Adverse Effects of Computing Technology and Their Mitigation: A Study on Bias in Artificial Intelligence

Submitted by: Blake Waldman

Date of Submission: April 7th, 2023

I. Introduction

In the rapidly evolving landscape of artificial intelligence (AI) research and development, the challenge of bias in AI systems has emerged as a topic of significant importance, warranting thorough investigation and the formulation of effective strategies for its mitigation. Furthermore, the nature of AI bias presents a complex web of interrelated factors contributing to biased outcomes, making it necessary for a comprehensive examination of the origins, implications, and potential solutions to this issue.

This paper presents in-depth analysis of the diverse sources of bias in AI systems, including biased training data, algorithmic bias, and measurement and confirmation biases. Through an exploration of literature, the various pathways in which biases can infiltrate AI systems are explained, ultimately leading to outcomes that may be detrimental to individuals and society. Moreover, the ethical, legal, and social implications of biased AI systems are examined, showing the potential ramifications that may ensue in domains ranging from healthcare to law enforcement and beyond. Finally, in response to the challenges posed by bias in AI systems, some popular mitigation strategies are investigated, encompassing conscientious data collection and management, fairness-aware machine learning, algorithmic transparency and explainability, and regular auditing and impact assessments of AI systems.

Furthermore, this paper covers real-world case studies, displaying the diverse manifestations of bias in AI systems and their consequential impacts on various sectors and initiatives that aim to counteract bias and promote fairness in AI development. Examining these real-world examples helps with an understanding of the practical implications of AI bias, highlighting the urgency of addressing this issue in AI research and development. Finally, recommendations for future research and development in AI

bias mitigation are presented, focusing on interdisciplinary collaboration, the development of new techniques and methodologies, evaluating and refining existing approaches, and exploring the potential benefits and challenges of emerging AI technologies. In articulating these recommendations, a collaborative, multifaceted approach to addressing AI bias is emphasized, acknowledging the complexities inherent in the subject matter and the necessity for concerted effort across disciplines.

In synthesizing the presented material, the aim is to contribute to the ongoing discussion surrounding AI bias, providing a comprehensive analysis of the diverse origins, implications, and mitigation strategies for bias in AI systems. By navigating the complex challenge of AI bias through a multi-dimensional lens, a deeper understanding of the issue is created, offering guidance for developing more inclusive, equitable, and socially responsible AI systems that cater to the needs of all members of society.

II. Sources of Bias in AI

Bias in AI systems can originate from many sources, leading to unfair and potentially harmful outcomes. These sources of bias can be broadly categorized into three main areas: biased training data, algorithmic bias, and measurement and confirmation biases. Each category plays a crucial role in shaping AI systems' behavior and decision-making processes, highlighting the need for a comprehensive understanding of their origins and implications to mitigate their impact.

One of the primary sources of bias in AI systems stems from the training data used to build and inform machine learning models. The quality and representativeness of the data used during the training process significantly influence the performance of AI systems. Biased data samples are a significant source of bias in AI systems, and they stem from the underrepresentation or overrepresentation of certain groups within data sets (Barocas, Hardt, & Narayanan, 2019). For example, when training data predominantly consists of samples from a specific demographic, the AI model may perform poorly on other demographics, marginalizing those who are underrepresented. Moreover, historical data can inherently reflect societal prejudices and discriminatory

practices, and AI models trained on such data may inadvertently perpetuate and amplify these preexisting biases (Mitchell et al., 2021).

Similarly, data preprocessing decisions, including feature selection and data cleaning, can also introduce biases. During feature selection, data scientists decide which features to include in the model. These selections can inadvertently introduce bias if certain attributes disproportionately affect specific demographic groups or exclude relevant features (Kusner, Loftus, Russell, & Silva, 2020). In addition, decisions made during data cleaning and transformation, such as handling missing or inconsistent data, can inadvertently introduce bias when imputing missing values based on assumptions that do not hold for all subpopulations, consequently skewing the data and impacting the AI model's performance (Gebru et al., 2018). Addressing these issues in the data collection and preprocessing stages is crucial for developing fair and unbiased AI systems.

