MITIGATING BIAS IN ARTIFICIAL INTELLIGENCE

A Firms Complete guide

CONTENTS

01 Introduction

The potential benefits of AIs are immense; globally, AIs are projected to generate $15.7T to the global economy by 2030 regarding human resources sectors (PwC, 2020). Across all industries, firms will be looking to leverage AIs; in the next two years, IBM (2020) estimates AIs will be utilized by 90% of businesses worldwide, automating complex-decisions requiring significant capital, like hiring.

However, AI-usage is a double-edged sword; while it presents an attractive opportunity, there lurks the potential to proliferate existing societal biases, resulting in discriminatory classifications. Currently, many businesses overlook or are unaware of such danger. This guidebook examines implications of such biases manifesting and exacerbating in society and businesses.

Main functionalities of this guidebook won't be to only provide firms with awareness, but rather provide businesses with practical incentives to engage in de-biasing, while addressing the current hardships businesses face. Being in an entrepreneur's shoes, this guidebook will assist your business's journey of tapping opportunities of AIs in an equitable and responsible way, while reaping the financial benefits of automation.

CURRENT BACKGROUND

The problem of biases isn't perceived as prevalent in Korea and is rather trivial, due to absence of strict and systematic anti-discrimination measures, or practices like Disparate Impact. The law only states to not engage in discriminatory hiring regarding gender, age, and religion, and fails to quantify the standards and consequences of biased hiring. Consequently, discriminatory hiring is prevalent in Korea – ranking 30th among 36 OECD countries in women employment (OECD, 2017).

While creating this guidebook, four firms providing AI-based hiring services have been consulted (Anonymous Firm A, B, C, D, personal communication, March, 2022); we identified that even firms who specialize in AI-utilization are inconsiderate of the biases. These four anonymous indicated how they have never considered the need of de-biasing; their only concern was accuracy, regarding how accurately they mimic past hiring-decisions. Considering how these firms' algorithms are currently employed by tech-giants of Korea like Samsung and LG, the urgent need for de-biasing was illuminated.

AIs are all involved within the hiring pipeline regarding sourcing, screening, interviewing and selection. This guidebook will focus on screening, where qualifications are used as features to predict whether a candidate is eligible for hiring/interviews.

DEFINING THE PROBLEM

What is bias in AI?

Bias in AI would mean how AIs reproduce biased decisions stemming from human subjectivity. We define bias as the 'prejudice in favor of or against a certain demographic group that is considered to have no actual effect on the outcome'.

How does bias in AI occur?

Machine-learning models make predictions for an outcome, by learning from a training dataset involving features of people, and their actual outcomes. Since the 'answers' are given, the learning is supervised. Algorithmic biases predominantly come from underlying datas, not the algorithm itself. These data in hiring-contexts would refer to historically discriminatory hiring data, and biases occur when the algorithm associates protected attributes as causes of a certain positive outcome. For example, hiring algorithms may identify being male as a contributor to being hired. Protected attributes are attributes that should be 'protected' where discrimination based on these attributes shouldn't be allowed.

Businesses may assert that they encourage fairness by educating, but human-hiring is inherently biased due to cognitive biases. Even after training courses on objective hiring, results still preferred to hire men over women (Moss-Rascusin et al., 2012). This phenomenon can be explained by the existence of confirmation bias. Confirmation bias arises from the tendency to actively search for information that aligns with one's preconceptions. Historically, from the 1980s, Korean firms have been criticized of favoring men over women, due to the perception that women are incompetent, arising from confucianist values (Ministry of Employment, 2020). The negative stereotypical features of a certain demographic may be unconsciously ingrained within the examiner's mind; to confirm their implicit beliefs, they may actively search for candidate's negative information, while overlooking positive attributes that disconfirm their beliefs. This disproportionate tendency to detect positive and negative traits leads to biased decisions. Confirmation biases lead to attentional biases, like inattentional blindness; although the positive features of the candidates are present, the lack of motivation to detect positive traits leads to attention gaps. There also is a top-down attentional bias, as we allocate more attention towards particular features associated with the person, based on our prior-preconceptions. Essentially, observing a certain protected attribute will serve as an endogenous cue, which makes us prioritize paying our peripheral attention towards other stimuli associated with the original cue (Banerjee et al. 2019).

