MG 492: Data Governance: Privacy, Openness and Transparency (2022/23)

Candidate Number: 43366

Teacher: Dr Edgar A. Whitley

Number of Words (including footnotes): 5192

Algorithmic Bias Solutions: The Case for Socio-Technical and Institutional Considerations

Abstract

Algorithms are present in all aspects of society. They are pervasive in daily life but also have large impacts on the livelihoods of humans. Despite the supposed objectivity of machines, algorithms have consistently produced biased outcomes that enhance the inequalities and unfair social state of society. In response to this, a wide variety of mathematical solutions have been proposed to treat algorithmic bias. Though ambitious and technically excellent, these computational fixes often fall short for a variety of reasons, especially considering the large potential field of algorithmic bias reduction ranging from software toolkits to the reduction of discrimination in recommendation and predictive AI. Rooted in a fairness framework, I discuss the benefits and drawbacks of current algorithmic bias techniques and propose several recommendations on how to enhance anti-discrimination in algorithms, namely by emphasizing the sociotechnical aspect by increasing human subject involvement. Recommendations in all stages of AI development are proposed, with a structural injustice approach serving as the guideline, with the exception of institutional bias serving as the caveat to an otherwise well-rounded solution analysis to algorithmic bias.

1. Introduction

Since 2014, Amazon has been utilizing experimental recruiting tools that employ artificial intelligence to assign stars to rate their applicants, not unlike how customers rate products on their website (Dastin 2018). It was discovered that the algorithm was not evaluating candidates in a gender-neutral way, with the system effectively teaching itself that male candidates were better (BBC 2018). After it was found that the tool was penalizing resumes with the word *women*, they tried to change the algorithm, but it was clear that the computer models were reflecting the dominance of male workers in the software development industry and projecting it onto the system. By 2017, the effort was dissolved. Even so, blunders like these are symptomatic of the sizeable increase in the role of AI in all aspects of life and the likely bias that will follow it.

In a world before algorithms, humans were the ones that decided who to loan to, who received longer criminal sentences, and who was being hired. These decisions were decided by laws and values that emphasized fairness and equity. Today, these decisions are made by machines under the guise that statistical and mathematical rigor can be substitutes for objectivity. Indeed, algorithms are harnessing huge amounts of power, from tracking students who apply to private universities to allocating creditworthiness, healthcare treatments, and welfare benefits (Panch & Mattie 2019). On a more macro level, algorithms have gone as far as transforming societal and institutional relationships because there is a belief that the only variable aspect of algorithms are how humans utilize them.

However, algorithms are still responsible for implementing values and biases. Not only can they be technically biased by using skewed data, but the very concept of algorithms embodies the idea that a performance is more worthy or important compared to others (Fazelpour & Danks 2021, 3). Rather than eradicating the inequalities that humans perpetuate, they simply "put a magnifying glass up to the issues that riddle society," (Herzog 2021). Algorithmic technologies have been created within the frameworks of societies of power and privilege and therefore reflect the value-laden perspective that certain inputs are better than others. In this way, when algorithms start to control the distribution of public goods and resources, algorithmic bias can result in cruel injustices if certain moral values are prioritized over others.

The goal of this essay is to shed light on solutions that attempt to mitigate algorithmic bias and discuss the highlights and drawbacks of each one. This paper draws on both technical and socio-institutional solutions to algorithmic bias, ultimately concluding that to combat algorithmic bias, there needs to be a computational algorithmic fix that emphasizes fair AI as well as ethical checks that incorporate diverse communities of people in the AI development process. The essay will proceed as follows: first, I will provide the definition of algorithmic bias, why it occurs, and how algorithmic bias fits in an AI fairness framework. Then, I will delve into four technical solutions (AI Fairness 360, discrimination-aware data mining or DADM, FairRecSys, COMPAS) to discuss the drawbacks of purely technical fixes to algorithmic bias while also acknowledging their utility in bias mitigation. I chose these four technical solutions because they exemplify the range of algorithmic bias mitigation techniques in four adjacent but varied manners: software toolkits, data-mining, recommendation systems, and prediction AI. I will

then discuss a structural-injustice approach to reducing algorithmic bias that contrasts the technical solutions, before ending on an acknowledgement of the wider institutional lens in attempting to eradicate algorithmic discrimination.