Another source of bias in AI systems is algorithmic bias, which can arise from the design and implementation of the algorithms themselves. For example, selecting an algorithm for a particular AI task can introduce bias due to the inherent properties of the algorithm itself, as some algorithms may be more susceptible to bias or overfitting (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). Additionally, inherent biases in the learning algorithms can contribute to algorithmic bias. Sometimes, algorithmic bias can result from including seemingly innocuous features that may inadvertently correlate with sensitive attributes, such as race or gender. This can lead to AI systems that make decisions based on these proxy variables, resulting in biased outcomes even when the model does not directly consider the sensitive attributes. Understanding and addressing these potential sources of algorithmic bias is essential for developing fair and unbiased AI systems.

Finally, bias can arise from the model evaluation process, such as selecting performance metrics and choosing training and validation data sets. Performance metrics play a crucial role in determining an AI model's success and performance; selecting metrics that do not adequately account for disparities across different groups can introduce bias. Such metrics might prioritize overall accuracy at the expense of

fairness, failing to identify or address the discriminatory behavior of the model (Corbett-Davies & Goel, 2018). Furthermore, the choice of training and validation data sets can significantly impact the generalizability of an AI model. If these data sets do not adequately represent the diversity of the target population, the model may not generalize well to real-world scenarios. This can lead to biased outcomes that disproportionately affect marginalized communities and individuals. Addressing bias in the model evaluation process involves careful consideration of performance metrics and the composition of training and validation data sets.

III. Types of Bias in AI

In understanding the biases that can infiltrate AI systems, it is essential to recognize the different types that can manifest, leading to discriminatory outcomes and reinforcing existing inequalities. Various forms of biases, such as direct and indirect discrimination, confirmation bias, and measurement bias, play significant roles in impacting the accuracy and fairness of AI systems. The following sections will delve deeper into these types of bias, examining their effects and consequences in AI systems across different domains.

Direct discrimination is a form of bias when AI systems treat individuals or groups differently based on race, gender, ethnicity, or other protected characteristics. For example, facial recognition algorithms have been found to exhibit higher error rates for individuals with darker skin tones and female subjects, leading to erroneous identifications and potential harm to those individuals (Grother, Ngan, & Hanaoka, 2019). Direct discrimination can severely affect marginalized groups, perpetuating systemic biases and reinforcing existing inequalities.

Indirect discrimination, on the other hand, occurs when AI systems generate biased outcomes that disproportionately harm certain groups, even if the system itself does not explicitly consider protected characteristics. For example, a loan-approval algorithm may inadvertently disadvantage certain groups based on historical lending practices that discriminate against those groups, despite the algorithm not considering race,

gender, or other protected characteristics (Angwin et al. 2016). Unfortunately, indirect discrimination can be challenging to detect and address, as it is often the result of systemic biases and historical practices that have been normalized over time.

Confirmation bias is another type of bias that can emerge in AI systems. This occurs when AI systems reinforce existing stereotypes or beliefs about certain groups, even if those beliefs are inaccurate. For example, a job candidate screening algorithm may systematically reject applicants from specific universities or neighborhoods, perpetuating the false belief that those individuals are not qualified. Confirmation bias can lead to harmful outcomes, particularly for marginalized groups who may already be subject to negative stereotypes and discrimination.

Measurement bias, on the other hand, is a type of bias that occurs when AI systems generate inaccurate or incomplete measurements of certain variables, leading to biased outcomes. For example, an AI-based healthcare system that relies solely on electronic health records may not capture important information about a patient's health that is not documented in their records, leading to inaccurate diagnoses or treatments (Obermeyer, Powers, Vogeli, & Mullainathan, 2019). As a result, measurement bias can have severe implications for the accuracy and fairness of AI systems, particularly in critical domains such as healthcare and criminal justice.

Overall, these types of bias can have severe implications for the accuracy and fairness of AI systems, leading to discriminatory outcomes and reinforcing existing inequalities. Therefore, it is essential to address these issues through various mitigation strategies, such as fairness-aware machine learning, data preprocessing and augmentation techniques, algorithmic transparency and explainability, and AI auditing and impact assessment, to ensure that AI systems serve the needs of all members of society.

