# When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation

CLARICE WANG, University of Pennsylvania, USA
KATHRYN WANG, Carnegie Mellon University, USA
ANDREW Y. BIAN, University of Maryland, College Park, USA
RASHIDUL ISLAM, KAMRUN NAHER KEYA, JAMES FOULDS, and SHIMEI PAN,
University of Maryland, Baltimore County, USA

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 humans and 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 career recommendations without sacrificing its accuracy in prediction. 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 perceived gender disparity is a determining factor for the acceptance of a recommendation. In other words, we cannot fully address the gender bias issue in AI recommendations without addressing the gender bias in humans. We conducted a follow-up survey to gain additional insights into the effectiveness of various design options that can help participants to overcome their own biases. Our results suggest that making fair AI explainable is crucial for increasing its adoption in the real world.

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

Additional Key Words and Phrases: AI, machine learning, fairness, gender bias, career recommendation

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.

Authors' addresses: C. Wang, University of Pennsylvania, 3820 Locust Walk, Philadelphia, PA 19104-6134, USA; email: clarice7@seas.upenn.edu; K. Wang, Carnegie Mellon University, Pittsburgh,112-01 Queens Blvd. Apt. 23G, Forest Hills, NY 11375, USA; email: kathryn1wang@gmail.com; A. Y. Bian, University of Maryland, College Park, 4603 Calvert Rd, College Park, MD 20740, USA; email: abian18@terpmail.umd.edu; R. Islam, K. N. Keya, J. Foulds, and S. Pan, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA; emails: {islam.rashidul, kkeya1, jfoulds, shimei}@umbc.edu.


2160-6455/2023/09-ART17 $15.00
https://doi.org/10.1145/3611313

**ACM Reference format:**
Clarice Wang, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2023. When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation. *ACM Trans. Interact. Intell. Syst.* 13, 3, Article 17 (September 2023), 28 pages.
https://doi.org/10.1145/3611313

## 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 [Angwin et al. 2016; Dastin 2018; Noble 2018]. 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 [Angwin et al. 2016]. In other domains including personalized search, ads and recommendation, AI systems lead to skewed outcomes [Ali et al. 2019]. For example, on Facebook, bodybuilding ads are much more likely to be delivered to men, while cosmetics ads are much more likely to be delivered to women [Ali et al. 2019]. AI models trained on text data have been found to encode gender stereotypes such as associating computer programming with men and homemaking with women [Bolukbasi et al. 2016], 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 [Dastin 2018].

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 [Campolo et al. 2017]. We argue that there are two aspects which must be considered when it comes to mitigating AI bias: the algorithmic aspect and the human-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 define and quantify fairness [Dwork et al. 2012; Foulds et al. 2020a; Kusner et al. 2017], and *(b)* new machine learning techniques that remove bias [Dev and Phillips 2019; Foulds et al. 2020b; Zhang et al. 2018]. On the other hand, the human- fair-AI interaction aspect as well as its broader social context is 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 [Miller and Hosanagar 2010], 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 [Barocas and Selbst 2016], will bias mitigation be effective if we focus only on removing the bias in AI algorithms without addressing the bias in humans?

In this research, we systematically study the interplay between AI debiasing techniques and humans, using AI college major recommendation as a case study. College major selection is one of the most important career choices one has to make that can impact one's economic success, social standing, and quality of life. Making a college major decision is difficult. Currently students rely on school counselors for advice. However, according to the latest data from **National Association for College Admission Counseling (NACAC)** and the **American School Counselor Association (ASCA)** [ASCA 2021], the typical high school counselor in the US is responsible for 415 students in the 2020-2021 school year. This workload leaves them little time to address students' needs, let alone provide personalized advice on college major decisions to large numbers of students. An AI-based recommender system has the potential to mine a large amount of student data to gain deep insight into a student's interests

and personality to provide more accurate and personalized college major recommendations [Stein et al. 2020; Yadalam et al. 2020].

In choosing a college major as well as other career decisions, 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 [White and White 2006]. Machines that learn from existing career data are thus expected to be influenced by the resulting gender gaps in career choices unless bias mitigation techniques are performed [Dastin 2018].

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 debiased fair AI system where gender stereotypes are systematically removed over a biased one. Users may have their own biases, and our results show that users' acceptance of a debiased AI system can be influenced by their own biases. While this result may not be entirely unexpected, it has largely been neglected by the AI research community, and to the best of our knowledge, we are the first to show it via rigorous experimental study. We then further investigate a few design options that can be used to "nudge" users into more often accepting a fair AI recommender system.

The following are the main contributions of the paper.

- To the best of our knowledge, this is the first study on how human bias interferes with the effectiveness of a fully implemented fair/debiased AI recommender and how to better design such a system to achieve its intended societal impact.
- As a case study, we develop a debiased college major recommender which mitigates gender bias in recommendations [Islam et al. 2019a]. Our offline evaluation shows that the debiased recommender is fairer than the gender-aware (biased) recommender without any loss of prediction accuracy, an ideal "fairness for free" situation [Islam et al. 2021].
- 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 an equally accurate gender-debiased 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 in a recommended college major is significantly correlated with its acceptance.
- We perform a follow-up survey to gain more insights into the effectiveness of various design options for a fair college major recommender which may encourage users to accept gender-debiased recommendations more often.

In the rest of the paper, we describe related literature, the implementation of a debiased **machine learning (ML)** 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 recommender system, and a follow-up survey that provides additional insights into the design of such a 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 **human-computer interaction (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, AIF360, the IBM open source platform for fair machine learning, has 77 metrics for fairness/bias [Bellamy et al. 2019]. Among them, some are about *individual fairness* and others are about *group fairness*.

Individual fairness seeks to ensure similar individuals get similar outcomes ("An algorithm is fair if it gives similar predictions to similar individuals") [Dwork et al. 2012]. 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") [Dwork et al. 2012], *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") [Hardt et al. 2016], *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") [Hardt et al. 2016] 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") [Foulds et al. 2020a].

Other widely used fairness definitions include *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") [Gajane and Pechenizkiy 2017], 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") [Kusner et al. 2017]. Moreover, most fairness measures are defined for classification tasks. There are fairness measures for other tasks. For example, *Non-parity Unfairness* [Yao and Huang 2017] is designed to evaluate the fairness of recommender systems.

In terms of mitigating bias in AI systems (a.k.a. debiasing AI), although the occurrence of bias in data and algorithms arises from myriad socio-technical factors and is difficult to disentangle and eliminate these impacts, there has been much progress on AI debiasing methods in the fair ML community and these techniques can be effective at mitigating bias and creating more equitable solutions. One bias mitigation method is to remove the bias from the data that is used to train the AI models. For example, vector projection-based bias attenuation method is used to remove bias from word embeddings [Dev and Phillips 2019]. Since word embeddings are widely used as features to train downstream natural language processing models, debiased word embedding improves the fairness of the downstream applications. There is also a large body of work that optimizes ML models under both traditional accuracy-based and new fairness-based objectives [Agarwal et al. 2018; Foulds et al. 2020b; Woodworth et al. 2017; Zafar et al. 2017]. Furthermore, adversarial learning is 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) [Zhang et al. 2018].