2. Concepts and Theories

An algorithm is just a set of rules that are used to complete a task (Davis et al., 2021). At the computational level, they are mathematical derivatives that used to be written by programmers; now, machine learning uses special algorithms drawn from large data sets and automated statistical procedures (Kearns & Roth, 2019). Machine learning is used in daily life via social media platforms and search engines or more systematically, by major institutions to determine access to loans or welfare distribution. Though the goal of algorithms is to make decisions more objective and convenient, the execution of algorithms consistently reflect patterns of gross inequality, such as racist facial recognition software that leads to wrongful arrests (Hill, 2020).

Algorithmic bias, defined as a bias that results in systemically unfair or harmful outcomes for certain groups of individuals, repeatedly occurs because the seemingly technical bodies of algorithms are intrinsically social (Bucher, 2018). Because algorithms are created by data and data originate from people, bias will naturally occur from the output of this fundamentally people-driven model. Historical human bias is often amplified in computer models due to the lack of diversity in datasets or the very real institutional inequalities that are reflected in training data (Lee et al., 2019). Especially when the data that is being used includes some sort of bias, naturally, the outcome of the algorithm will likely follow the same trend (Dastin, 2018).

Even if there has been a concerted effort made to create data patterns that do not adhere to bias, due to the nature of pattern recognition within algorithms, they are designed to notice connecting attributes and discriminate against certain features: they are perfect vehicles for inherent bias. Indeed, by using pre-existing data to make new predictions, they will discover patterns that are correlated with gender, class, or race and will treat fresh datapoints in likeness (Herzog, 2021). For example, on CV selection processes, the word *woman* can be unintentionally penalized because it originates from data taken from finance or tech industries, sectors that have typically been male dominated (Shields, 2015). In this way, even if a sensitivity

attribute, a variable that is fed into an algorithm, is intentionally ignored, algorithms can detect and isolate these social categories, whether the computers have been fed the data about the given social category or not.

Mitigating bias and unfairness within algorithms is essential to preventing the impact of perpetuating inaccuracies that could be harmful to large groups of people. Before deciding how to do so, it's important to define fairness within the framework of algorithms and artificial intelligence. This is surprisingly difficult because if the statistical definition of fairness values treating minority groups the same as majority groups are treated, this may affect the broader implications of equity because majority and minority groups often do not have the same background, statistical implications aside (Alikhademi et al., 2021: 2). Additionally, the parity of selected statistical measures between majority and minority groups may give contradictory results to statistical fairness, and statistical fairness is limited to clearly defined classifications (Alikhademi et al., 2021: 3). Furthermore, researchers have pointed out how many notions of fairness are in opposition, not only in how to measure fairness but how drastically different outcomes result from different definitions (Ryan, 2006). The answer to this disagreement does not lie in merely creating better algorithms because the very normative standard for "better", assessing an algorithm's fairness, is the position of disagreement (Wong, 2020). Not only this, but researchers have shown that it is not possible to satisfy fairness definitions simultaneously, so the clarity on which bias mitigation strategies can be considered "best" is foggy and uncertain (Miconi 2017).

In this way, if it is mathematically impossible for algorithms to adhere to multiple fairness measures, it may be more helpful to tailor different fairness measures to different interests and stakeholders that are affected by the algorithmic bias (Narayanan, 2018). For example, when the fairness of COMPAS, an algorithm that helps judges decide a defendant's recidivism before they stand trial, was put into question by companies ProPublica and Northpointe, they had two very different understandings of fairness. Northpointe argued that the algorithm was not biased because the reoffending rate was the same on the COMPAS scale regardless of race, while ProPublica highlighted that for those who were not in the reoffend group, black defendants were more likely to be classified as high risk of reoffending (Wong, 2020). In this instance, the two companies have different understandings of violations of fairness, with one focusing on

disparate impact (the decision especially impacts protected groups) and the other emphasizing disparate treatment (protected features are explicitly used in the decision), (Wong, 2020).

