Kelly Wang, James A Landay, and Jeff Hancock. 2018. Gender-inclusive design: Sense of belonging and bias in web interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[60] Alex P. Miller and Kartik Hosanagar. 2010. How Targeted Ads and Dynamic Pricing Can Perpetuate Bias. *Harvard Business Review* (2010).

[61] S.U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[62] B. A. Nosek and M. R. Banaji. 2001. The go/no-go association task. *Social Cognition* 19, 6 (2001), 161–176.

[63] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[64] Frederick L Oswald, Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E Tetlock. 2013. Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of personality and social psychology* 105, 2 (2013), 171.

[65] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 6620–6631.

[66] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J Jansen. 2020. Analyzing Demographic Bias in Artificially Generated Facial Pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[67] Joni Salminen, Soon-Gyo Jung, and Bernard J Jansen. 2019. Detecting Demographic Bias in Automatically Generated Personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[68] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[69] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8, 9 (2013), e73791.

[70] Zoe Skinner, Stacey Brown, and Greg Walsh. 2020. Children of Color's Perceptions of Fairness in AI: An Exploration of Equitable and Inclusive Co-Design. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[71] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*. 125–132.

[72] Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[73] Susan Kochenberger Stroeher. 1994. Sixteen kindergartners' gender-related views of careers. *The Elementary School Journal* 95, 1 (1994), 95–103.

[74] Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv:1901.10002 [cs.LG]

[75] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[76] Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[77] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[78] Michael White and Gwendolen White. 2006. Implicit and Explicit Occupational Gender Stereotypes. *Sex Roles* 55 (08 2006), 259–266.

[79] 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*. 1–14.

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

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

[82] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 2925–2934.

[83] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040.

[84] 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*. 1171–1180.

[85] 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*. 335–340.