An example would be how when the examiner sees a female, they may unconsciously associate them with features of passivity and lacking initiatives, overlooking activities like sports that imply leadership skills. Instead, when they see that the applicant has worked as a part of a team, they may prone in on how they lack experience in leading a team.

Impacts of Bias

B Biases perpetuated by AIs have formidable societal implications, ranging from worsening inequality, obstructing opportunities for minorities, and hindering efficiency (Mahoney et al., 2020). The biases in AIs generate further biased hiring data which are again used as future inputs.

It isn't only society who suffers; for businesses, being uncovered that biases exist within their hiring practices leads to harmed corporate reputation. Additionally, a biased algorithm may potentially be scrapped, as seen in Amazon's case when their algorithm revealed biases towards women (Villanova University, 2018). This incurs significant costs down the drain.

02 Challenges

What is Fairness?

Quantifying/measuring bias is often conducted through measuring degree of fairness. Fairness itself is also a complex and multidimensional concept, that is defined differently in different domains of studies. This guidebook will pertain to definitions of fairness regarding group-fairness, which assess whether the privileged-group and unprivileged-group are subject to same results.
Regarding group fairness, these fairness metrics are used to quantify the degree of fairness (ada, ada).

1. Average odds difference - average of difference in false positive rates and true positive rates between unprivileged and privileged groups.

2. Statistical Parity difference - difference of the rate of positive outcomes received by the unprivileged group to the privileged group. 0 implies fairness.

3. Equal Opportunity difference - the rate of success for different gro
ups is the same. 0 implies fairness.

4. Disparate impact -

$$\frac{\%\ of\ positive\ outcomes\ for\ the\ unprivileged\ group}{\%\ of\ positive\ outcomes\ for\ the\ privileged\ group}$$

1 implies fairness.

5. Chi-squared test for association - conducts hypothesis-testing to see if certain traits are associated with the positive outcome.

CHALLENGES IN ADDRESSING BIAS

There are several structural factors making biases in AIs exceptionally difficult to address. First factor would be that there often are more lucrative market-priorities than de-biasing. Tech-sector is a competitively-fierce market where being behind on development may be detrimental for businesses; diverting human resources to de-biasing may be risky. The second factor is the lack of actionable guidance towards de-biasing. De-biasing and AI-ethics is a relatively new concept, requiring extensive research and background knowledge on determining most suitable approaches to take and measuring/interpreting progress. Since our objective is to facilitate the process of de-biasing to be accessible for your business, addressing these difficulties will guide our approach.

Why should firms address?

We acknowledge that de-biasing may be challenging. However, there also are potential advantages that businesses can exploit only by engaging in de-biasing.

Main incentives for firms include fulfilling their Corporate Social Responsibility (CSR). CSR refers to promoting ethically-oriented campaigns contributing to corporate ethics. In this case, we specifically refer to promoting fair recruiting processes and societal diversity. The importance of diversity and non-discrimination within workplace is becoming prominent in Korea, as seen in the recent presidential election, where potential enactment of anti-discrimination law was a major controversy (Heo, 2022).

Firms can leverage these aspects as awareness increases. The main benefits can be broken down into three parts:

1. R

ECRUITMENT AND RETENTION:

Firms can gain competitive advantage in the graduate market. Studies by Nikolic et al. (2022) revealed Gen-Zs are more sensitive about social issues, and corporate ethics play a significant role in their consideration. Promoting diversity within the workplace is known to promote higher job-attitudes and increase work performances too (Valentine et al., 2010).

2. R

ISK MANAGEMENT:

CSR can promote brand differentiation, by enhancing brand reputation from firm's positive impacts on society. In Korea, promoting diversity is yet to be the focus of major firms, signifying untapped opportunities regarding this marketing strategy.