IV. Adverse Effects of Biased AI

AI systems can potentially revolutionize various aspects of modern society; however, the presence of bias in AI systems can lead to a wide range of adverse effects, undermining the potential benefits that these technologies can bring. These adverse

effects can span ethical, legal, social, and economic dimensions, impacting both individuals and society. Understanding the full extent of these adverse effects is crucial to develop comprehensive strategies to mitigate bias in AI systems and ensure that these technologies promote fairness, inclusivity, and social progress. In addition, biased AI systems can lead to unintended consequences on social dynamics, reinforcing existing stereotypes and exacerbating social divides. For example, AI-driven advertising algorithms may perpetuate gender stereotypes by targeting ads for traditionally gendered products or services based on users' demographics (Lambrecht & Tucker, 2019). Such consequences may further entrench existing biases and hinder social progress and the dismantling of harmful stereotypes.

The presence of bias in AI systems can result in misallocating resources and opportunities, adversely affecting individuals and society. Biased AI in education, for instance, could lead to the unequal distribution of educational resources or opportunities, disadvantaging students from marginalized groups and contributing to the achievement gap (O'Neil, 2016). Similarly, biased AI in resource allocation for public services may disproportionately affect vulnerable communities, perpetuating systemic inequalities and hindering social mobility. In addition, biased AI systems can erode public trust in institutions and decision-making processes, mainly when AI is used to inform critical decisions in areas such as criminal justice, healthcare, and social services. When AI-driven tools exhibit discriminatory behavior, they can undermine the perceived fairness and legitimacy of the institutions that utilize these tools. Gaining public trust may require significant efforts to address the underlying biases and ensure that AI systems are employed in ways that promote fairness, transparency, and accountability.

Finally, biased AI systems may hinder international cooperation and policy development by exacerbating disparities between countries and regions. For example, AI systems developed in one country or region may not account for cultural, social, or economic differences in other areas, leading to biased outcomes when deployed globally (Vinuesa et al., 2020). This can create challenges for international cooperation, as countries may be unwilling to adopt AI-driven solutions that do not adequately account for their unique

context or that perpetuate existing disparities. Addressing these issues may require collaborative efforts to develop and deploy AI systems that account for diverse perspectives and contexts.

The adverse effects of biased AI highlight the critical need for comprehensive strategies to address and mitigate these biases, ensuring that AI technologies promote fairness, inclusivity, and social progress. By acknowledging and addressing the far-reaching implications of biased AI, stakeholders can work together to develop more equitable AI systems that account for diverse perspectives and contexts. This collaborative effort can help contribute to the realization of AI's potential as a force for positive change in the global landscape.

V. Mitigation Strategies for Reducing Bias in AI

To effectively mitigate the adverse effects of bias in AI, it is crucial to implement a comprehensive range of strategies that address the various sources of bias. These strategies must consider each stage of the AI development process, from data collection to model evaluation.

This begins with adopting responsible data collection and management practices to ensure that training data is representative of the target population and does not perpetuate historical biases (Barocas, Hardt, & Narayanan, 2019). In addition, employing techniques such as oversampling underrepresented groups, data augmentation, and synthetic data generation can help enhance the diversity and balance of training data. These approaches, when executed correctly, can ultimately lead to the development of more fair and equitable AI models.

Fairness-aware machine learning is another critical strategy for mitigating bias in AI systems. By incorporating fairness considerations into the model development process, data scientists can design algorithms that explicitly account for and minimize biases (Zafar, Valera, Rodriguez, & Gummadi, 2017). Techniques such as re-sampling, re-weighting, and adversarial training can be employed to optimize AI models for both accuracy and fairness, ensuring that these systems do not systematically disadvantage

certain groups or demographics. Additionally, fostering algorithmic transparency and explainability is critical for addressing bias in AI systems. This approach enables stakeholders to understand the decision-making processes of AI models and identify potential sources of bias (Guidotti et al., 2018). Developing interpretable AI models and providing clear explanations for their outputs can foster trust in AI systems and facilitate identifying and mitigating biases, ultimately leading to more equitable outcomes.

Regular AI auditing and impact assessment can help identify and address biases in AI systems, ensuring that these technologies are used responsibly and ethically (Raji, Smart, White, Mitchell, & Gebru, 2020). In addition, by evaluating the performance of AI models across different demographic groups and measuring potential biases, organizations can monitor and address potential discriminatory effects, fostering a culture of continuous improvement and responsible AI deployment. Finally, collaboration among stakeholders, including developers, data scientists, policymakers, and civil society, is essential for promoting fairness and inclusivity in AI systems. By engaging diverse perspectives and fostering interdisciplinary dialogue, stakeholders can identify and address the complex ethical, legal, and social implications of AI bias, ultimately promoting the development of AI systems that serve the needs of all members of society.

