Mitigating Cognitive Biases in Machine Learning Algorithms for Decision Making

Christopher G. Harris
School of Mathematical Sciences, University of Northern Colorado, Greeley, Colorado 80639 USA
christopher.harris@unco.edu

ABSTRACT

Cognitive biases are an ingrained part of the human decision-making process. Nearly all machine learning algorithms that mimic human decision-making use human judgments as training data, which propagates these biases. In this paper, we conduct an empirical study in which 150 applicants are rated for suitability for three separate job openings. We develop an algorithm that learns from human judgments and consequently develops biases based on these human-generated inputs. Next, we explore and apply techniques to mitigate these algorithmic biases, using a combination of pre-processing, in-processing, and post-processing algorithms. The results from our study show that biases can be mitigated using these approaches but involve a tradeoff between complexity and effectiveness.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**;

KEYWORDS

Algorithmic Bias, Data Science, Fairness, Machine Learning, Decision Making, Artificial Intelligence, Cognitive Bias, Human Resources Tasks

ACM Reference Format:
Christopher G. Harris. 2019. Mitigating Cognitive Biases in Machine Learning Algorithms for Decision Making. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366424.3383562

1 INTRODUCTION

Over the past few decades, dozens of cognitive biases that affect human judgment and decision making have been identified by notable behavioral economists such as Kahneman and Tversky [1]. When humans make judgments or decisions, they frequently use heuristic strategies (i.e., decisional short cuts) These heuristics often lead to cognitive biases, which are systematic and predictable errors in judgment that result from over-reliance on these heuristics. For example, the anchoring effect demonstrates that humans tend to be heavily influenced by the first piece of information they hear, such as the list price for a car they wish to purchase; future negotiations deviate from that initial amount which serves as an anchor.

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383562

Cognitive biases become an unintentional input into these decisions. Researchers have discovered that many humans are often unaware of their own cognitive biases [2]; moreover, even when they are made aware of a specific cognitive bias, i.e., anchoring, they often cannot correct or eliminate them [3]. Cognitive biases in decision-making have material impacts on people's lives - examples of these include decisions in hiring, advertising, criminal justice, granting credit, personalized medicine, and targeted policymaking.

Machine learning algorithms were designed to make decisions not only faster but also more accurately and fairly; in other words, these algorithms are designed to eliminate or reduce cognitive biases. However, since human judgments typically serve as inputs to decision-making machine algorithms, these cognitive biases become integrated into the resulting algorithms, propagating the biases. Recently, with the advent of machine learning algorithms, increasing attention has focused on identifying, reducing and eliminating these biases and ensuring fairness.

We conduct an empirical study to evaluate strategies of bias mitigation in algorithms trained on human-generated data in a human resources task. In this task, we ask humans to judge a pool of fictional job-seeking applicants either with or without an image of the applicant's face. We then build a convolutional neural network (CNN) and train it on these human-generated ratings and evaluation or findings. Our contribution in this paper is the empirical evaluation of several different types of methods to reduce these biases using a combination of techniques.

2 BACKGROUND AND RELATED WORK

To understand the long-term implications of implicit and explicit biases in decision-making in machine learning, increasing attention has been put on the fairness, accountability, and transparency of algorithms, particularly in the steps used in making decisions. However, this area of research is relatively new, finding its roots in a 1996 paper on bias in computer systems by Friedman and Nissenbaum [4] and a 2008 paper discussing discrimination in machine learning by Pedreshi et al. in [5]. Discrimination, defined by [6], is defined as the preference (or bias) either for or against a set of social groups that result in the unfair treatment of its members with respect to some outcome. Fairness, in simplest terms, can be viewed as the inverse of discrimination; however, a complete definition of fairness in data is influenced by both culture and context and thus hard to define; Narayanan identified 21 definitions of fairness [7] while in [8] Kleinberg et al. indicated that all definitions of fairness could not be simultaneously satisfied.