In this case, judges will focus on positive predictive values of COMPAS, such as how accurately recidivism can be predicted while defendants will be concerned of their chance of being misclassified as medium or high-risk and facing a worse penalty (Narayanan, 2018). Thus, it is important to go beyond algorithmic fairness as a merely technical measurement and instead highlight the relationship between the interest of the stakeholders and fairness measurements. Any fairness measurement will inevitably favor one group over another and so, a fairness framework in AI should instead look at the groups involved and optimize a diversity of political, social, and ethical balancing of the stakeholders involved. AI fairness should work within the structure of an "all-affected principle" which expresses how those who are affected by a decision should be a part of the decision-making process (Dahl, 1990: 49). Indeed, in the following solutions proposed, the all-affected principle is considered as a metric within the fairness framework because it seeks to emphasize how researchers and developers must consider the impact of their algorithm and work with the affected populations to mitigate bias. Only when the people who are affected by the algorithmic bias are considered in the artificial intelligence development process is true fairness optimized.

3. Solutions

3.1 AI Fairness 360

To address the question that burden AI developers on how and whether data should be debiased, IBM created the AI Fairness 360, an open-source toolkit whose goal is to detect and mitigate algorithmic bias (Bellamy et al., 2019: 41). Toolkits are functions that can be accessed through programming languages that takes steps to lessen bias; they are often used in conjunction with checklists that are written by AI practitioners that ensure that ethical thought is included in machine learning pipelines (Richardson et. al, 2021). Focused on the developer-side of algorithmic bias, IBM created AI Fairness 360 to ensure that engineers have an open-source platform that prioritizes code quality while also implementing bias metrics through

demos and documentation. This software toolkit comes in the form of a Python packages that includes 71 bias detection metrics, 9 bias mitigation algorithmics, and an explanation workbook that allows users to understand the meaning of the algorithm results (Bellamy et al., 2019: 42). It also contains tutorials on credit scoring, predicting medical costs, and ordering face images by gender (Panch & Mattie, 2019).

While this software toolkit is impressive in that it prevents the development of machine learning models that privilege certain groups systemically and seeks to treat bias in training data from under or over sampling, flaws in AI Fairness 360 can reveal broader insufficiencies of software toolkits in mitigating algorithmic bias. Because AI Fairness focuses on implementation of large amounts of debiasing techniques on a group level for classification issues, its utility is only in a very limited setting (Lee & Singh, 2021). Even IBM Fairness warns that it should only be used for risk assessment problems that have well-defined protected attributes that have a level of statistical sameness (Bellamy et al., 2019: 48). Additionally, it is extremely data-driven and unable to focus on problems with more than just a binary classification. If the data-mining process is flawed or the challenge requires interacting sensitive attributes, it could result in intersectional discrimination (Lee & Singh, 2021: 5). Furthermore, other issues have been identified like its steep learning curve, lack of a tailored user experience, imbalances in overload and oversimplication, limited coverage of fairness considerations beyond the model testing step, and limited ability to integrate in real-life circumstances (Lee & Singh, 2021: 7). Drawing from these conclusions, though software toolkits can be helpful in base-level bias mitigation of simple datasets, IBM Ai Fairness 360 and other similar software toolkits are unlikely to completely transform the industry in identifying and mitigating unfairness in more complicated and impactful models.

3.2 DADM

To address data mining that naturally creates discrimination, the Fairness, Accuracy, and Transparency in Machine Learning (FAT/ML) research community proposed discrimination-aware dating mining (DADM). Data scientists employ this technique in the hopes that it would reduce bias by starting with an older model and using that to search for other models that are

comparatively non-discriminatory (Schmidt & Stephens, 2019: 140). The very point of data mining is to use data to create discrimination: the goal is to find a rule that makes distinctions based on certain attributes. DADM seeks to prevent creating these distinctions by identifying "bad patterns" and filtering them out (Berendt & Preibusch, 2014: 178).The "good" patterns are kept. In a typical data mining process, a descriptive analysis sheds light on imbalances that identify a feature that predicts a poor outcome (like loan applicants who cannot pay back their loans). People with this feature are then separated from the population, and the decision will then discriminate against customers with this feature, thus reducing the undesirable outcome (Berendt & Preibusch, 2014: 178). DADM would be incorporated into this process by constraining the decision that discriminates against the people that holds that feature.

