An Ethical Turing Test:

Analyzing Structural Injustice in Algorithmic Fairness Measures

Becca Smith

Philosophy Honors Thesis

Professor Kok-Chor Tan

February 24, 2024

Abstract

In this paper, I use Iris Young's conception of structural injustice to analyze the drawbacks and limits of algorithmic justice measures in machine learning, specifically as it is used regarding racial equality in hiring, admissions, and prison risk assessments. I discuss the examples of these three institutions in detail to show how they demonstrate inherent structural injustice. I argue that the current use and optimism of measuring justice using algorithms is incomplete due to its failures of recognizing the deeper structural injustices. Citing specific ways algorithmic justice is measured, I will analyze how their idealized approaches give a false sense of justice by oversimplifying fairness. To conclude I discuss how a solution to a more just predictive algorithm will have to take into account nonideal theory.

State of Play

As our world becomes increasingly reliant on artificial intelligence, predictive algorithms are becoming more useful in making jobs efficient such as hiring, college admissions, or determining jail time. Machine learning, by nature, looks at existing data, finds patterns, and makes predictions using the patterns it sees. It associates features with outcomes, so then it can see features of new instances and predict outcomes. For instance, the way hiring algorithms work is by looking at a CV or resume and predicting the success of the candidate based on the history. If many existing employees were on a debate team in college, a resume that mentions 'debate' would be more likely to move forward in a machine hiring process that parses through resumes and makes decisions based on patterns. In theory, it seems machine learning would more fairly make decisions due to lack of human implicit bias and subjectivity, but if data is historically biased, machines are not smart enough to fight this on their own.

The problem is that there is bias in the existing data that these machine learning algorithms use. It is no secret that there is bias in employment and hiring practices, along with the other social institutions that utilize machine learning. To give some examples, under 11% of senior leadership positions in computer science-related fields belong to women, and less than 6% of doctors in the U.S. are Black (Hubbert 2023, Howard 2023). Clearly, there is an injustice. As a result, it should be no surprise that algorithms trained on this data tend to perpetuate the existing bias.

In recent years, there has been more research done in algorithmic justice, where scientists use computational strategies to measure the fairness of an algorithm with intentions of combatting this bias. This research varies in its measures of fairness, such as checking for the same demographic proportions before and after the algorithmic processing stage or equalizing

the false positive and false negative rate for different populations. This emerging field of algorithmic justice shows promise as many computer scientists with good intentions are finding ways to quantitatively make fairness judgments about existing algorithms. Yet I believe there is more work to be done.

This paper explores an aspect of bias that is underappreciated in accounts of algorithmic fairness: structural injustice. Basic measures of objectivity or proportionality are not guaranteed to suffice for justice because they implicitly assume that the reason for existing hiring bias is the fault of a wrongdoer. While this is sometimes the case, structural injustice reaches further than morally wrong agents and causes unfairness in social institutions due to action patterns and structures over time, which is an idea formulated by Iris Young in *Responsibility for Justice*. In this paper, I will describe and motivate Iris Young's account of structural injustice and explain how it applies to social institutions that are being taken over by predictive algorithms. I then argue how any approach to algorithmic fairness must take into account structural injustice, with case studies of hiring practices, college admissions, and recidivism. I will explore how the current analyses might work within these institutions and why they fail. I end with a discussion of ideal versus non ideal theory to motivate the idea that current algorithmic fairness measures are flawed in their over-idealized approach to bias.

Iris Young's *Responsibility for Justice*

In *Responsibility for Justice*, Iris Young argues for the existence and relevance of structural injustice in social institutions. She defends the claim that there are deep, structural inequalities that are perpetuated by actions of actors that are not individually blameworthy. Through continuous activity in a normal, socially acceptable manner, somehow injustice results, that is not a matter of simple bad luck or malicious actors. Young's main motivating example is of a poor woman named Sandy, who lives in an apartment building being bought and demolished by a developer who is converting it into condos. Sandy works near the apartment, and there are few other living places close to her workplace. She has no choice but to put in a deposit at a place 45 minutes from her workplace, meaning she also needs to buy a car so she can drive to work. To afford the car, Sandy applies for a housing subsidy and waits as her eviction date draws closer. Sandy faces near homelessness due to her grave situation, and she is stuck in a situation of financial stress. As Young notes, this hypothetical situation is not far from the true economic struggles of poor people in America, as many people work jobs yet struggle to afford rent. This is a clear example of a structural injustice, since it is a moral wrong due to an inherent structure that is "distinct from wrongs traceable to specific individual actions or policies" (Young 2011). Structural injustice plagues people of certain positions, most clearly those of low socioeconomic status, and is not caused by a moral agent's wrongdoing. This kind of injustice is created when people act, within reasonably accepted norms, to pursue their particular goals and interests. We would not say the land developer is blameworthy for Sandy's poverty or that she inflicted her situation upon herself. It is a result of the recursive structures that continue as people act as expected within the structures.

As Young motivates her conception of structural injustice through an example of the housing market, I will use her ideas to justify structural injustice in the areas relevant to predictive algorithms: hiring, college admissions, prison sentences. One way to define structural injustice is an injustice that is "explained by the presence of a structure rather than by an aggregate of individual decisions operating independently" (Sangiovanni 2018). This definition provides a good outlook for the three examples I will work through.