3.

BRAND DIFFERENTIATION:

Current business climate regarding employment processes is under scrutiny. By demonstrating CSR, businesses are able to prepare ahead for likely structural changes in Korea; sudden enactment of anti-discriminationlaws will require resources for revision of algorithms, which might destabilize the organization's finance; early-adopters

In the US, a famous firm who leverages the benefits from CSR regarding diversity would be Walmart. Walmart's workforce consists of 21% Black, 17% Latino, 5% Asian associates. In addition, 55% of all associates are women. By leveraging CSR, Walmart is better positioned to tap into the talented pool of minorities, as well as attract socially-aware consumers (Walmart, 2022).

Overall, examining these hidden benefits highlights the positive incentives for firms to engage in de-biasing. This is a direct response to the challenge of there being more weighty market priorities, as firms promoting CSR can reap long-term benefits which normal R&D programmes cannot bring about.

03

CURRENT SOLUTIONS AND APPROACHES

To move from the current-state of having biased-AIs to the goal-state of implementing a fair unbiased algorithm, multiple approaches have been conducted. Studies by Park (2021) has thoroughly examined two popular existing approaches: skill-based assessment approach, and game-based approach.

SKILL-BASED APPROACHES

Here, protected attributes are simply removed from the dataset to prevent the algorithm from classifying applicants based on these attributes, and only skill-related attributes remain. The problem with this approach is that there are many features that act as a proxy. For example, Barocas and Selbst (2016) identified how when Amazon had eliminated gender in their hiring algorithm, the algorithm then gave higher scores to applicants playing lacrosse, a sport typically associated with affluent white men. Simply removing the protected attributes will not remove biases, as the algorithm will continue making predictions based on proxy variables.

GAME-BASED APPROACHES

Other approaches include the game-based assessment approach, which treats the existence of historically discriminative data as a constraint, and rather attempts to work around the constraint by generating new skill-based data. Bogen et al.'s report (2018) outlined that Pymetrics, a dominant firm within this area, customizes games based on their corporate values for applicants to play and generate new data. Candidates have full access to the results, and are able to acquire transparent information regarding their rejection. However, Tambe et al. (2019) criticized how this transparency led external firms to engage in counseling services for applicants at high prices. The firm 'CK Pass' in Korea is an example, who trains applicants to perform well in external tests provided by organizations like Pymetrics (CK Pass, n.d.). Since these services are expensive, it may also exacerbate the biases by systematically putting people of lower-income at a relative disadvantage. Moreover, the fundamental weakness of these tests is that it doesn't necessarily reflect the firm culture or values, as past decisions are totally disregarded; hence, there is no quantifiable measure of 'accuracy' for these methods.

GAP ANALYSIS

Observing the downfalls of two existing approaches and the previously outlined challenges, these conditions for a novel solution can be outlined:

1. Removal of sensitive attributes won't be sufficient, as we can't eliminate all the associated proxy variables.
2. Historically hiring data should not be treated as constraints, and should rather be treated as an obstacle to overcome, as it is necessary to reflect the historical decisions of the firm hiring practices for accuracy; what matters is mitigating the bias involved.
3. De-biasing should be easy/low-cost enough for firms.

AI Fairness 360

With these conditions, we can evaluate AI Fairness 360 (AIF360), a free open-source toolkit regarding de-biasing (IBM, 2021). The toolkit includes several bias detection metrics, and also 10 different bias-mitigation algorithms as a Python library. These bias-detection metrics include metrics like re-weighing and adversarial-debiasing, which satisfies condition 1 and 2, as they do not alter the original datasets and still make predictions based on historical data. However, despite recent advancement within the field of bias-mitigation and the fact that AIF360 provides easy access, AIF360 is not actively used since the official guide provided by AIF360 is mediocre and requires extensive machine-learning knowledge for practical applications. Thus, the third condition is not satisfied. Considering Korea's current employee pool where only 5.1% of firms possess programmers specializing in AI development (KDI, 2020), AIF360's jargon-packed instructions make it difficult for most firms to use.