Normal data mining is helpful in the sense that it detects discrimination in a data set and identifies statistical imbalances that originate in the data. DADM utilizes extra background knowledge about sensitive attributes to detect discrimination, taking it one step further than traditional data mining. However, DADM is so constrictive that it prevents new insights and hypotheses to be tested because it is not exploratory and therefore, it prevents socially relevant discoveries from being tested (Berendt & Preibusch, 2014: 199). For example, though DADM has been able to detect discrimination against women in the workplace by detecting discrimination-indexed features and highlighting awareness about it, the classification of "mothers" who experience discrimination has been an emerging pattern that has gone unnoticed by DADM. In this case, discrimination indexed attributes refer to discrimination grounds that are legally coded and may include other attributes (Ruggieri, 2010). DADM has identified "having children" as not being classified as predictive of gender, so the risks of "having children" was not grouped within "female" to be discovered as the feature of "mother" (Berendt et al., 2008, 120). In this way, it shows the limitations of DADM and how feature constructions within data mining require lots of background knowledge within the algorithm, something that may not be possible without the extra help of human indexing.

These flaws in DADM reflect how algorithms that claim to be blind to certain attributes can yield bias even without the input of sensitive attributes. For example, when Amazon decided to exclude specific neighborhoods from its Prime delivery system, the parameters of the algorithm followed several considerations: enough Prime members with that zip code, being close to a warehouse, and sufficient labor to deliver to that location (Lee et al., 2019). Though

these factors never outlined any racial or socioeconomic metrics, they excluded poor neighborhoods made up of people of color, creating data points that were just proxies for racial and economic classifications. Thus, in both cases of intentional exclusion of sensitive attributes, and conscious constriction of discrimination in DADM, the possibility of creating biased results still occurs, ultimately pointing to flawed models of mitigating algorithmic bias from a purely technical perspective.

3.3 FairRecSys

Another solution addressing algorithmic bias proposes transforming recommendation and personalization systems through the development of an entirely new algorithm FairRecSys. FairRecSys attempts to postprocess recommendation matrices while maintaining the utility of personalizations (Edizel et al., 2020: 197). In the context of algorithms, postprocessing procedures refer to the pruning, rule filtering and knowledge integration after the algorithm has been run (Bruha & Famili, 2000: 112). The issue that is identified in recommendation matrices is that it is possible to predict a user's sensitive attribute, like gender or race, by simply looking at the recommendation matrix, which includes the recommendation for each individual user. To address this bias, the goal of the algorithm is to limit the predictability of these sensitive attributes from the results (Edizel et al., 2020: 201). The FairRecSys works by going through the post-process of a recommendation matrix and correcting the potential biases of the output of the algorithm.

To test the algorithm, the developers of FairRecSys used two real-world data sets from Reddit and MovieLens, prevalent datasets in recommender system scholarship (Harper & Konstan, 2016). Using the demographic of gender as the sensitive attribute for both datasets, they ran it through the FairRecSys algorithm and found that the Reddit dataset carried a higher bias than the MovieLens because the Reddit dataset had smaller balance error rate values, indicating lower rates of predictability and higher rates of fairness (Edizel et al., 2020: 210).

Though this algorithm is a very good attempt at mitigating bias in algorithms, it only achieves measuring fairness and bias between two sensitive attributes. This makes it difficult to apply because users that articulate preferences in recommendation systems will often belong to

multiple discriminated groups. This is not the only drawback of this algorithmic bias proposal. When FairRecSys incorporated the algorithmic constraint for fairness, this caused the algorithm to sacrifice utility and accuracy. This brings about a larger discussion of an accuracy tradeoff that will often occur in algorithmic predictions that prioritize anti-discrimination, privacy, or fairness (Edizel et al., 2020: 211). When there is an optimization for ethical concerns and mitigating bias, in the case of the FairRecSys, there was a deterioration in the recommendation quality. I will be looking further at the fairness and accuracy tradeoff in the following section that discusses the COMPAS algorithm.