### 2.2 Social Science Research on Gender Bias and Careers

According to Glick and Fiske [1999], 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 [Correll 2001]. 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 [Stroeher 1994]. Scottish pupils were found to perceive Truck Driver, Engineer, Plumber/Electrician, Laborer, Armed Forces as "male" jobs while Nurse and Care Assistant as "female" jobs. Boys, but especially girls, have strong preferences against working in sectors and industries that are traditionally the domain of the opposite sex [McQuaid and Bond 2004].

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* [Bandura 1977] and *Social Cognitive Career Theory* [Lent et al. 1994] are the most influential social cognitive frameworks for understanding individual human behavior as well as career decisions. They posit that human behaviour 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 [Bandura 1978; Lent et al. 1994].

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" [Hyde et al. 1990; Meece et al. 1982]. 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) [Correll 2001].

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 [Hill and Giles 2014; Lockwood 2006; McQuaid and Bond 2004]. Encouragement from friends and family [Leaper and Starr 2019] also improves self-efficacy.

### 2.3 HCI and AI Fairness/Bias

Recent work on AI fairness/bias in the HCI community mostly focuses on identifying and analyzing biases in AI systems such as image search [Kay et al. 2015; Otterbacher et al. 2017], social media analysis [Johnson et al. 2017], image and persona generation [Salminen et al. 2020, 2019], sentiment analysis [Díaz et al. 2018], text mining [Cryan et al. 2020] and natural language generation [Strengers et al. 2020].

The HCI community also worked on fairness perception and definition. Wang et al. [2020] 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. Hou et al. [2017] and Woodruff et al. [2018] explored how intended users, especially those marginalized by race or class, feel about algorithmic fairness. Dodge et al. [2019] conducted an empirical study on how people judge the fairness of ML systems and how explanations impact that judgment. Chen et al. [2020] 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 [Madaio et al. 2020; Metaxa-Kakavouli et al. 2018; Skinner et al. 2020; Vorvoreanu et al. 2019]. Furthermore, better tooling [Cramer et al. 2019; Yan et al. 2020] and better algorithms [Barbosa and Chen 2019; Strengers et al. 2020] also helps address the problem.

So far, there is little 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 COLLEGE MAJOR RECOMMENDATIONS

We use an AI college major recommendation system as a case study to illustrate how an AI 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 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 [Sarwar et al. 2001]. 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 to students who share similar interests.

A major source of bias/unfairness in machine learning outcomes arises from biases in the data [Barocas and Selbst 2016; Mehrabi et al. 2019; Suresh and Guttag 2019]. 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 system may inherit the bias into its recommendations, for example, by suggesting, as career choices, 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. Figure 1 shows the system architecture. The input to the system is a user's interests indicated on Facebook. We chose the Facebook data because it contains a wide range of items a person can like such as books, hobbies, thoughts, brands, movies, music, celebrities and sports. The output contains the top academic majors/concentrations recommended by our system such as computer science, biology and psychology. Since the academic majors/concentrations in our data are declared by Facebook users, they are quite noisy (e.g., "Defense Against the Dark Arts" is a declared academic concentration). To prepare the Facebook data to train our recommender, we filtered out academic majors/concentrations 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 academic majors/concentrations, 140K+ unique Facebook items that a user can like and 3.5 million+ user- like item pairs.

To develop a college major recommender, first, we train a **neural collaborative filtering (NCF)** [He et al. 2017] model for predicting the Facebook 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

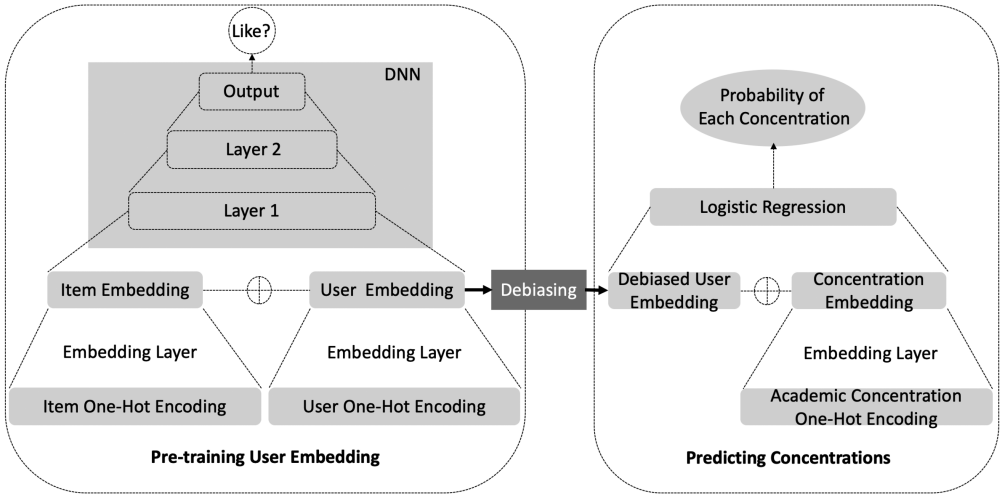


Fig. 1. The architecture of a gender-debiased career recommender.

also included 10% negative instances from those user-interest pairs marked as "0."[1] In the input layer, the users and items are one-hot encoded.[2] 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 combined by concatenation. 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 **mean squared error (MSE)** loss[3] 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 a de-biasing step prior to applying logistic regression. Our debiasing approach adapts the work on attenuating bias in word vectors [Dev and Phillips 2019]. 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 [Buonocore 2019]:

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

The user embeddings we have trained experience a very similar issue. 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

[1]While observed interactions (e.g., thumbs up) reflect users' interests, unobserved interactions are either missing data (a user didn't rate an item) or negative examples (dislike an item). Given this, treating all unobserved interactions as negative examples is not ideal. Following [He et al. 2017, 2016], we randomly sampled unobserved interactions as negative examples. This approach also reduces training time.

[2]One-hot encoding is a common way of representing categorical variables in ML as vectors of size N, where N is the number of unique values for each variable (e.g., number of users or items). For each vector, only one element can be "1" and the rest are "0."

[3]Here we optimize the MSE loss because we treat the predicted outcome: "0" and "1" as numerical ratings, as what is typically done in collaborative filtering-based recommender systems [Koren et al. 2009].

gender bias in our system. We then debias $p_u$ by removing $p_u$'s projection on the 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, 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:

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

In the career 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 Facebook items of the user using the pre-trained NCF model. Then the embedding of the new user is used as input features in the pre-trained logistic regression classifier directly. For the gender-debiased system, we dropped the intercept terms during prediction to further remove popularity bias [Steck 2011]. Finally, the logistic regression model recommends top-$N$ academic majors/concentrations computed by sorting the probabilities of the 1,380 majors/concentrations for a given a 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 only differences between these two variations are:

(1) The **gender-aware system** makes recommendations based on the choices by the people of the same gender (e.g., recommending to girls based on the college major choices of other girls) while the **gender-debiased system** makes recommendations based on the choices by the people of both genders.
(2) The gender-debiased system employs the linear projection-based gender de-biasing strategy to systematically remove gender stereotypes from user embeddings, while the gender-aware system bypasses the debiasing step.

### 4.1 Evaluation Metrics

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

***Normalized discounted cumulative gain at position K (NDCG@K).*** This is a well-known metric for assessing the quality of a *ranked* list of results (e.g., recommendations) [He et al. 2015].