ADDRESSING THE GAPS

Hence, currently there is a gap regarding the accessibility of de-biasing tools without complex technical jargons, especially considering the lack of AI-developers in Korea. To address this gap, this guidebook will aid the process of de-biasing to be easily conducted by laymen, opening the door of accessible de-biasing. In response, the next section proposes a comprehensive framework which shows an example of applying AIF360 library for real-life datasets, which can be re-applied in your businesses' context with ease.

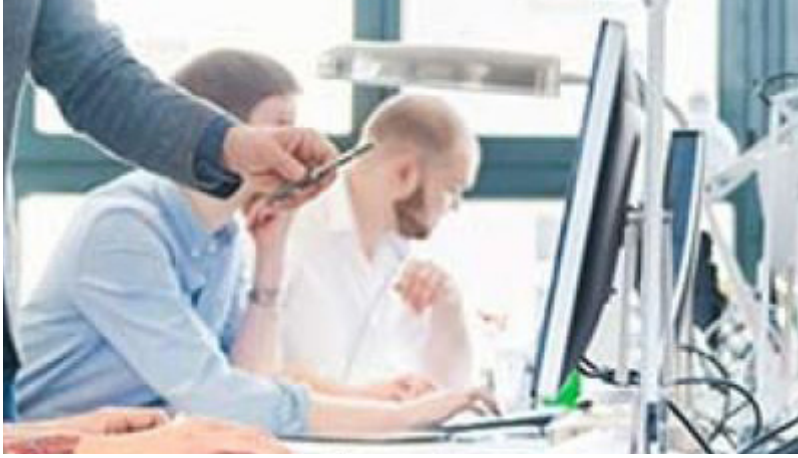
04

PROPOSING THE FRAMEWORK

Before delving into the de-biasing process, it is crucial to learn about the basic machine learning pipeline to understand where and how we will intervene.

The Machine-Learning Framework at its simplified form involves:
1. Splitting the original dataset into the training dataset and the test dataset.
2. Using the training dataset to train a machine learning algorithm
3. Generating the model with the tra ined machine-learning algorithm and test-dataset
4. The model is deployed to real-world data to produce predictions.

As we recognize that it is imperative for firms to prioritize the accuracy and accurately reflect the corporate hiring-culture, among different bias-mitigation methods, methods with the most accuracy were selected. To demonstrate that our approach indeed is effective, we have conducted a test-run with the Adult Dataset from the UCI Machine Learning Repository (Dua and Graff, 2019). The prediction task will be to predict whether an individual makes over 50K USD a year or not, based on their features like education years, gender, marital status, etc. While we couldn't conduct a direct test on hiring data due to privacy concerns, the same logic should be applied to hiring practices.



RECOGNIZING THE BIAS

Firstly, before engaging in de-biasing, it is essential to identify whether a bias exists in the first place. Based on the aforementioned definition of fairness, we may observe whether the protected attributes are independent from the predictions. In our approach, we have conducted a chi-squared test for association. A chi-squared test was specifically chosen, since labels and features were categorical variables. An example of a chi-squared test being conducted on the Adult Dataset with the protected attribute of gender is presented:

1. Identify how many people lie within each category (e.g. how many males have more than 50K income?).
This can be easily done in Python, by importing the data. A simple tutorial is recorded here: (link)
For convenience, these are the values:

Female with >50K - 1769
Female with <50K - 14423
Male with >50K - 9918
Male with <50K - 22732

2. Set our hypothesis:

Null Hypothesis: There is no association between certain gender and whether income is greater than 50K; they are independent.
Alternative Hypothesis: There is an association between a certain gender and income being greater than 50K.

3. Craft a contingency table:

	Income > 50k	Income < 50k	Row Totals
Male	9918	22732	32650
Female	1769	114432	16192
Coulmn Total:	11687	37155	Total: 48842

4. Calculate expected counts for each quadrant:

Expected Values are calculated by $\frac{row \times column}{Total}$.
rowcolumnTotal. An example of the expected count for quadrant Male and Income > 50K would be $\frac{32650 \times 11687}{48842}$.