Defining fairness and detecting possible biases in datasets are important first steps in addressing them. Consequently, developing metrics to detect bias has been the work of a growing number of research scientists (e.g., [9–11]). Other researchers have developed

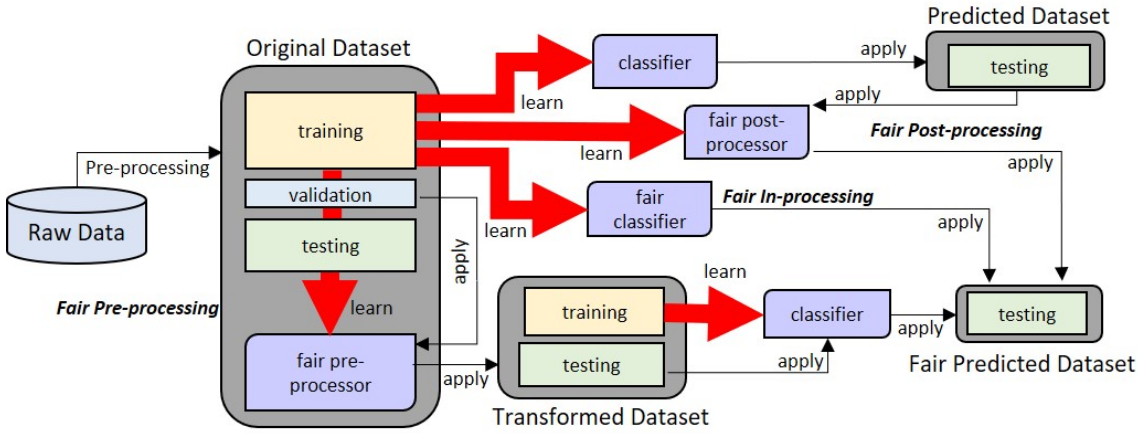

Figure 1: An example instantiation of a generic fairness pipeline consists of loading data into a dataset object, transforming it into a fairer dataset using a fair pre-processing algorithm, learning a classifier from this transformed dataset, and obtaining predictions from this classifier. Adapted from [12].

open-source toolkits to identify the potential for bias, which makes detection of fairness in algorithms easier to achieve (see [12] for a detailed discussion and list of toolkits).

Monitoring biases and fairness helps us address flaws in an algorithm once it is built and trained, but can these biases, once identified, be mitigated? As illustrated in Figure 1, 2these may occur during three separate stages: during algorithm pre-processing, during algorithm in-processing, and during algorithm post-processing. We examine several of the techniques included in the toolkit described in [12].

2.1 Algorithm Pre-processing

Pre-processing is the most effective stage to address biases and involve transforming the training data set. One limitation of pre-processing is that it is often challenging to eliminate biases until they are seen. Learning fair representations [13] develops a new dataset that encodes the data well but obfuscates information about protected attributes; in other words, the model loses any information that can identify whether the person belongs to the protected subgroup while retaining as much other information as possible.

Optimized pre-processing [14] uses a data-driven optimization framework to transform data to reduce algorithmic discrimination probabilistically. The authors apply a randomized mapping to transform the raw dataset into a new unbiased dataset that is used to train the model. This randomized mapping involves editing the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives. Although effective, their results understandably illustrate a trade-off between algorithmic distortion and accuracy.

In [15], three methods are proposed to remove discrimination from the training dataset. A classifier is subsequently learned on this discrimination-free dataset. This approach's rationale is that, since the classifier is trained on discrimination-free data, it is likely that its predictions will be (more) discrimination-free as well. The first, massaging the data, is based on changing the class labels in order to remove the discrimination from the training data. Instead of changing the labels, with re-weighting, the tuples in the training dataset are assigned weights. By carefully considering the weights of each tuple, the training dataset can be made discrimination-free without having to change any of the labels. Not all classifier learners can incorporate weights in their learning process; therefore, they also apply a sampling approach. The weighted dataset is transformed by sampling the objects with replacement according to their determined weights.

In [16] a method is introduced to reduce disparate impact, which we (and the authors in [16]) define as "the 80 percent rule" advocated by the US Equal Employment Opportunity Commission [17]:

Given data set D = (X, Y, C), with protected attribute X (e.g., race, sex, religion, etc.), remaining attributes Y, and binary class to be predicted C (e.g., "will hire"), we will say that D has a disparate impact if

$$\frac{P(C = YES|X = 0)}{P(C = YES|X = 1)} \leq t = 0.8$$

for positive outcome class, YES and majority protected attribute 1 where P(C=c|X=x) denotes the conditional probability (evaluated over D) that the class outcome is c ∈ C, given protected attribute x ∈ X.