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

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.

Table 1. Offline Evaluation Results on the *Facebook* Dataset

| | NDCG@3 ↑ | NDCG@10 ↑ | NDCG@20 ↑ | $U_{PAR}$ ↓ |
|---|---|---|---|---|
| GenderAware | 0.0016 | 0.0041 | 0.0065 | 1.4524 |
| GenderDebiased | **0.0024** | **0.0057** | **0.0089** | **1.1888** |

Higher values are better for NDCG; lower values are better for $U_{par}$.

*Non-parity unfairness ($U_{PAR}$).* This metric is designed to evaluate the fairness of recommender systems [Yao and Huang 2017]. 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 $N$ career 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 career concentrations, we need negative examples as well, that is, a job or career concentration that is not a good fit for a user. We generate random pairs $(u, c)$ as negative training instances, where $c$ is any career concentration not explicitly declared by $u$. Furthermore, we conducted cross-validation in offline evaluation experiments. Instead of $k$-fold cross-validation, we performed a repeated random sub-sampling validation [Kuhn et al. 2013] (also known as Monte Carlo cross-validation or the repeated hold-out method). In k-fold cross-validation, the likelihood of a single or extreme outlier skewing the results is increased. For example, if a model is significantly biased (in terms of statistical bias) by one extreme outlier, 9 out of 10 partitions in a 10-fold cross-validation will be impacted. The advantage of the Monte Carlo cross-validation approach over $k$-fold cross-validation is that the proportion of the training/test split is not dependent on the number of partitions. We created 25 random splits of the dataset where each split contains 70% of the data as a train set and 30% as a test set. For each such split, the model was fit to the training data, and predictive accuracy was assessed using the test data. The results are then averaged over the 25 random splits.

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

We also visualize the **principal component analysis (PCA)** projections of the male and female bias direction vectors ($v_{male}$ and $v_{female}$, respectively) before and after the linear projection-based debiasing embeddings method (BeforeDebiasing and GenderDebiased, respectively). *PCA* was performed based on the embeddings for all of the data. Figure 2 shows that the male and female vectors have very different directions and magnitudes for the BeforeDebiasing system. In contrast, the male and female vectors have a more similar direction and magnitude to each other for the GenderDebiased system.

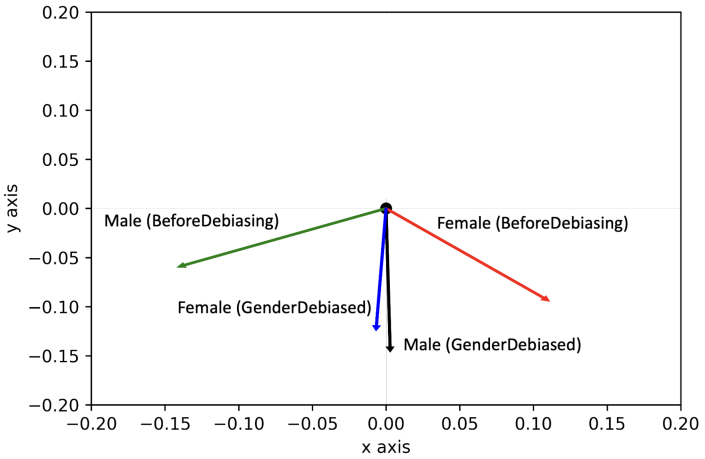


Fig. 2. Visualization of the bias direction vectors for male ($v_{male}$) and female ($v_{female}$) by PCA projection. PCA was performed based on the embeddings for all data points.

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, an ideal "fairness for free" situation [Islam et al. 2019b]. This is also a criteria frequently used in the fair machine learning community to indicate the success of a bias mitigation algorithm. In the following, we describe a user study to investigate whether the gender-debiased career recommender can achieve the desired 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 college major recommender over a gender-aware recommender, and *(b)* whether their own beliefs/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 case 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" 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 (both STEM majors such as Mathematics, Chemistry, and Computer Science as well as non-STEM majors such as History, English, Visual Art, and Music) at **University of Maryland, Baltimore County (UMBC)**, a mid-size public 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 **Institutional Review Board (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 $50 Amazon Gift Cards. Overall, we received responses from 202 participants. The entire survey took 5-10 minutes to complete for each participant.

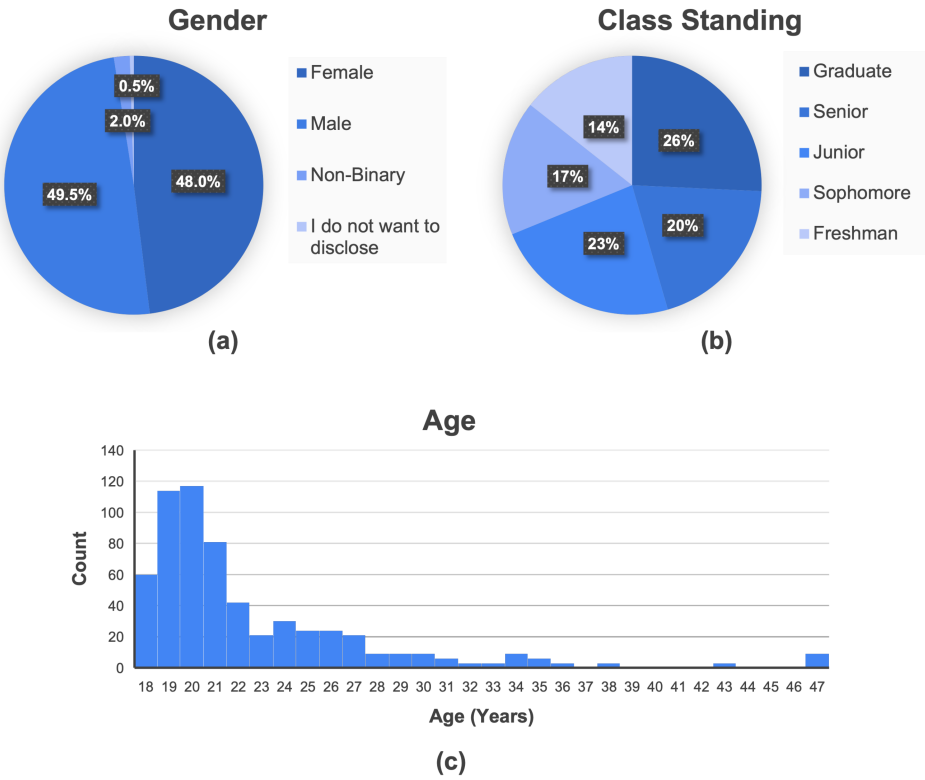


Fig. 3. Participant demographics: (a) gender, (b) academic standings, (c) age.

In the following, we describe the questionnaire used in the study. The questionnaire included five sections: demographics, user interests, beliefs about gender-bias in career choice, 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 student)?
(3) What is your age?

As we randomly assigned participants to use either the gender-aware or the gender-debiased system, the demographics of the participants in each experiment condition (gender-aware versus gender-debiased) are quite similar. Here, instead of group-wise statistics, we summarize the demographics of all the participants to reduce redundancy.