	Income > 50k	Income < 50k	Row Totals
Male	7812.5	24837.5	32650
Female	3874.5	12317.5	16192
Coulmn Total:	11687	37155	Total: 48842

5. Calculate the test statistic:

To calculate the test statistic, we would calculate the difference between the actual observed values and the expected values, and square that difference. Then, we would divide the value by the expected value. An example for the quadrant Male and Income >50K would be $\frac{(9918-7812.5)^2}{48842}$.

The contingency table of the test-statistics is given here.

	Income > 50k	Income < 50k
Male	567.4	1144.1
Female	178.5	359.9

6. Calculate the P-value:

Then, we can add up the test-statistics altogether. The addition returns 2249.9. The degree of freedom given by
(row-1) X (column-1), is 1. Then, we can compute the p-value from these values in online calculators, accessed here.

Regarding the interpretation of the p-value, we should set the significance-level. While 0.05 is used as a rule of thumb, we should consider the implications of Type-1 errors and Type-2 errors. In this case, Type-1 errors would mistakenly affirm the relationship between gender and income level, while Type-2 errors would fail to reject the null-hypothesis, when there exists an association between gender and income level in reality. Here, committing type-2 errors would be more costly, failing to recognize the bias. Transferring the context to job-hiring, this means we would overlook gender biases within our data. In response, the significance-level should be greater, perhaps 0.1 rather than 0.05.

As the calculated p-value is 0.00001 and 0.00001 < 0.1, we reject the null-hypothesis; there is statistically significant evidence for association between gender and income being greater than 50K, indicating a historical bias within the data.

However, it is necessary that we also calculate the practical significance of our results to observe whether the bias is large enough to be meaningful in real life context. For Chi-squared test for association, Cramer's V is used (Zach, 2021), with the equation

$$\sqrt{\frac{\chi^2}{n*df}} \cdot \sqrt{\frac{2249.9}{48842*1}} = 0.215.$$

In interpretation, considering the degree of freedom, this can be interpreted as a moderately strong association

between gender and income levels being over 50K. This indeed shows that there is a bias within the dataset.

The following chi-squared test for association should be iteratively repeated for different protected attributes to identify all potential biases present in the dataset.

REWEIGHING: PRE-PROCEESSING

The technique of reweighing involves providing 'weights' to different privileged and underprivileged groups. This is a pre-processing method, intervening in the dataset itself. If we upweight positive outcomes of the underprivileged group, this would generally lead to less false-negatives, addressing how certain privileged groups receive more positive outcomes. Reweighing ensures that, within the dataset, the proportion of the unprivileged group that has been classified as the positive outcome is the same as the proportion of the privileged group that has been classified as the positive outcome.
Demonstration: https://www.youtube.com/watch?v=MDQyMrhx-W0
The code is available at: shorturl.at/mISV2 (AIF360, n.d.)

The advantage is that since it does not directly alter the dataset, it is effective in maintaining accuracy. In addition, since this is a pre-processing step, it can easily be combined with other post-processing or in-processing techniques for a more robust de-biasing process. While re-weighing can address different fairness metrics, it is particularly useful for addressing average odds differences (Kamiran & Calders, 2021)

ADVERSARIAL DE-BIASING - IN-PROCESSING

The technique of adversarial de-biasing introduces another classifier that is based on the same machine learning algorithm of the original classifier. While the objective of the original classifier would be to 'predict correct labels based on features', the new classifier would have the goal to 'predict which protected attribute led to the label'. If the classifier is able to predict the protected attribute, that would signify that the original classifier is biased. Essentially, you can think of the second classifier as an 'adversary' that tries to spot bias in your algorithm. The ultimate goal of your algorithm would be to reduce the accuracy of the adversary algorithm, while maintaining the accuracy of the original algorithm. The advantage of this de-biasing method is that it recognizes how accuracy is the paramount interest of firms, and de-biases with minimal accuracy-loss. While adversarial debiasing can address different fairness metrics, it is particularly useful for addressing demographic parity (Zhang et al. 2018).
Demonstration: https://www.youtube.com/watch?v=BkHyvAv1Ais
The code is available at: shorturl.at/ejszM (Hoffmansc, 2021)