VI. Case Studies

The following case studies provide real-world examples of biased AI systems and their consequences and initiatives that aim to counteract bias and promote fairness in AI development.

1. Racial Bias in Healthcare Algorithms

    In a study conducted by Obermeyer et al. (2019), researchers discovered that a widely used healthcare algorithm exhibited racial bias, resulting in unequal treatment recommendations for Black and White patients. The algorithm prioritized patients for high-risk care management programs based on their predicted healthcare costs, which were inherently biased due to historical

healthcare access and spending disparities. Consequently, despite having higher health needs, Black patients were systematically underrepresented in high-risk care programs. This case highlights the importance of evaluating AI systems for potential biases and implementing fairness-aware algorithms in critical domains like healthcare.

2. Gender Bias in Natural Language Processing

Natural language processing (NLP) models, such as those used for text completion and translation, have been found to exhibit gender bias. In one instance, an NLP model was shown to produce gender-biased translations, associating male pronouns with specific professions and female pronouns with others (Stanovsky, Smith, & Zettlemoyer, 2019). These biases can reinforce stereotypes and limit the representation of women in specific fields. Addressing this issue requires developing models that are more sensitive to gender bias, along with implementing data augmentation techniques to ensure diverse and representative training data.

3. Bias in Facial Recognition Technology

Facial recognition technology has been criticized for its biased performance, with higher error rates for individuals with darker skin tones and women (Grother, Ngan, & Hanaoka, 2019). These biases can have significant consequences in law enforcement, where misidentification can lead to wrongful arrests or other legal ramifications. To mitigate these issues, it is crucial to develop facial recognition algorithms trained and tested on diverse and representative data sets, ensuring equitable performance across different demographic groups.

4. Countering Bias in AI-driven Recruitment

AI-driven recruitment tools have the potential to streamline the hiring process but can also perpetuate biases present in historical hiring data. In response to this issue, some companies are actively developing solutions to counteract bias in AI recruitment. For example, a company called Pymetrics employs

neuroscience-based games and AI algorithms designed to minimize demographic bias in candidate assessment (Dastin, 2018). This case study demonstrates how AI developers can proactively address bias by designing models and tools that prioritize fairness and reduce the influence of demographic factors in decision-making processes.

These case studies demonstrate the various manifestations of bias in AI systems and their real-world consequences. They also showcase the potential for AI developers to create solutions that actively combat bias and promote fairness. By understanding these examples and learning from them, researchers and practitioners can work towards developing AI systems that serve the needs of all members of society.

VII. Recommendations for Future Research and Development

As AI continues to play an increasingly significant role in various aspects of modern society, addressing the issue of bias in AI systems becomes critical to ensure the development and deployment of inclusive and fair AI technologies. This section presents recommendations for future research and development in AI bias mitigation, focusing on interdisciplinary collaboration, the development of novel techniques and methodologies, evaluating and refining existing approaches, and exploring the potential benefits and challenges of emerging AI technologies.

One recommendation is to foster interdisciplinary collaboration, bringing together experts from diverse fields, such as computer science, psychology, sociology, ethics, and policy studies, to address the multifaceted nature of AI bias in real-world applications (Whittlestone et al., 2019). Such collaborations can facilitate a deeper understanding of the societal, cultural, and historical contexts that contribute to biased outcomes in AI systems and can help develop more effective and comprehensive mitigation strategies that consider these complexities.

Another area of focus should be on the development of novel techniques and methodologies for bias mitigation in AI, including fairness-aware machine learning, data preprocessing and augmentation, and explainable AI. Developing new approaches to

address different types of bias, such as measurement and confirmation biases, can lead to more robust and fair AI systems (Grgić-Hlača et al., 2018). Recent advances in this area include using adversarial learning for fair representation learning (Zhang, Lemoine, & Mitchell, 2018) and developing fairness-aware data synthesis techniques (Kearns, Neel, Roth, & Wu, 2018). Further research should investigate how to effectively integrate these techniques into existing AI development pipelines, ensuring that bias mitigation becomes more integral to AI system design and deployment.