As shown in Figure 3, 48% of our participants are females and 49.5% males. Unlike the Facebook dataset used to train the backend algorithm, our online user study also 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 terms of age, the range was between 18 and 47 with most participants between 18 and 22.

### 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 16$K$ Facebook users' *likes* of 140$K$ online items (e.g., books, brands, movies, TV shows, celebrities). 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 online study participants were not among the 16$K$ users in the Facebook dataset. 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 140$K$ 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/topic, and finally we asked our participants how they liked those representative items.

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 a 100-topic **Latent Dirichlet Allocation (LDA)** [Blei et al. 2003] analysis to automatically identify 100 latent topics in all the documents. Each of the latent topics 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. We finally decided on 48 well-known items, which are not too many to elicit a participant's interests in them during the online user study.

The strategy we used in eliciting user interests in our online user study is based on the following assumptions: (1) each user only rates a small fraction of the 140$k$ items on Facebook. Among the large number of items on Facebook, users are more likely to rate popular items due to awareness; (2) items in the same LDA topic (e.g., "The Lord of the Rings" and "The Hobbit") are highly correlated, so we only need to rate one of the most representative items in each topic (e.g., only rate " The Lord of the Rings").

### 5.3 Personal Belief on Gender Roles in Career Choice

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

(1) **(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."
(2) **(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 Recommendation Acceptance

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

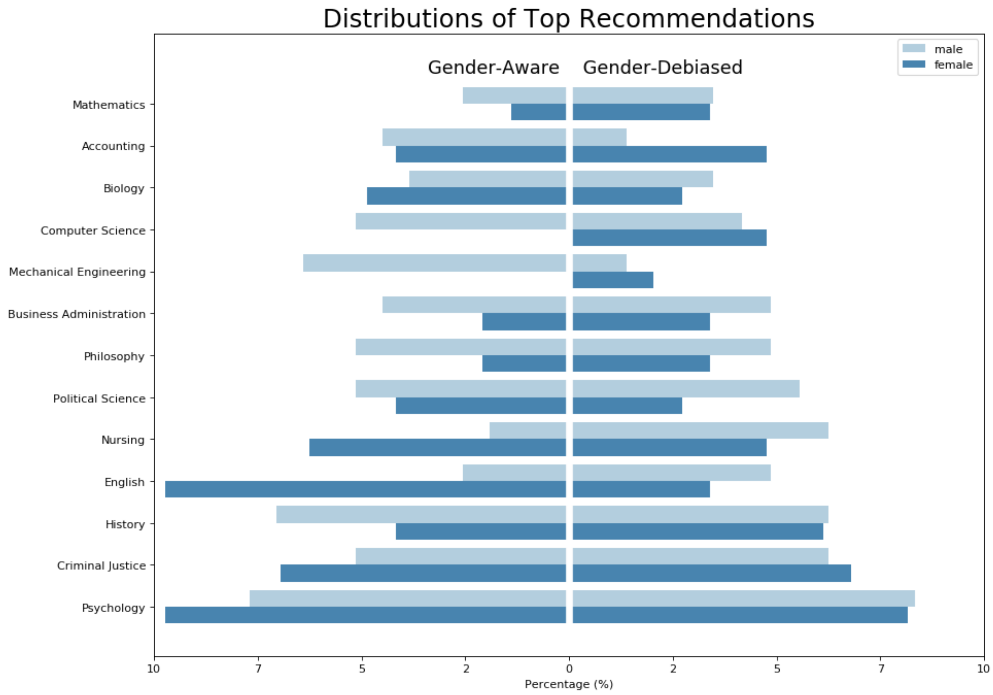


Fig. 4. Distributions of career recommendations by gender from the *Gender-Aware* and *Gender-Debiased* systems.

**(Q-DisparityCareer)** whether they perceive it to be a female- or male-dominated career or I do not know. The main difference between **(Q-DisparityCareer)** and **(Q-DisparityPersonal)** is that **(Q-DisparityPersonal)** expresses a general personal opinion, while **(Q-DisparityCareer)** is a participant's perception specific to a career.

### 5.5 General Usability

We also asked about a participants' agreement with two general usability-related statements:

(1) **(Q-UseAgain).** "I would like to use a career recommendation system like this in the future,"
(2) **(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 from strongly disagree to strongly agree.

## 6 RESULT ANALYSIS

We summarize the main findings of the online user study, with a focus on *(a)* user acceptance of gender-debiased versus gender-aware recommendations, and *(b)* whether a user's own belief or 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 academic majors/concentrations (606 recommendations in total for 202 participants). Figure 4 summarizes the recommendations by the gender-aware and the gender-debiased systems. It shows the top 13 most

recommended academic majors/concentrations 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 majors/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 fewer gender stereotypes. As shown in the right chart of Figure 4, Computer Science was recommended to both males and females with similar probability (4.8% versus 4.1%). This was also the case for Mechanical Engineering. Based on this analysis, it seems the gender-debiased system is capable of mitigating some existing gender biases in its recommendation.

In addition to the examples from the top recommendations by each system, we also computed $U_{PAR}$ to assess the overall fairness of all the recommendations made by each system: 3.5833 for the gender-aware system and 2.2083 for the gender-debiased system. Since the $U_{PAR}$ value for the gender-debiased system is lower than that for the gender-aware system, this confirms that in the online user study, the recommendations made by the gender-debiased system were fairer (or less biased) than those made by the gender-aware system.

### 6.2 Do People Prefer a Gender-Debiased Recommender?

To test this, for each of the top three recommended concentrations, 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 findings.

### 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 5 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. As we randomly assigned the participants to either the gender-aware or the gender-debiased condition, the distribution of Q-Stereotype and Q-DisparityPersonal is quite similar in these two conditions. 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.

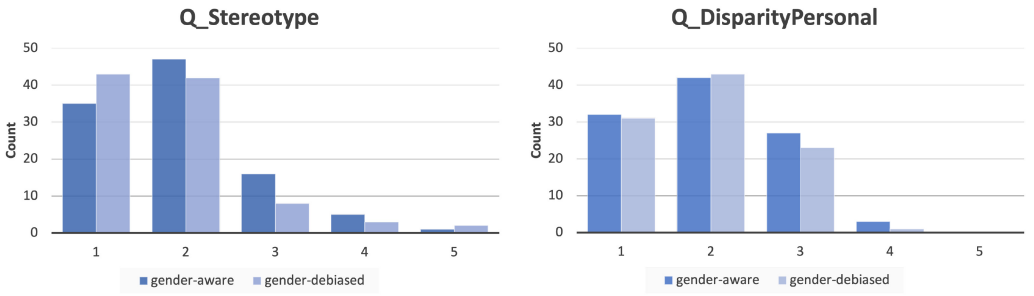


Fig. 5. Self-reported belief about the gender role in career choices. Q-Stereotypes (1: strongly agree or least biased and 5 strongly disagree or most biased) and Q-DisparityPersonal (1: strongly disagree or least biased and 5 strongly agree or most biased).

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 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) [Krumpal 2013]. 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 shows 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 [Edwards 1953; Fordyce 1956]. To address this problem, in the following, we measure a form of implicit bias based on perceived gender-disparity of an academic major/concentration. Since perceived gender-disparity of a career (e.g., to ask a person whether Computer Science is a male or female-dominated career) is a less personal and less sensitive question, we expect to get a more objective assessment of the participants' attitudes toward gender bias in career selection.