3.4 COMPAS

The COMPAS algorithm is an algorithm utilized by judges that helps them decided whether a defendant should be released or not while they stand trial (Angele & Rosenblatt 2015). It assigns a risk score between 1 and 10 to determine how likely they are to commit a violent crime based on various sensibility attributes like age and sex; in this case, violent crime is also a variable. An example of its application is if a defendant scores 4, they are half as likely to reoffend compared to a defendant that scores an 8 who will have a higher likelihood of being detained while awaiting trial (Angele & Rosenblat, 2015). Though race is not an explicit input, black defendants have been shown to be classified as high risk at a higher rate: being classified as high risk meant receiving harsher treatments by the court (Angwin et al., 2016: 5).

There have been reformulations on making the COMPAS fairer by constraining optimization. The goal of one study published in ACM reformulated the algorithmic fairness by constraining for optimization; the goal was to maximize public safety while also satisfying fairness requirements (Corbett-Davies et al., 2017: 798). The results of the study found that the optimal algorithm required using multiple race-specific thresholds to produce a defendant's risk score. For example, if a white defendant scored above a 4, they would be detained, compared to a black defendant who would only be detained if their score was above a 6 (Corbett-Davies et al., 2017: 799). In contrast, an optimal unconstrained algorithm would require a single equal threshold to all defendants: all are held to the same standard regardless of race. Here, there is tension between constrained and unconstrained algorithms due to the tradeoff that happens

when reducing racial discrimination and optimizing public safety – in essence, it's a fairness-accuracy tradeoff. Indeed, the conclusion found that maximizing public safety would penalize communities of color while adhering to legal and institutional fairness would correlate to the releases of high-risk defendants that could endanger the public (Lee et al,, 2019).

In this way, it is easy to treat data discrepancies in the COMPAS algorithm, but fairness is still difficult to define and measure. As shown, companies and national organizations have been seeking to measure fairness that a developer can simply incorporate, but this becomes increasingly difficult when determining the right levels of trade-off between accuracy and fairness. In this way, perhaps human discretion cannot be replaced by the decision-making made by algorithms, and the ethical discretion of developers, policymakers, creators, and government workers should not be underestimated in the execution and stages of algorithmic development.

Though the argument could be made that it is possible to change variables based on the needs of the algorithm, with the problem lying in data sets that are not representative of certain groups, this ignores the fundamental tension between metrics for fairness and the difficulty in choosing between them. Different aspects of an algorithm and its fairness emphasize a variety of performances: the very act of choosing a certain fairness metric is a political choice that actively decides to uphold one point of view while silencing another one (Friedler et al., 2021). The task of choosing the correct fairness metric is in the hands of developers, making the machine-learning inherently wrought with bias and varying levels of priorities. Indeed, the impossibility of solving algorithmic bias lies in the task of constantly having to choose between fairness measures – fairness definitions will often restrict the development of algorithms and will therefore yield drastically different models.

3.5 Structural Injustice Approach

In contrast to attacking the issue of algorithmic bias with technical recommendations, solutions have been proposed that address the structural and systemic discrimination that algorithms exist within, instead of just transforming the way algorithms mathematically function. Increasingly, there has been a shift from individual evaluations of AI systems to a more structural approach to how algorithms function within the broader social world. In this way, if

algorithmic bias is just a form of repackaged institutional bias, a form of structural injustice "that exists when AI systems interact to merely exacerbate existing inequalities", then the way to attack algorithmic bias is to take a systemic approach (Lin & Chen, 2022).