SUMMARY

Debiasing through re-weighing and adversarial debiasing are two compatible techniques whereby running both techniques on your dataset and algorithm will ensure a robust algorithm and produce fair predictions. Since they specialize in addressing different fairness metrics, they should be used in tandem. A comprehensive flowchart of the de-biasing approach is presented below:

ORGANIZATIONAL CHANGES

Lastly, it is crucial for businesses to understand that promoting a fair hiring process isn't enough in promoting fairness and diversity within the workplace. A structural change at the organization level must also be made, in order to sustain the diverse environment you create through fair algorithms.
Firstly, they have to be systematically 'aware' of the status-quo of their hiring processes. They can refer to reliable guidebooks in the industry – including our team's – to conduct more organized analysis of fairness in their hiring process. Further, they can benchmark other firms in the forefront of the industry to customize the internal fairness guidebooks taking the 'fit' of their firm in consideration. For example, Google has an internal Diversity Equity & Inclusion department that issues Annual Diversity Report to share the level of diversity present in the firm with their employees. The department also suggests specific feedback for Google to move toward (Google, n.d.).
Next, the increased awareness should prompt concrete action-items. For instance, Twitter has Twitter Academy Program to promote diversity and inclusion, which is to onboard underrepresented demographics in the tech-industry, to their firm (Twitter Academy, n.d.). Twitter can save hiring costs and ensure diversity with the program.

An advantage of a diverse environment would be that it provides scope for detecting biases beyond US-specific protected attributes in the anti-discrimination law, as more perspectives are reflected. Individuals may suggest often-overlooked and context-specific protected attributes like completion of military service, contextualized to Korea.

BUSINESS STRATEGIES

After establishing a fair hiring algorithm and a diverse workplace, businesses can start leveraging this as part of their marketing strategies. A comprehensive initial diagnosis is present through a SWOT analysis.

Strength: De-biased algorithms that other firms lack / diverse workplace environment

Weaknesses: General public's lack of awareness of the problem of biases in hiring algorithms - strengths may not be perceived as impactful

Opportunities: portraying firms as a revolutionary firm / attracting applicants from minority groups / diverse workforce with increased efficiency and better decision-making

Threats: Continued failure of the law to quantify the standard of fairness - other firms may advertise themselves as being 'fair' while in reality they did not engage in de-biasing.

From the SWOT analysis, we can tap into the relative strengths of having an impartial algorithm that other firms lack. This strength can be utilized to exploit the opportunities of attracting minority group applicants. A guiding policy that draws from the diagnosis and the relative strength can be devised, being that the firm should focus on presenting themselves as an ethical and socially-aware firm who recognizes the value of diversity. With the guiding policy, a coherent action can be devised.

1. Evaluate current employer branding

Employer branding is needed to show benefits of being employed, to potential employees. (Ambler and Barrow, 1996). It is a set of beliefs about the firm held by potential employees. It is vital to analyze and identify current trends as to how it is perceived by potential employees so specific improvements in employer branding can be brought. This is achieved by surveys and observational studies.

2. An improved value proposition is decided and marketed to potential employees

Using inferences drawn from surveys, the firm will decide what audience marketing campaigns need to be aimed at. Eg. If a specific firm is perceived to be biased against women but not people of color, the audience will be women, and vice versa. The 4 P's of Marketing will be kept in consideration.

PRODUCT **FAIR SCREENING PROCESS** **DIVERSE WORKPLACE**	**PROMOTION** **ADVERTISEMENT BUDGET BASED ON SURVEY INFERENCES**
PRICE **WORKING FOR THE FIRM**	**PLACE** **THE FIRM**

3. Competitor Research

Keeping Sun Tzu's art of war in mind, the firm would proceed by researching their competitor: Other technology firms in Korea and observe trends in their techniques of advertising openings. In specific, focusing on how and whether they focus on minority groups. Differentiating the firms from other non-debiasing firms can potentially end up in high applicant numbers from minority groups as most firms do not follow these practices as there are no strict laws against hiring biases in Korea.