### 6.4 Perceived Gender Conformity and Acceptance

We define **perceived gender conformity (PGC)** based on a participant's own gender and 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 the 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."

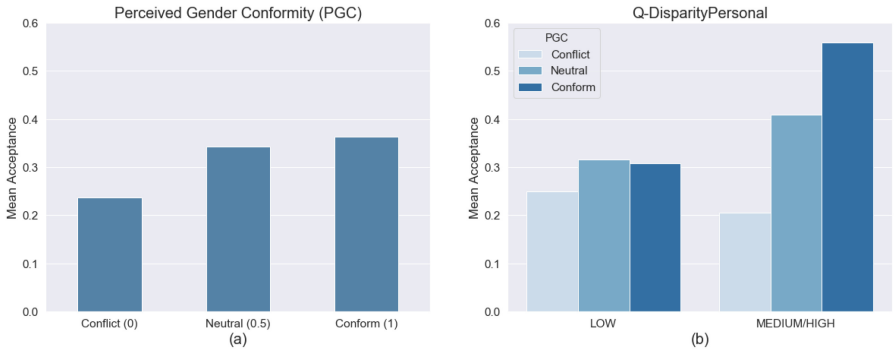


Fig. 6. (a) The relationship between user acceptance (y-axis) and perceived gender conformity (PGC). (b) The interaction between Q-DisparityPersonal and PGC.

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 6(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 for a career or the participant was gender non-binary or non-disclosed (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 that between *PGC=0.5* and *PGC=1*, the observed correlation 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 the system intentionally tried to overcome some of the gender stereotypes and was more likely to recommend careers dominated by the opposite gender to study participants.

We also grouped the participants based on experiment conditions (gender-aware group versus gender-debiased group). We repeated the same analysis within each group. Our results showed a positive correlation between user acceptance and PGC in both groups. However, due to a reduced sample size when performing group-wise analysis (e.g., the size of each group is roughly half of the size of the combined data), the correlation was marginally significant for the gender-debiased group ($p < 0.060$) and not significant for the gender-aware group ($p < 0.349$). Since gender was taken into consideration when making recommendations by the gender-aware system, the number of recommended majors in the "conflict" category was lower (15%) than that in the gender-debiased group (25%). This may partially explain the reduced aversion effect in the gender-aware group.

### 6.5 Interaction Between Personal Belief and Perceived Gender Conformity

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 disparity 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

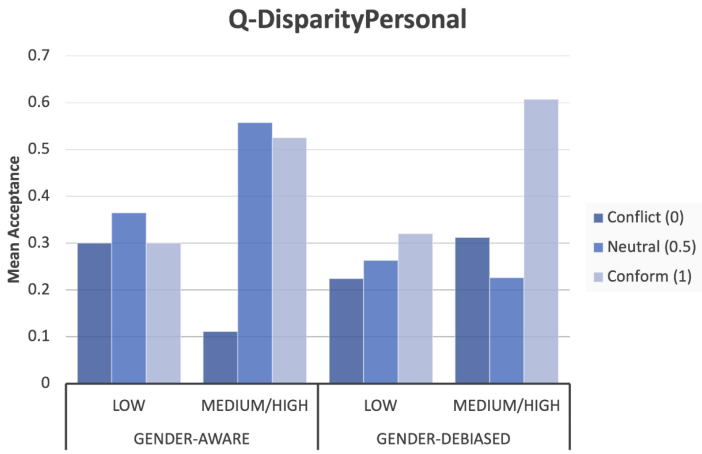


Fig. 7. The interaction effect of Q-DisparityPersonal and perceived gender conformity (PGC) on user acceptance (y-axis) for each experiment condition.

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 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 6(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 with 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).

We also repeated the same analysis for each experiment condition. For the gender-debiased group, the interaction effect was not significant for both Q-Stereotypes (p<0.922) and Q-DisparityPersonal (p<0.680). For the gender-aware group, the interaction effect was not significant for Q-Stereotypes (p<0.926). For Q-DisparityPersonal however, the interaction effect was significant with p<0.018. Figure 7 shows the group-wise interaction effect between PGC and Q-DispartityPersonal on user acceptance. The chart on the left shows the significant interaction effect between PGC and Q-DisparityPersonal on user acceptance.

### 6.6 Interaction between Gender and Perceived Gender Conformity

We are also interested in understanding whether the impact of PGC on recommendation acceptance varies across different genders (e.g., whether females are more likely to avoid a career dominated by a different gender than males?).

To study this, we performed a new GLM analysis where the dependent variable was Q-Acceptance and the independent variables were gender, PGC, and gender*PGC. We also controlled for other demographic factors such as age and academic standing. Our results show that the interaction effect of gender and PGC on Q-Acceptance was indeed significant ($p < 0.005$).

Figure 8 shows the acceptance rates of different genders and different PGC categories. As only 2.5% of the participants belong to the gender "non-binary" or "I do not want to disclose" group and

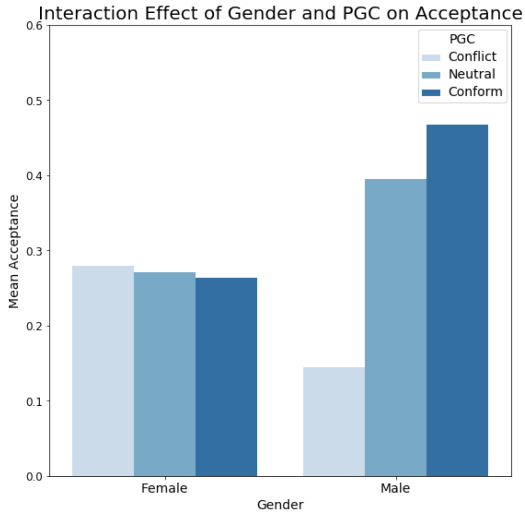


Fig. 8. The interaction effect of gender and perceived gender conformity (PGC) on user acceptance (y-axis).

all of them had a PGC value of "neutral," here we only focus on the differences between females and males. As shown in Figure 8, there was not much difference for females to accept college major recommendations conflicting with or conforming to their own gender. Male participants however showed a significant aversion to careers dominated by a different gender (e.g., the acceptance rate of the "conflict" group was 0.145% versus 0.467% for the "conform" group).

This result is interesting as the observed aversion to careers dominated by a different gender was mainly because of males. Females in our study did not exhibit the same kind of aversion. We hypothesize that careers perceived to be associated with males (e.g., engineers and physicians) frequently are highly paid with higher social status than those associated with females (e.g., teachers and nurses). These incentives may motivate females to consider male-dominated careers more, and consequently partially mitigate females' aversion to male-dominated careers.

### 6.7 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 junior) was not significantly correlated with the acceptance ($p < 0.911$).