Similar to the massaging technique used in [15], the disparate impact reduction method used in [16] edits the feature values to increase group fairness while preserving rank-ordering within groups.

2.2 Algorithm In-processing

In-processing is the most efficient stage to handle bias since it is often unsupervised, and it can thus be self-correcting. Moreover, it does not involve adulterating the underlying training dataset. In [18], adversarial debiasing applies a technique called adversarial training first pioneered in [19] in which multiple networks with competing goals to force the first network to "deceive" the second network. This technique learns a classifier to maximize prediction

accuracy and simultaneously reduce an adversary's ability to determine (and thus exploit) the protected attribute from the predictions, which leads to a fair classifier.

Prejudice remover [20] adds a discrimination-aware regularization term to the learning objective. This method employs two regularizers: an L1 regularizer to avoid over-fitting and an L2 regularizer to enforce fair classification.

2.3 Algorithm Post-processing

With large or complex datasets, the post-processing stage may be an ideal time to handle biases: first, metrics can be applied most accurately at this stage; second, the algorithm does not need to be rerun. In [21], a technique called equalized odds post-processing determines optimal probabilities to change output labels in order to optimize equalized odds, which ensures that no error type disproportionately affects any particular group (as opposed to demographic parity, which requires that a decision be independent of a protected attribute). One benefit is that equalized odds enforces both equal bias and equal accuracy in all demographics, punishing models that perform well only on the majority.

Calibrated equalized odds post-processing [22] takes the work examined in [21] a bit further by optimizing calibrated classifier score outputs to find probabilities with which to change output labels. We find that calibration of equalized odds is often challenging to do in practice.

Reject option classification [23] works by boosting (providing favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups) by examining a confidence band around the decision boundary with the highest uncertainty. It applies massaging and re-weighting techniques like those used in [15], but in post-processing.

3 EXPERIMENTAL DESIGN

Our design objective was to develop and apply a dataset that demonstrated algorithmic bias. This section describes the steps used to create our dataset used to train our machine algorithm and evaluate bias.

3.1 Obtaining Data

We began by obtaining three job descriptions describing jobs for two early-career, and one mid-career job positions from an executive search firm. From this, we obtained information (resumes and cover letters) for a pool of 50 actual job applicants for each position (150 total). We de-identified each applicant (i.e., we anonymized information that might convey the protected classes of age, race, gender, national origin and religion). Job positions and applicants were all located in the United States.

We asked three human resource (HR) recruiters (average experience in evaluating applicants = 10.7 years) to independently rate each applicant's suitability for the corresponding job description on a five-point Likert scale (1 = a poor fit, 5 = an excellent fit). For each applicant, we averaged their score. Inter-annotator reliability, determined by Fleiss' κ, was 0.71, representing 'substantial' agreement [24]. We used the ratings provided by human resource professionals as our gold standard data.

Next, using the Amazon Mechanical Turk (MTurk), a crowdsourcing website[1], we asked human assessors to rate a pool of applicants on the same five-point scale using the job description and the de-identified applicant resumes and cover letters. Each applicant was evaluated by 5 MTurk assessors, and the average of these scores was recorded (the need of 5 assessors was empirically determined in a pilot study). Fleiss' κ, for this set of assessors, was 0.62, also representing a 'substantial' inter-annotator agreement. We also obtained self-provided demographic information about the MTurk assessors. The results of this evaluation became our baseline.

Last, as our treatment, we obtained random facial images from the Chicago Face Database (CFD) [25] and added them to the de-identified information for job applicants used in the baseline. The CFD contains photographs of 158 males and females from four races (Asian, Black, Latinx, and White) as well as normalized subjective information about each image; however, none of the subjective CFD information was used in our training data or provided to the assessors. The distribution of races was applied evenly across gender and race. The applicant's de-identified information and a randomly selected facial image was given to a separate set of MTurk assessors, which rated each applicant on a five-point Likert scale. Fleiss' κ, for this set of assessors was 0.51, representing a 'moderate' inter-annotator agreement.