Indeed, social institutions are created in a manner that produces discriminatory outcomes regardless of the efficacy of automated systems (Creel & Hellman, 2021: 3). In taking a structural-injustice approach to AI fairness, there are several measures that can be taken to encourage fairness within AI system development. A structural-injustice approach is an approach to AI performance that considers the power relations that are laced into AI development (Lin & Chen, 2022: 14). A perpetuator of the structural injustice framework when looking at AI bias, Lin & Chen argue that all participants in a social structure have a hand in AI bias, and the bearers of responsibility not only include developers but the CEOs of tech companies, policymakers who influence AI law, and those funding the algorithm development (Lin & Chen, 2022). Thus, the core recommendation for addressing AI fairness seeks to reconcile this structural injustice model by emphasizing the need for diverse human involvement at all stages of algorithmic development.

For the first stage, problem selection, the recommendation focus is on assessing social contexts and existing social inequalities by identifying the benefits and supposed risks that an AI would create. It also emphasizes attention to the hierarchies of power within the decision-making process, like who decides what issues to address; for example, in the tech industry, it should be those in marginalized positions that should be the drivers towards diverse representation, (Lin & Chen, 2022: 22). This would encourage developers to take a big-picture approach to AI development and be more intentional to the full impact of their algorithms, avoiding the unintentional outcomes of discriminatory AI. The issue with this recommendation is a lack of regulatory sandboxes that would allow for innovation in algorithms but also limit the legal discriminatory capabilities of AI. As of right now, policymakers have been unable to keep up with the high rate of innovation in machine learning and have not enacted regulations that sufficiently deal with the nuances of algorithmic bias; in this way, this recommendation is very idealistic in believing that developers will have the internal drive to thoroughly assess institutional effects of their algorithm without the pressure of formal government regulations.

In the data curation stage, efforts should be made to level out resource distribution and create more representational unbiased input data by having an institutional level reflection on

collecting accurate, available, and indicative datasets (Lin & Chen 2022). This would entail collecting higher amounts of datapoints with racial and socioeconomic diversity to prevent algorithms from developing predictions based on incomplete sources that could result in uncertain outcomes (Pierson et al., 2021). This would address the issues surrounding data mining as more representational data distribution would translate into less bias being embedded into the data itself. For example, when examining AI that helps those with osteoarthritis, algorithms that use existing datasets of knee X-ray images resulted in inaccurate diagnoses due to predictions based on insufficient available X-ray images of communities of color (Pierson et al., 2021). Indeed, though there is a desire to have more representational bias, the solutions are more institutional, such as creating more inclusive clinical guidelines or giving underserved communities more access to healthcare.

In the third stage, algorithm development, it is important to have critical assessments of potential associated factors to avoid the reoccurrence of existing inequality, such as the link between zip codes and racial makeups of neighborhoods (Cowgill & Tucker 2017). This is where members of the public should collaborate with developers and specialists to create demographically representative dialogue, deliberation, and input on the process of AI development. A model of this has been done before: the National Institute for Health Research created Citizens' Juries on Artificial Intelligence gathered groups of citizens to produce a final report on AI and accuracy & explainability tradeoffs (Van der Veer et al., 2021). This allowed for a larger sample of people's preferences on accuracy & accessibility, successfully creating representational guidelines for machine learning systems. Indeed, because definitions of fairness are so variable and developers are unable to simply incorporate it in algorithms like COMPAS, involving community deliberation on AI development can reconcile this gap.

The last stage should peer into the system's true impacts on the real-word and make necessary adaptations (Lin & Chen 2021). As articulated by the structural-injustice analysis, unjust outcomes will form even when careful examinations and guidelines have been adopted. This would also require the collaboration of different trades, from policymakers to business corporation owners, to the developers, and to the CEOs who distribute the resources that allow AI development to get made. In this way, the four stages of AI development should not be run by single drivers. Collective action should be mobilized to address structural injustice from the beginning to the end in pursuit of the goal of AI fairness.