### 6.8 General System Usability

Finally, based on user responses to two general 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 9 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

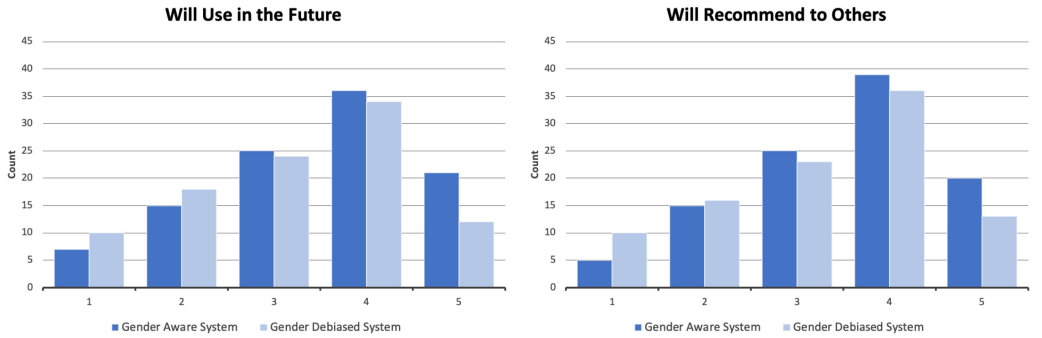


Fig. 9. Distributions of system usability scores.

significant (for Q-UseAgain, $p < 0.095$; for Q-RecommendToOthers, $p < 0.093$). This result is also consistent with the Q-Acceptance-based evaluation measure, which further confirms our main finding: the study participants preferred the gender-aware system more than the gender de-biased system.

## 7 A FOLLOW UP SURVEY ON DESIGNING A MORE EFFECTIVE FAIR AI RECOMMENDER

Although the gender-debiased recommender was fairer without any loss of prediction accuracy, our participants preferred such a system less than a gender-aware recommender. Since the potential societal benefits of fair AI systems can only be realized if users accept gender-debiased recommendations, an important question we need to answer next is how fair AI systems can be designed to help users overcome their own biases. To gain additional insight on this, we designed a follow-up survey with our study participants to get some early feedback, especially on the role "fairness/bias explanation" plays in a fair AI recommender system.

### 7.1 Participants

We invited all the participants in our first user study to the follow-up survey, and received the responses from 20 participants. We invited the same group of participants because they had prior experience interacting with the AI career recommender, which is the context of the follow-up survey. Figure 10 shows the demographics of the study participants. Among them, 8 were females (40%), 10 were males (50%), and 2 were gender non-binary (10%). In terms of academic standing, 15% were freshmen, 20% were sophomore, 20% were juniors, 5% were seniors, and 40% were graduate students. In terms of age, the range was between 19 and 44 with most participants between 18 and 24. The average age was 24.8. They were from diverse majors such as history, public health, and computer science.

Since AI bias/fairness is an abstract concept unfamiliar to many lay people, we hypothesize that users may be more motivated to accept the recommendations of a debiased/fair AI system if they understand AI bias/fairness and the goals a fair AI system tries to better achieve.

### 7.2 Questionnaire

Our first set of questions were designed to investigate the effectiveness of different **fairness/bias explanation strategies (FESs)** that can be used to help users understand AI fairness as well as its potential societal impact. The FES statements are in the form of: "It would help me better understand the bias/fairness issues in the system if the AI recommender [$FES_i$]" where $FES_i$ is one of the following:

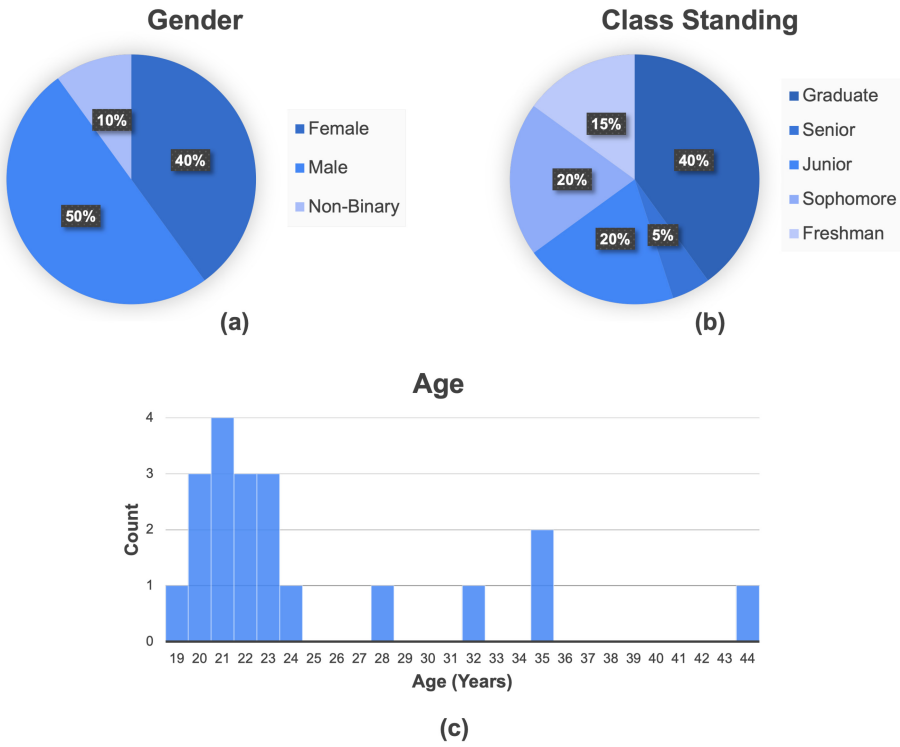


Fig. 10. Participant demographics in the follow-up survey.

- (FES1) explains why a traditional recommender is unfair (e.g., the traditional recommender is more likely to recommend computer science to males and nursing to females).
- (FES2) explains how bias/fairness are defined in the system (e.g., an AI system is considered fair with regard to gender if other things being equal, males and females are equal likely to be recommended a particular career, e.g., computer science).
- (FES3) provides a "process-oriented" explanation (e.g., The system ensures that people sharing similar interests are recommended similar careers regardless of their genders).
- (FES4) explains why bias exists in data and how it impacts the behavior of the system (e.g., if in the data most of the computer scientists are males, the system will learn to recommend computer science to males more often than to females).
- (FES5) provides a side-by-side comparison of the recommendations from a gender de-biased system and a gender-aware recommender.

The second set of questions was designed to help us understand what are the effective **design choices (DCs)** a fair AI system can implement to encourage users to accept gender-debiased recommendations more. Since the DCs were designed to "nudge" people to overcome their aversion to careers dominated by a different gender, we also explored those not directly related to AI fairness.

The DC statements are in the form of "I would be more likely to accept the recommendations by the system if [$DC_i$]" where $DC_i$ is one of the following:

- (DC1): the system employs FES1-FES5 to explain the fairness/bias issues to me.
- (DC2): the system informs me of the positive societal impact of debiased recommender systems (e.g., research shows that increasing the number of women in STEM can equate to an additional $12 trillion in the global GDP by the year 2025).