3.2 Developing our Models

We developed a CNN to incorporate the textual features of each applicant's materials as derived from the 2015 version of the Linguistic Inquiry and Word Count software tool (LIWC 2015) [26] and use them, along with the ratings provided by crowdworkers, for training. The objective of using two models is to examine if the addition of the CFD portrait images biased the findings. In our first model, we used the baseline data; in our second model, we used the treatment data. Each model's output was the predicted rating (reflecting the five-point scale of the gold standard and the MTurk assessors), representing five separate classes. These were compared to the ratings for each applicant from the gold standard data, rounded to the nearest integer value.

3.3 Evaluating our Models

We developed a CNN model using the gold standard rating data as well as features from the LIWC using a 2:1 split between training and test sets (the development of the CNN is part of a larger extension of this study to be published later). Our training dataset was balanced among the five classes by resampling using SMOTE [27]. We applied 5-fold cross-validation and used a SoftMax activation function in the output layer. We find that our model provided a low error rate (indicating the absence of bias, or a 'fair' model), which we attribute to the substantial inter-annotator agreement between our human resource professionals. The accuracy of predicting test labels on our gold standard model is 0.969.

Next, we wish to see if we can replicate this information using our baseline data. The only difference between this baseline model and the gold standard model evaluated previously is the use of different assessors; however, in this model, we train on the baseline

[1] www.mturk.com

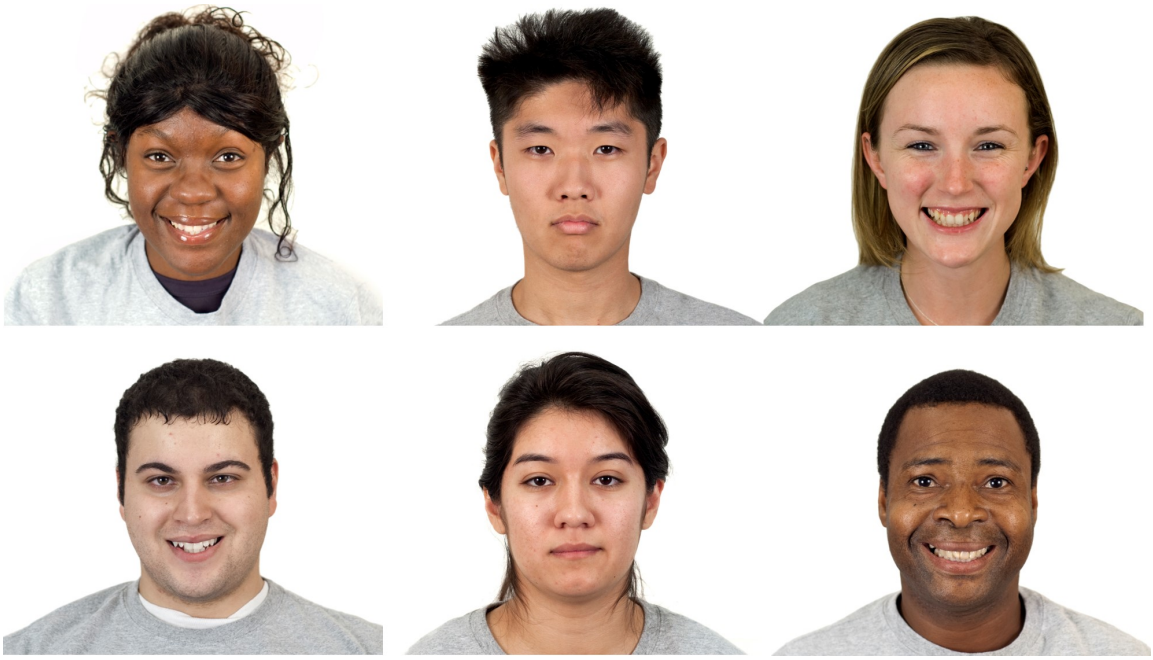

Figure 2: Representative images from the Chicago Face Database (CFD) used in our study.

model data and test on the corresponding unseen applicants in the gold standard data. Our model provided a low error rate, indicating that the use of 5 MTurk assessors can provide similar results as three HR recruiters. The accuracy of predicting test labels on our gold standard model is 0.952.