The continuous evaluation and improvement of AI bias mitigation approaches are where future research can make significant strides. To further enhance these efforts, researchers can perform experiments and empirical studies, allowing them to assess the effectiveness of existing techniques in different contexts. In addition, the growth of standardized benchmarks and evaluation metrics that have a focus on bias in AI could facilitate comparisons between different approaches and identify areas for improvement. Furthermore, researchers should explore potential trade-offs between fairness, accuracy, and other performance objectives to develop context-specific recommendations for AI developers and practitioners.

Lastly, emerging AI technologies, such as quantum computing and explainable AI, offer opportunities and challenges in addressing AI bias. Therefore, researchers should investigate how these technologies can be utilized to enhance the fairness and inclusivity of AI systems while remaining vigilant to potential pitfalls and new forms of bias that may arise. For instance, explainable AI methods can offer insights into the decision-making process of AI models, enabling developers and stakeholders to identify and address potential biases more effectively (Adadi & Berrada, 2018).

Addressing bias in AI systems is a complex and ongoing challenge that necessitates the collaboration of researchers, developers, and stakeholders from various fields. Future research and development in AI bias mitigation can contribute to the creation of more inclusive, fair, and socially responsible AI systems by focusing on interdisciplinary collaboration, the development of novel techniques and methodologies, the evaluation and refinement of existing approaches, and the exploration of emerging AI technologies.

VIII. Conclusion

In synthesizing the presented material, it is evident that the complex challenge of bias in artificial intelligence (AI) systems warrants comprehensive examination and thoughtful solutions. The focal point of this inquiry comprised the diverse origins of bias in AI systems, the potential ramifications stemming from such biases, and the formulation of efficacious strategies to ameliorate them.

This paper's principal arguments and discoveries underscore the diverse origins of bias in AI systems, encompassing prejudiced training data, algorithmic bias, and measurement and confirmation biases. Moreover, the repercussions of biased AI systems permeate ethical, legal, social, and economic domains, impacting individuals and society collectively. In response to these quandaries, various mitigation strategies have been proposed, such as conscientious data collection and management, fairness-aware machine learning, algorithmic transparency and explainability, and regular auditing and impact assessments of AI systems. Moreover, the indispensability of interdisciplinary collaboration and the engagement of various stakeholders have been accentuated.

The ramifications of this research transcend the technical facets of AI development, underscoring the necessity for an all-encompassing and cooperative approach to engender more inclusive, equitable, and socially responsible AI systems. By concentrating on interdisciplinary collaboration, cultivating novel techniques and methodologies, evaluating and refining extant approaches, and exploring emerging AI technologies, future research and development in AI bias mitigation can contribute to AI systems that cater to the needs of every member of society. This endeavor ultimately fosters social progress and advances the principles of fairness and inclusivity in AI-driven applications, striking a delicate balance between maintaining academic rigor and embracing the complexity of the subject matter.

IX. Works Cited

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. https://doi.org/10.48550/arXiv.1808.00023

Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women. Reuters. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. D., & Crawford, K. (2021). Datasheets for Datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 51-60. https://doi.org/10.1609/aaai.v32i1.11296

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT)

Part 3: Demographic effects. National Institute of Standards and Technology (NIST). https://doi.org/10.6028/nist.ir.8429.ipd

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys (CSUR), 51(5), 1-42. https://doi.org/10.1145/3236009

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and learning for subgroup fairness. Proceedings of the 35th International Conference on Machine Learning, 2564-2572.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2020). Counterfactual Fairness. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4069–4079.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An Empirical Study of Apparent Gender-based Discrimination in the Display of STEM CCareer Ads. Management Science, 65 (7). pp. 2966-2981. https://doi.org/10.1287/mnsc.2018.3093

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2021). Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm used to Manage the Health of Populations. Science. 2019 Oct 25;366(6464):447-453. https://doi.org/10.1126/science.aax2342

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., & Gebru, T. (2020). Closing the AI Accountability Gap: Defining an end-to-end Framework for Internal Algorithmic Auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in

Machine Translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1679-1684. https://doi.org/10.18653/v1/P19-1164

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Langhans, S. D. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. https://doi.org/10.1038/s41467-019-14108-y

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research. Nuffield Foundation.

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for Fair Classification. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 962-970.

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, 335-340. https://doi.org/10.1145/3278721.3278779