3.6 Limitation: Institutional Bias

Though the structural injustice approach rightfully emphasizes the human involvement in AI development, sometimes bias is so intrinsic that algorithms cannot be unbiased until institutional norms are repaired (Flowerman, 2023). A paper released in *Ethics and Information Technology* examines this bleak reality that perhaps there is no actionable solution to algorithmic bias because discrimination is so embedded in our social world. If we are to use the example of credit, an institution that assigns norms on how finances are managed, recovered, and lent by companies, algorithms are used to perform risk assessments on how likely a person will default on a loan. Using a borrower's debt loan and bill-paying history, financial algorithms determine their credibility by outputting a FICO score (O'Neil, 2016).

It's true that creditworthiness is no longer determined by local bankers who are vulnerable to personal prejudice. However, no matter the accuracy of its data mining, the algorithm will always have a degree of bias and discrimination because the institution of credit is fundamentally classist (Flowerman, 2023). Being able to build credit is necessary to buying property or getting access to housing; this structure disadvantages many who are not financially able to participate in purchasing and re-paying because they belong to systematic cycles of poverty and debt (Flowerman, 2023). In this way, it is not the algorithm which limits certain groups but the institution itself that closes off opportunities to those who cannot build credit. Thus, an extreme fix to algorithmic bias would be to completely abolish certain institutions instead of treating minute aspects of the problem.

4. Conclusion

Regardless of the ever-presence of institutional bias in algorithmic development, there is still hope for algorithmic bias mitigation. The awareness of institutional discrimination should only encourage the future of AI development to be more mindful of looming power and social structures to create algorithms that are fit to adequately deal with them. In the advent of unparalleled AI development and technical algorithmic excellence, a multidisciplinary approach

that emphasizes human involvement in the execution of all stages of algorithms has been the missing link in reducing algorithmic bias – rather than replacing human choice, algorithms work best by working alongside our decisions.

In examining four different technical fixes to algorithmic bias, AI Fairness 360, DADM, FairRecSys, and COMPAS, all of which span across far reaches of the field of algorithmic discrimination and mitigation, I attempt to point out that there is dimension beyond computational excellence in fostering algorithmic fairness. Using a fairness framework that incorporates an all-affected principle, it becomes clear that algorithmic reparation cannot have a one-dimensional fix and requires the nuance and intersectional solutions of approaches that acknowledge the embedded inequalities that translate into technology. Because algorithms embody the stratified hierarchies of human life, it makes sense that, in conjunction with software toolkits and discrimination aware data mining, there must be human feedback that provides checks on the efficiencies of algorithms. This is why, even with all the precautions of varied computational algorithmic mitigation tactics, people must play an active role in dynamically correcting bias before, during, and after the development, testing, and launch of algorithms. From the data-mining stage to the post-processing step, algorithms are meant to complement human judgement. To exacerbate the accuracy of algorithmic effectiveness, the consultation of wide varieties of people must be consulted: not only those involved, but those who are the subject. In this way, it is increasingly important to promote the norm of collaboration between engineers, developers, policymakers, and investors to optimize algorithmic models that reflect the well-roundedness of those who were involved.

At the core, the answer to algorithmic bias lies in the fairness framework: the subjects of "robotized" decisions deserve to know the bias that affects them and contribute to the system that could be marginalizing them. When feedback occurs, there is a higher likelihood of anticipating origins of bias, creating less of a chance of repeat in the future. In this way, a sense of justice is preserved, as those who are potentially the most affected by biased decision-making can now take on the powerful role of mitigating it. It is true that humans are inherently biased, and as much as we had hoped that our algorithms would not be, they have unconsciously absorbed the structural inequalities of society. However, in concurrence with specific discrimination aware algorithmic measures, it may seem that the answer to the solution

to algorithmic bias can be found in the increased involvement of the very arbiters of the bias in the first place.

Works Cited

Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2021). Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv.*

Angele, C. & Rosenblat, A. (2015). Courts and Predictive Algorithms. *Data & Civil Rights: Criminal Justice and Civil Rights Primer*.

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

BBC. (2018). Amazon scrapped 'sexist AI' tool. *BBC.*

Bellamy, R. K. E., et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63 (4), pp. 41-55.

Berendt, B. & Preibusch, S. (2014). Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence Law.* 22, pp. 175–209.