Last, we wish to see if we can replicate this information using our treatment data. The only difference between this model and the baseline model is the inclusion of CFD images; however, in this model, we train on the treatment model data test on the corresponding unseen applicants in the gold standard data. Our model provided a higher error rate. The accuracy of predicting test labels on our gold standard model is 0.767. Thus, the introduction of CFD face images provided bias in the assessment phase.

3.4 Assessing Bias

We initially investigate the LIWC features for the applicant information that differs between the baseline and the gold standard data. LIWC features were not provided to any of the assessors, so this only provides an indirect examination of the rating differences between the two datasets. Given the small difference in the accuracy rates, the findings were minor. We did find that applicants that provided cover letters that contained more words, more analytical terms, and more authentic terms (as reported by LIWC) were rated slightly more highly (0.15 points) by MTurk assessors than HR recruiters. Likewise, the differences in ratings due to LIWC features between the treatment group and the expert group was also minor, with MTurk assessors rating resumes and cover letters with more analytical terms more highly (0.21 points) than HR recruiters and those using more pronouns were rated 0.14 points lower than the HR recruiters.

More important to determining bias was the impact of the CFD face images on rating each applicant. MTurk assessors provided with face images of Black, and Latinx faces rated them lower, on average, by 0.61 points and 0.43 points, respectively. Applicants who were given Asian faces were rated higher by 0.29 points. These reinforce findings that demonstrate the use of facial characteristics in evaluating others [28, 29].

With respect to age and gender, females were rated lower, on average, by 0.23 points in the treatment dataset. We used CFD face images that contained people ranging in age from 18 to 56, with an average age of 28.8; assessors rated applicants aged 34 or older lower by 0.24 points, those that were 23 or younger rated lower by 0.56 points. This effect implies that our assessors maintain an anticipated age range in mind for each position and infer an age of each candidate from their facial characteristics. This backs up findings in the literature (e.g., [30, 31]). Thus, the differences between the baseline and treatment groups indicate bias based on the face images.

We examined the trustworthiness feature provided for each image by CFD. Although assessors were not provided with this value, we found a strong Pearson's correlation between the rating given and the trustworthiness of the face provided. ($r = 0.734$). This backs up research which finds that trustworthiness can be inferred from facial characteristics [32].

4 MITIGATING BIASES

The primary objective of this paper is to evaluate methods to increase fairness in our biased dataset. In this section, we discuss preprocessing, in-processing, and post-processing methods to mitigate biases from the treatment dataset. We use an increase in accuracy

Table 1: Accuracy and increase in accuracy for each pre-processing algorithm evaluated.

Algorithm	Accuracy	Δ Accuracy
OPP	0.841	0.074
MRS-M	0.844	0.077
MRS-R	0.873	0.106
MRS-S	0.857	0.090
DIR	0.887	0.110

Table 2: Accuracy and increase in accuracy for each in-processing algorithm evaluated.

Algorithm	Accuracy	Δ Accuracy
AD	0.910	0.143
PR	0.869	0.102

as a reasonable proxy for reducing bias. Accuracy is defined as:

$$Acc = \frac{total\ \#\ applicants\ rated\ correctly}{total\ \#\ applicants\ evaluated}$$

where the correct rating is determined by the average rating provided in the gold standard, as rounded to the nearest integer between 1 and 5.

4.1 Pre-Processing Algorithm Evaluation

The most obvious method, as shown with the similarity of the baseline and the gold standard data, is not to use facial images in application materials; however, this is not practical in the context of this job-seeking task for several reasons. First, in many cultures, providing a facial image along with the application materials is standard practice. Second, presuming that application materials are not de-identified, images of many of those applying for a job can be found through social media. Third, if an applicant is invited to interview in person, the same types of rating biases, as seen in our treatment data, will occur.

We now turn our attention to the methods discussed in Section 2.1. We find that it would be challenging to apply the Learning fair representations (LFR) algorithm even with a CNN because the facial characteristics are not provided by discrete features. This leaves three approaches to examine: optimized pre-processing (OPP), massaging-reweighting-sampling (MRS), and disparate impact reduction (DIR). For MRS, the massaging, re-weighting, and sampling apply to three distinct techniques and are applied independently; we abbreviate them as MRS-M, MRS-R, and MRS-S respectively.