.

4. Marketing value proposition to current employees:

Lastly, the firm will ensure the new values associated with a more diverse workforce eg. a welcoming environment is reflected by all current employees across all aspects of work. We should ensure that potential employees are aware their needs will be met beyond the screening process. This will give the firm a high employer brand value and will be more attractive to potential employees (Berthon et al. 2005). This step emphasizes on raising awareness through several platforms ranging from word of mouth to online platforms like Reddit and Quora. The goal is to embody the change through example and persuade potential applicants of the authenticity of the firm's campaign, which addresses the weakness section in the SWOT analysis.

CONCLUSION

AI is not just a trend; It is a foundational technology that is penetrating into every field, even disrupting conventions in a novel way. The development of AI is proceeding at an exponential speed. On the brink of the critical point, we have to be aware and beware of the collateral damage it might pose to the society, otherwise it will be very difficult for us to set the flow into the right way in the future. It would be especially ironic that the tool, AI, we implement for societal efficiency might undermine the overriding requirement for a better society - fairness for diversity and inclusion towards social justice. Defining 'fairness' to systematically de-bias AIs cognitively as well as technically will take a lot of effort from diverse stakeholders. However, we can guarantee that the effort will bring about even greater benefits.

REFERENCES

AIF360. (n.d.). AIF360 Python. AIF 360. https://nbviewer.org/github/IBM/AIF360/blob/master/examples/tutorial_medical_expenditure.ipynb#

Ambler, T., & Barrow, S. (1996). The employer brand. Journal of Brand Management, 4(3), 185–206. https://doi.org/10.1057/bm.1996.42

Banerjee, S., Grover, S., Ganesh, S., & Sridharan, D. (2019). Sensory and decisional components of endogenous attention are dissociable. Journal of Neurophysiology, 122(4), 1538–1554. https://doi.org/10.1152/jn.00257.2019

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2477899

Berthon, P., Ewing, M., & Hah, L. L. (2005). Captivating company: dimensions of attractiveness in employer branding. International Journal of Advertising, 24(2), 151–172. https://doi.org/10.1080/02650487.2005.11072912

Bogen, M., & Rieke, A. (2018). Help wanted: an examination of hiring algorithms, equity, and bias. Upturn. https://apo.org.au/node/210071

CKPASS. (n.d.). CKPASS: AI hiring practice. https://www.ckpass.copykiller.com/?gclid=CjwKCAjwu_mSBhAYEiwA-5BBmf3v25qZsTrRt8bDVUkMptg3jwqVfe-AY4hr_gPOy_hzT6_xiB-9qPBoCHAAQAvD_BwE

Google. (n.d.). Building a Sense of Belonging at Google and Beyond. https://about.google/belonging/

Heo, B. Y. (2022, February 14). Two presidential candidates argue over anti-discrimination law. Seoul News. https://www.seoul.co.kr/news/newsView.php?id=20220214500061

Hoffmansc, T.-A. (2021). AIF360 Adversarial Debiasing. GitHub. https://github.com/Trusted-AI/AIF360/blob/master/examples/demo_adversarial_debiasing.ipynb

IBM. (2020). The Business value of AI. https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic

IBM. (2021). GitHub - Trusted-AI/AIF360: A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models. GitHub. https://github.com/Trusted-AI/AIF360

Kamiran, F., & Calders, T. (2011a). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kamiran, F., & Calders, T. (2011b). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

KDI. (2020). Research on firm's cognizance on AI. KDI Journal for Economics and Information, 3. https://eiec.kdi.re.kr/publish/reviewView.do?idx=43&ridx=7&fcode=000020003600004