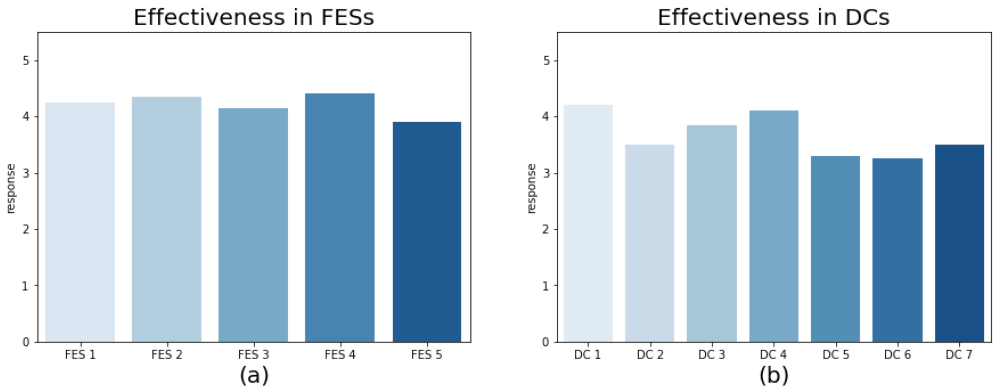


Fig. 11. Summary of survey results.

- (DC3): the system informs me of the positive personal impact of a career. (e.g., pursuing a computer science career may bring individual financial success)
- (DC4): the system suggests the field as a minor or interdisciplinary subject in addition to my current major, instead of asking me whether I would change my current career/major.
- (DC5): the system provides success stories and anecdotes of people of my gender in that field.
- (DC6): the system provides the actual gender disparity and it isn't as bad as I had presumed.
- (DC7): if the system provides a trend that shows more people of my gender are entering this field than before.

All the FES and DC statements were rated on a 5-point Likert scale, "1" being strongly disagree and "5" being strongly agree. For each of their responses, we also asked the participants to justify their selection (which was optional). At the end, we also included open-ended questions to allow users to share additional thoughts/feedback with us.

### 7.3 Result Analysis

Figure 11(a) shows that our participants were positive about all the FESs we investigated. The average scores were mostly above 4. Among all the FESs, FES4 (explain the bias in data) and FES2 (explain the definition of fairness) had the highest mean scores (4.4 and 4.35, respectively) while FES5 (side-by-side comparison) and FES3 (explain process) had the lowest scores (3.9 and 4.15, respectively). However, only the following differences are statistically significant based on a paired sample $t$-test: (a) FES4 is significantly better than FES5 ($p < 0.033$), and (b) FES2 is significantly better than FES5 ($p < 0.035$).

Based on this preliminary analysis, our participants seems to recognize the value of a fairness explanation. Based on their feedback, many believe this can help increase their trust to the system and its recommendation. For instance, User 1 (U1) "If it shows that it is unbiased and explains how, I would trust it more." Or (U2) "If the processes behind the AI recommender are more transparent, easier to understand, then they'll also be easier to trust and more genuinely helpful if the bias is removed." Except for FES5, the mean scores for the other FESs were above 4.0. Other than FES5, there were no statistically significant differences among FES1-FES4).

In terms the effectiveness of various design choices (DCs) in encouraging users to accept de-biased recommendations, DC1 (explaining fairness) received the highest scores (mean 4.20), while DC6 (actual disparity) was least effective (3.25).

Table 2. DC Differences: Statistical Significance Analysis based on Paired Samples $t$-Test

| | DC2 | DC3 | DC4 | DC5 | DC6 | DC7 |
|---|---|---|---|---|---|---|
| DC1 (4.20) | **0.002** | 0.092 | 0.358 | **0.004** | **0.010** | **0.037** |
| DC2 (3.50) | | **0.035** | **0.007** | 0.165 | 0.165 | 0.500 |
| DC3 (3.85) | | | **0.048** | **0.006** | **0.007** | 0.055 |
| DC4 (4.10) | | | | **0.001** | **0.001** | **0.005** |
| DC5 (3.30) | | | | | 0.402 | 0.179 |
| DC6 (3.25) | | | | | | 0.086 |

The numbers in parentheses in the first column are the means and the numbers in other columns are the $p$-values. Bold-faced numbers are significant.

We performed a paired sample $t$-test to assess the significance of their differences. As shown in Table 2, we observed that DC1, DC4 and DC3 form the "top group" while DC6, DC5, DC2 and DC7 form the "bottom group." The differences between the strategies in the top and the bottom groups are mostly significant while the differences within each group are mostly insignificant. These results seem to suggest:

- Our participants value fairness/bias explanation and consider this to be one of the most effective strategies to help them accept debiased college major recommendations.
- Although our participants may prefer not to change the current major based on the recommendation, this doesn't necessarily mean they see no value in the recommendations. In fact, they may still consider the recommended academic concentrations as minors.
- Positive personal impacts such as financial success seem promising as a useful incentive to nudge people to consider majors dominated by an opposite gender.

Some of the comments provided by our participants confirm the above findings, including: (U3) "If it shows that it is unbiased and explains how, I would trust it more" (supports DC1); (U4) "... I would rather add it on as a minor instead of changing my major because I have already put 3 years of work into my major and would not want to completely change majors..." (supports DC4); (U5) " ... with more incentive and display of benefits, I would want to be a part of it." (supports DC3); (U6) "When recommending career paths, I don't need to know who currently dominates the field. I only want to know what matches my interests. Adding these anecdotes introduces doubt to the individual and the inclusion of success stories reinforces the idea that the underrepresented group is an anomaly" (against DC6 and DC5).

In their responses to the open-ended question in the follow-up survey, the participants also pointed out some limitations in the current study. For instance, (U7) "Though the survey asked about gender. Is this the only source of bias? I believe mentioning other sources of bias would help in pushing the use of a fair/debiased AI system"; (U8) "There are a lot of other biases you'd need to correct outside of gender bias to really make a career AI trustworthy for users, though I'm guessing that's not part of the research project."

In addition to various biases, career decisions can be influenced by many other factors, which were not accounted for in our current study. To quote several of our participants, (U9) "Detailing the impacts of a chosen career based on gender would definitely impact my choice. Yet, as a man I generally don't take gender as a consideration in my job. Location, Pay, Potential for progression and my current credentials would come first and foremost"; (U10) "maybe include some additional information on professional groups/societies around that career, and more specifically, student subsections, where they can connect with other students interested in pursuing that career"; (U11)

"Addressing salary/wage gaps for women in male-dominated careers"; (U12) "if the system shows 'empathy' …"

Some participants also expressed doubt regarding an AI career recommender. For example (U13) "These additional pieces of information about a recommended career would be helpful, but more than likely, I would research the career on my own"; (U14) "I always take AI suggestions with a grain of salt because there is no way of knowing every details of me as a person, not because of gender bias."

## 8 DISCUSSION

In this section, we discuss the main 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.

### 8.1 What Have We Discovered?

We have discovered three main findings in this research:

**(1) Our study participants did not prefer a de-biased/fair AI recommender**. 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 college major recommender we developed. What we have discovered is that our *participants prefer biased recommendations over debiased ones*. The debiased/fair recommendations were accepted less often. The system was also rated lower based on two usability measures: Q-UseAgain and Q-RecommendToOthers. *For debiasing algorithms such as our fair 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.*

**(2) Aversion to college majors dominated by a different gender seems to be one of main reasons why the participants did not prefer gender-debiased recommendations**. To understand why participants did not prefer debiased recommendations, our study results suggest that their unconscious bias may play a role. Our participants, primarily male participants, seemed to avoid careers that are dominated by a different gender. Similar results were found in previous research [Stroeher 1994] where, even kindergartners would select mostly careers conforming to their own genders. In other words, the bias is so deeply ingrained that people subconsciously shun a career dominated by the opposite gender regardless whether it matches their interests 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.