Berendt, B., Preibusch, S. & Teltzrow, M. (2008). A privacy-protecting business-analytics service for online transactions. *International Journal of Electronic Commerce*, 12, pp. 115–150.

Bruha, I., & Famili, A. (2000). Postprocessing in machine learning and data mining. *ACM SIGKDD Explorations Newsletter*, *2*(2), pp. 110-114.

Bucher, T. (2018) *If… Then: Algorithmic Power and Politics*. NY: Oxford University Press.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM International Conference on Knowledge Discovery and Data Mining*, pp. 797-806.

Cowgill, B., & Tucker, C. (2017). Algorithmic bias: A counterfactual perspective. *NSF Trustworthy Algorithms*.

Creel, K. & Hellman, D. (2022). The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy,* pp. 1-18.

Dahl, R. A. (1990). *After the revolution? Authority in a good society*, Revised Edition. New Haven: Yale University Press.

Dastin, J. (2018, October 8). Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters.*

Davis, J. L., Williams, A., & Yang, M. W. (2021). Algorithmic reparation. Big Data & Society, 8(2).

Edizel, B., Bonchi, F., Hajian, S. *et al.* (2020). FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science & Analytics,* 9, pp. 197–213.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), e12760.

Flowerman, C.H. (2023). (Some) algorithmic bias as institutional bias. *Ethics of Information Technology,* 25 (24).

Friedler, S., Schneidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making. *Communications of the ACM,* 64 (4), pp. 136-143.

Harper, F.M. & Konstan, J.A. (2016). The movielens datasets: history and context. ACM Transaction on Interactive Intelligent Systems, 5(4), pp. 1-19.

Herzog, L. (2021). Algorithmic Bias and Access to Opportunities, in C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics*. Oxford Academic.

Hill, K. (2020). Another arrest, and jail time, due to a bad facial recognition match. *The New York Times*

Kearns, M. & Roth, A. (2019) *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

Lambrecht, A. & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads, *Management Science*, 65:7, pp. 2966-2981.

Lee, M. S. A. & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. *Proceedings of the 2021 CHI conference on human factors in computing systems.* 699, pp. 1-1 3.

Lee, N.T., Resnick, P., & Barton, G. (2019). Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms. *Brookings Institute*, Washington, DC.

Lin, T. A., & Chen, P. H. C. (2022). Artificial Intelligence in a Structurally Unjust Society. *Feminist Philosophy Quarterly, 8(4)*, Article 3.

Miconi, T. (2017). The impossibility of "fairness": a generalized impossibility result for decisions. *arXiv.*

Narayanan, A. (2018). 21 fairness definitions and their politics. https://www.youtube.com/watch?v=jIXIuYdnyyk. Accessed 23 April 2023.

Noble, S.U. (2018). Algorithms of Oppression. New York University Press, New York.

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

Panch T. & Mattie, H. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*. 9(2).

Pierson, E, Cutler, D., Leskovec, J., Mullainathan, S. & Obermeyer, Z. (2021). An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations. *Nature Medicine.* 27 (1), pp. 136–40.

Richardson, B. & Gilbert, J.E. (2021). A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. *ArXiv.*

Ruggieri, S., Pedreschi, D. & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data*, 4(2), pp. 1–40.

Ryan, A. (2006). Fairness and philosophy. *Social Research, 73*(2), pp. 597–606.

Schmidt, N., & Stephens, B. (2019). An introduction to artificial intelligence and solutions to the problems of algorithmic discrimination. *arXiv preprint arXiv:1911.05755*.

Shields, M. (2015). Women's participation in Seattle's high-tech economy. *Georgia Institute of Technology.*

van der Veer, S.., Riste, L., Bozentko, K., Atwood, S., et al. (2021). Trading off Accuracy and Explainability in AI Decision-Making: Findings from 2 Citizens' Juries. *Journal of the American Medical Informatics Association. JAMIA.* 28 (10) pp. 2128-2138.

Wong, P.H. (2020). Democratizing Algorithmic Fairness. *Philosophy & Technology,* 33, pp. 225-244.