Applying each of these to the treatment test dataset, we obtain the following improvements in accuracy (corresponding to a reduction in bias), as shown in Table 1

From Table 1, using the best of these pre-processing steps, we can increase the accuracy by 11% over the raw accuracy score.

4.2 In-Processing Algorithm Evaluation

Turning our attention to the two in-processing methods described in Section 2.2, we now evaluate the adversarial debiasing (AD) and prejudice remover (PR) on the treatment data. As with the pre-processing algorithms used in Section 4.1, each of these is applied independently of any other treatments. The accuracy for each of these in-processing algorithms is given in Table 2

According to Table 2, the best of these in-processing approaches, AD, increased the accuracy (and removed the bias) by over 14%.; however, the integration of these in-processing steps into our CNN involved significant additional complexity (using a generative adversarial network, or GAN, would have reduced the complexity)

4.3 Post-Processing Algorithm Evaluation

We now examine the three post-processing algorithms described in Section 2.3. These include the equalized odds post-processing (EOP), the calibrated equalized odds postprocessing (CEOP), and reject option classification (ROC). Since these apply to the post-processing step, they are the most straightforward to implement. The accuracy for each of these three is provided in Table 3

From the pre-processing approaches we evaluated, as shown in Table 3, we can increase the accuracy by a minimum of 13%.

4.4 Combining Algorithms

Each of these bias mitigation techniques shows promise independently. What happens when we combine methods? We now turn our attention on combining each of the three stages into a single model. We test each of the 72 algorithmic combinations (6 pre x 3 in x 4 post, which includes the omission of using a technique at each stage) and provide results for the best five combinations in Table 4

Examining Table 4 in more detail, the best of these combinations uses the DIR technique for pre-processing and AD technique for in-processing and achieves accuracies (0.949) that approach those in our baseline group (0.952). This indicates for our biased dataset, it is possible to mitigate the biases, but at what cost?

As mentioned earlier, the in-processing algorithms added significant complexity to our CNN model in that they required several steps that required extensive re-tuning. If we just applied pre- and post-processing algorithms, we can achieve slightly lower accuracy, as observed in Table 5

Table 3: Accuracy and increase in accuracy for each post-processing algorithm evaluated.

Algorithm	Accuracy	Δ Accuracy
BOP	0.903	0.136
CEOP	0.901	0.134
ROC	0.897	0.130

Table 4: Accuracy and increase in accuracy for the top 5 pre-, in- and post-processing algorithm combinations.

Algorithms (pre, in, post)	Accuracy	Δ Accuracy
DIR,AD, EOP	0.949	0.182
DIR,AD, CEOP	0.947	0.180
DIR, AD, ROC	0.941	0.174
MRS-R, AD, CEOP	0.933	0.166
MRS-R, AD, EOP	0.931	0.164

Table 5: Accuracy and increase in accuracy for the top 5 pre- and post-processing algorithm combinations.

Algorithms (pre, in, post)	Accuracy	Δ Accuracy
DIR, EOP	0.942	0.175
DIR, CEOP	0.941	0.174
DIR, ROC	0.937	0.170
MRS-R, CEOP	0.927	0.160
MRS-R, EOP	0.927	0.160

By avoiding the in-processing algorithms, we can obtain a significant reduction in bias for our dataset, while incorporating far less complexity in the model.

5 CONCLUSIONS AND FUTURE WORK

Although bias is inherent in human decision making, one objective of machine learning algorithms is to provide accurate results. To accomplish accuracy, there is a need to maximize fairness and minimize biases. In this paper, we show that by using several pre- in- and post-processing algorithms, we were able to mitigate biases effectively, and these were most effective when they are combined.

Our empirical examination involved creating a dataset where information from 150 job applicants was rated with respect to a set of job descriptions. This data exhibited biases due to the inclusion of randomly selected faces from an established dataset. By training our algorithm (a CNN) on this biased data and demonstrate that it provided biased results. We then evaluated ten algorithms that occur during three different stages of our decision-making model (pre-processing, in-processing, and post-processing) to increase fairness in our biased dataset. We saw that every one of the approaches exhibited the ability to increase accuracy (which we use as a proxy for reduction of bias and increasing fairness); however, the pre- and in- processing methods of disparate impact reduction and adversarial debiasing, respectfully, provided the most impactful on mitigating biases. By combining methods at different stages, we were able to reduce nearly all biases introduced by the facial images. We then looked at the additional complexity each algorithm would add and found that we could achieve very good results by eliminating the complex and cumbersome in-processing algorithms. This result implies that there is a tradeoff between more accuracy and more complexity.