Since systemic bias and prejudice due to humans are root causes of inequities in our society [Crenshaw 1989] 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.*

**(3) Our study participants considered fairness/bias explanation to be one of the most effective strategies that can potentially "nudge" users to accept the debiased/fair recommendations more often.** To get early feedback on designing a more effective fair AI recommender, we conducted a follow-up survey. Based on the survey results, our participants considered the FESs valuable. Most of the FESs received a mean score that is above 4.0. Among the 7 DCs in

the survey, fairness/bias explanation was ranked the highest in its potential to help increase user acceptance of debiased recommendations.

### 8.2 Implications on Fair AI System Design

Identifying, addressing, and removing human bias is clearly a challenging problem. Governments and corporations conduct bias training routinely to raise awareness. But this process can be costly and only large institutes can afford it. How can AI and HCI help? First, we could potentially use AI to help humans to detect 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 [Greenwald et al. 1998, 2003; Nosek and Banaji 2001]. We believe that a promising new research direction in fair AI system design could be the integration of novel automated bias measurement instruments to assess implicit human bias and to raise bias awareness.

Second, bias/fairness explanation and persuasion technology may play important roles in helping people overcome their biases. During our follow-up survey, fairness/bias explanation was considered one of the top strategies to encourage users to accept gender-debiased recommendations. We believe educating people by explaining fairness/bias as well as its societal impact may also raise awareness and slowly nudge people toward a fairer society.

Third, 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 to reduce/correct human biases. On the other hand, since AI bias frequently originates from human bias and prejudice which impacts 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.

Finally, it is well-known that AI systems frequently learn and inherit the human biases in their training data [Howard and Borenstein 2018], and so the biases captured in an AI system often mirror the biases in our society. For instance, as our college major recommender was trained using the career decisions made by the people in our society, the biases captured in the model may reflect the career-related gender biases in our society. As shown in the left side of Figure 4, before debiasing the system, the most male-dominated majors are mechanical engineering and computer science while the most female dominated careers are English and nursing. This implies that AI models trained with data on human decision making can potentially be used as tools to help us analyze, interrogate and quantify societal biases. As a potential future research direction, we would like to study the correlation between a user's perception of career-related gender bias and that encoded in the system. We also want to explore whether and why males and females may perceive them differently.

### 8.3 Limitations of the Current Study

One limitation of the current study is the small sample size in the follow-up survey. Although the survey results indicate that our participants consider fairness explanation to be one of the top strategies to "nudge" them to accept gender-debiased recommendation more, the data is insufficient to answer more specific questions such as which of the FESs are more effective. In addition, due to the small sample size, the representativeness of the sample population could be a concern (e.g., 10% participants in the follow-up survey are gender non-binary versus 2% in the first user study). We plan to extend the current effort to include more participants in the future.

Another limitation of the current study is that the current FESs in the follow-up survey were designed based on the key elements in a fair AI system including fairness definitions, fairness measures and bias mitigation processes, concepts unfamiliar to many lay people. Although we tried to use examples to illustrate the concepts, lay people may still find it difficult to differentiate them. As a part of our future research, we plan to explore effective ways to communicate and explain these concepts to people who are not AI experts.

In terms of its scope, our study focused solely on gender bias, which represents just one facet of bias that can influence career choices. It is important to note that Title VII of the Civil Rights Act of 1964 addresses employment discrimination not only based on sex/gender but also on other factors like race, color, religion, and national origin. Therefore, the study does not encompass the broader spectrum of biases that individuals may encounter.

Furthermore, we did not incorporate additional social and economic factors that can significantly impact an individual's career decisions, such as salary considerations, geographical location, and potential for career development. These factors can wield considerable influence on career choices.

It is essential to acknowledge that the current study's focus on gender bias is just one piece of the larger puzzle, and the omission of other biases and influential factors limits the comprehensive understanding of career decision-making processes.

## 9 CONCLUSIONS

In this paper, we demonstrated that it is not sufficient to simply perform AI/ML debiasing algorithms to achieve the desired societal outcomes that the field of fair AI/ML aims to produce, at least in the context of gender-debiased 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 just as accurate as the corresponding gender-aware system and participants generally self-reported that gender stereotypes in career selections are undesirable. Our results suggest that participants' own beliefs and 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 fairness system, it may be necessary to counter human bias as well as AI bias. Results from our follow-up survey suggested that providing additional fairness explanations may be helpful in addressing this issue. Going forward, it would be useful to validate the questionnaires and increase the number of participants in our follow-up survey. It is also 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 fulfil their potential benefits to society, more research on human- fair-AI interaction, an understudied area, is urgently needed.

## REFERENCES

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

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.

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).

ASCA. 2021. Student-to-School-Counselor Ratio 2020–2021. https://www.schoolcounselor.org/getmedia/238f136e-ec52-4bf2-94b6-f24c39447022/Ratios-20-21-Alpha.pdf

Albert Bandura. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* 84, 2 (1977), 191.

Albert Bandura. 1978. Reflections on self-efficacy. *Advances in Behaviour Research and Therapy* 1, 4 (1978), 237–269.

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.

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

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.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

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*.

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

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

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.

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.

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.

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.

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.

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

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

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.

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, 275–85.

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

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.

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

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

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020b. Bayesian modeling of intersectional fairness: The variance of bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 424–432.

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

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

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.

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.

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

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.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. 173–182.

Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 549–558.

Elspeth J. R. Hill and James A. Giles. 2014. Career decisions and gender: The illusion of choice? *Perspectives on Medical Education* 3, 3 (2014), 151–154.

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.

Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics* 24, 5 (2018), 1521–1536.

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

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

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

Rashidul Islam, Shimei Pan, and James R. Foulds. 2021. *Can We Obtain Fairness For Free?* Association for Computing Machinery, New York, NY, USA, 586–596.

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.

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.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

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

Max Kuhn and Kjell Johnson. 2013. *Applied Predictive Modeling*. Vol. 26. Springer.

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.

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.

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.

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.

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.

Ronald McQuaid and Sue Bond. 2004. Gender stereotyping in career choice. *European Commission. Employment Research Institute and Careers Scotland*.

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.

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).

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.

Alex P. Miller and Kartik Hosanagar. 2010. How targeted ads and dynamic pricing can perpetuate bias. *Harvard Business Review* (2010).

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

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

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.

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.

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.

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.

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.

Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems*. 125–132.

Samuel A. Stein, Gary M. Weiss, Yiwen Chen, and Daniel D. Leeds. 2020. A college major recommendation system. In *RecSys'20: Proceedings of the 14th ACM Conference on Recommender Systems*. 640–644.

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.

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

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

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.

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.

Michael White and Gwendolen White. 2006. Implicit and explicit occupational gender stereotypes. *Sex Roles* 55 (08 2006), 259–266.

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.

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

Tanya V. Yadalam, Vaishnavi M. Gowda, Vanditha Shiva Kumar, Disha Girish, and Namratha M. 2020. Career recommendation systems using content based filtering. In *Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)*.

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.

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.

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.

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.

Received 20 June 2022; revised 11 July 2023; accepted 17 July 2023