Ko, H. S. (2019, December 4). Individuals who didn't complete military service aren't eligible for Seoul Metro's full-time job. Yeonhap News. Retrieved April 20, 2022, from https://www.yna.co.kr/view/AKR20191204120400004

Mahoney, T., Varshney, K., & Hind, M. (2020, March). AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning. O'Reilly. https://krvarshney.github.io/pubs/MahoneyVH2020.pdf

Ministry of Employment. (2020). The guide for blind hiring. https://customerfile.incruit.com/2017/09/MOIS_PDF.pdf

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of Sciences, 109(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Nikolić, T. M., Paunović, I., Milovanović, M., Lozović, N., & ĐUrović, M. (2022). Examining Generation Z's Attitudes, Behavior and Awareness Regarding Eco-Products: A Bayesian Approach to Confirmatory Factor Analysis. Sustainability, 14(5), 2727. https://doi.org/10.3390/su14052727

OECD. (2017). The Pursuit of Gender Equality: An Uphill Battle. https://www.oecd.org/korea/Gender2017-KOR-en.pdf

Park, M. S. (2021). When using AIs for screening applicants to employ civil servants such as the police force of Korea, how can we tweak existing algorithms to promote diversity? Minerva University. https://docs.google.com/document/d/1jqXbAPZmZNiqdWG_AHK1ZK3B3YB-ycSoXqwamhOrop8/edit?usp=sharing

PwC. (2020). Sizing the Prize - What's the real value of AI for your business? https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. California Management Review, 61(4), 15–42. https://doi.org/10.1177/0008125619867910

Twitter Academy. (n.d.). Twitter Academy. https://twitteracademy21.splashthat.com/

Valentine, S., Godkin, L., Fleischman, G. M., & Kidwell, R. (2010). Corporate Ethical Values, Group Creativity, Job Satisfaction and Turnover Intention: The Impact of Work Context on Work Response. Journal of Business Ethics, 98(3), 353–372. https://doi.org/10.1007/s10551-010-0554-6

Villanova University. (2021, June 17). Fairness and Bias in Machine Learning. https://taxandbusinessonline.villanova.edu/blog/bias-in-machine-learning/

Wallace, M., Lings, I., Cameron, R., & Sheldon, N. (2013). Attracting and Retaining Staff: The Role of Branding and Industry Image. Workforce Development, 19–36. https://doi.org/10.1007/978-981-4560-58-0_2

Walmart. (2022, April 7). What is Walmart doing to promote a diverse workplace? Ask Walmart. https://corporate.walmart.com/askwalmart/what-is-walmart-doing-to-promote-a-diverse-workplace#:%7E:text=In%20our%202020%20Culture%2C%20Diversity,Islander%20associates%2C%20and%201%20percent

Zach. (2021, September 30). How to interpret Cramer's V. Statology. https://www.statology.org/interpret-cramers-v/

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3278721.3278779

APPENDIX

```
import numpy as np
import pandas as pd
import scipy.stats as stats


#import the csv data - this can be replaced your own data
datas = pd.read_csv("/Users/macbook/Desktop/adult.csv", chunksize=10000000)
#concatenate objects along x axis of male, creating a new dataframe that only includes males
df1 = pd.concat((x.query("sex == 'Male'") for x in datas), ignore_index=True)
#the "sex=='Male'" part can be changed, according to your protected attribute.

#count the values regarding class, which is the income level being >50K or <=50K
print("Male count:")
print(df1['class'].value_counts())


#Print Female index
female = pd.read_csv("/Users/macbook/Desktop/adult.csv", chunksize=10000000)
#concatenate objects along x axis of female, creating a new dataframe that only includes females
df2 = pd.concat((x.query("sex == 'Female'") for x in female), ignore_index=True)
#the "sex=='Female'" part can be changed, according to your protected attribute.

#count the values regarding class, which is the income level being >50K or <=50K
print("Female count:")
print(df2['class'].value_counts())
```

```
Male count:
<=50K    22732
>50K      9918
Name: class, dtype: int64
FeMale count:
<=50K    14423
>50K      1769
Name: class, dtype: int64
```