There are several limitations to our study. First, we examined a single dataset, so although our results hold promise, the external validity beyond this human resource dataset is uncertain. Moreover, this study is an initial examination of bias reduction and will be expanded to other datasets since fairness is dependent on context and culture.

In future work, we plan to examine other algorithms that show promise for deep learning, such as deep weighted averaging classifiers [33], as well as exploring the role of incentives to reduce bias and increase fairness. Fairness implies that all cognitive biases are equal, but some research shows humans (and the data they produce) are susceptible to certain cognitive biases more often than others, so therefore we plan to examine those cognitive biases that impact fairness the most. We also propose new metrics [34] to help identify these biases and plan to refine these metrics with large datasets.

REFERENCES

[1] Tversky, A., & Kahneman, D. (1986). Judgment under uncertainty: Heuristics and biases. Judgment and decision making: An interdisciplinary reader, 38-55.
[2] Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. Cambridge university press.
[3] Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. Journal of Experimental Psychology: General, 125(4), 387.
[4] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.

[5] Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 560-568). ACM.
[6] Bantilan, N. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. Journal of Technology in Human Services, 36(1):15–30, 2018. URL https://github.com/cosmicBboy/themis-ml.
[7] Narayanan, A. (2018, February). Translation tutorial: 21 fairness definitions and their politics. In Proc. Conf. Fairness Accountability Transp., New York, USA.
[8] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.
[9] Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., ... & Roth, A. (2017). A convex framework for fair regression. arXiv preprint arXiv:1706.02409.
[10] Fish, B., Kun, J., & Lelkes, Á. D. (2016, June). A confidence-based approach for balancing fairness and accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining (pp. 144-152). Society for Industrial and Applied Mathematics.
[11] Hajian, S., Bonchi, F., & Castillo, C. (2016, August). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2125-2126). ACM.
[12] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
[13] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In International Conference on Machine Learning (pp. 325-333).
[14] Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems (pp. 3992-4001).
[15] Kamiran, F., and Calders, T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1):1–33, 2012. DOI: https://doi. org/10.1007/s10115-011-0463-8.
[16] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-268). ACM.
[17] The U.S. EEOC. Uniform guidelines on employee selection procedures, March 2, 1979
[18] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340). ACM.
[19] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).
[20] Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 35-50). Springer, Berlin, Heidelberg.
[21] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).
[22] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems (pp. 5680-5689).
[23] Kamiran, F., Karim, A., & Zhang, X. (2012, December). Decision theory for discrimination-aware classification. In 2012 IEEE 12th International Conference on Data Mining (pp. 924-929). IEEE.
[24] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174.
[25] Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. Behavior research methods, 47(4), 1122-1135.
[26] Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2015). Linguistic inquiry and word count: LIWC [Computer software]. Austin, TX: liwc. net, 135.
[27] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
[28] Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of Neuroscience, 17(11), 4302-4311.
[29] Sergent, J., Ohta, S., & MACDONALD, B. (1992). Functional neuroanatomy of face and object processing: a positron emission tomography study. Brain, 115(1), 15-36.
[30] Roscigno, V. J., Mong, S., Byron, R., & Tester, G. (2007). Age discrimination, social closure, and employment. Social Forces, 86(1), 313-334.
[31] Rosen, B., & Jerdee, T. H. (1976). The influence of age stereotypes on managerial decisions. Journal of applied psychology, 61(4), 428.
[32] Van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. Cognition, 108(3), 796-803.
[33] Card, D., Zhang, M., & Smith, N. A. (2019, January). Deep Weighted Averaging Classifiers. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 369-378). ACM.
[34] Harris, C.G. (2020), Methods to Evaluate Temporal Cognitive Biases in Machine Learning Prediction Models. In Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), Taipei, Taiwan. ACM, New York, NY, USA. https://doi.org/10.1145/3366424.3383418