Structural Injustice in Algorithmic Decision-Making Through Case Studies

Case 1: Hiring Practices

It is clear that there is a system of elitism in many hiring practices. Structures of white supremacy, patriarchy, and classism cause certain groups to be subjugated in ways that surpass individual wrongdoers. That is, even if people reading applications and making hiring decisions are not doing anything wrong, due to the interconnectedness of race, gender, socioeconomic status, educational opportunity, and access to connection, there are ways that the hiring process disadvantages certain groups. For example, 10.4% of CEOs of the Fortune 100 companies are women, 3% of engineers in the workforce are Black, 6.4% of business owners are Asian, 2.8% of physicians are black women, and only 2 Latina women have ever been CEO of a Fortune 500 company (Hinchliffe 2023, Carnevale, Smith, and Quinn 2021). The requirements for these elitist job positions are often elite higher education that often includes graduate degrees, which are all very expensive. Ivy League and Ivy-plus students are 60% more likely to reach the top 1% in earnings distribution, the average employment rate for MBA graduates three months after graduation is 86.4%, and all of this education costs hundreds of thousands of dollars, making it not an option for many (Dillon 2023). The disadvantaged groups in the workforce are the same

groups that disproportionately are less likely to afford higher education, with some examples being that 1 out of every 3 Black children and 1 out of every 4 Latino children live in poverty, which is twice the rate of White children (Lin and Harris 2009). Again, even without individual wrongdoers and with the assumption that all agents are reasonable acting within social structures, there is a clear cycle of injustice perpetuated by these systems that impacts hiring.

Case 2: College Admissions

The structural injustice in college admissions works similarly. With the existence of legacy status, which could be said to not be intentionally malicious and instead a self-interested way for universities to maximize their profits, students whose parents attended elite institutions are enormously benefited. There is also the discrepancy due to not only the price of colleges but the price of what is often required to get into elite institutions. 89% of students from wealthy families attend college compared to 51% of students from low-income families (Reber and Smith 2023). SAT tutors, college counselors, and the luxury of free time to partake in impressive extracurriculars without needing to provide for a family, are all privileges of wealthy students that perpetuate these inequalities. These inequalities are also connected with other demographics, namely race. 50% of White adults have college degrees, where only 34% of Black adults and 28% of Hispanic adults do (Barshay 2023). Minorities lacking representation in higher education often view it as futile to apply to college, perpetuating this cycle. Crucially, these inequalities would exist even if no individual was particularly blameworthy, but if all actors acted acceptably within the social norms. As shown, inequalities in both hiring and college admissions are complex, deeply systemic, and require comprehensive understanding of the structural injustice to correctly handle.

Case 3: Prison Risk Assessment

The structural injustices missing in an understanding of recidivism are well-articulated in Michelle Alexander's *The New Jim Crow*. I will not attempt to cover all she does in the book in a paragraph here, but the main idea is the existence of a racial caste system that disproportionately subjugates Black men in the American prison system. Black men may commit crimes at the same rate as White men, yet they are not arrested at the same rates. Additionally, they are more likely to live in impoverished urban areas, be suspended from school, and not have equal resources to reintegrate into society or the workforce after prison. Alexander describes the difference between a crime for a White and Black men, where often crimes by a young White man are "just one mistake" where they can later on go to college or make a successful career (Alexander 2020). If a Black student is convicted of a drug crime, it can mean the end of their academic or professional career. These struggles after prison sentences also lead to disproportionate rates of recidivism. People labeled as felons are banned from jobs, housing, welfare, social services, and voting, making them essentially not a part of the American social sphere. Alexandra notes that while these policies are not overtly racist, the existing inequalities make all these policies lead to large numbers of Black men being excluded from many facets of American life. Clearly the inequalities are deep-rooted in interconnected social systems that make the implications of crime be caused by a lot more than individual responsibility. This is, again, a structural injustice.

The specific algorithmic focus related to prison sentences is the risk assessment process. Risk assessments are often given pretrial to inform jail sentences, determine programming inside jail, and influence parole decisions. The methods for risk assessments differ from that of a trial in that they use "predictive features" of individuals, such as living situation, drug history, mental health history, and other features historically correlated to crime.

The COMPAS algorithm is one of the most widely used in America's criminal justice in assigning risk assessments. A nonprofit called ProPublica found discrepancies in recidivism predicted by COMPAS were heavily racially biased and unreasonable. In one example, two 18-year-old Black girls stole a bike and were convicted of theft for $80 worth of goods, causing COMPAS to label them high risk. In another case, a White 41-year-old man who had a previous armed robbery conviction shoplifted $86 worth of goods for Home Depot and was labeled low risk by COMPAS (Angwin et al. 2016). The COMPAS algorithm uses features correlated with crime and recidivism, such as substance abuse, residence, and social isolation. Tim Brennan, the statistician behind the original algorithm, said himself that it is impossible to ignore qualities correlated with race and without them, accuracy goes down. What Brennan acknowledged was the exact problem with this algorithm: structural racial injustice. Due to the systems described earlier and highlighted in *The New Jim Crow*, recidivism measures and the features associated with them are so highly correlated with race. By using these features to predict recidivism, it will only perpetuate this cycle of unfairness and do nothing to address the moral wrong that is structural injustice. Again in this case, the algorithm itself was not made with obvious racism or bias malition; it merely used data to make predictions as accurately as possible. Yet, this objectivity and predictive accuracy is not enough to make a system just. In this case as with others, the failure to recognize structural injustice will only perpetuate it further.

Current Measures of Algorithmic Fairness

Up to this point, I have motivated that structural injustice exists in three areas that currently use algorithms in decision-making. I have argued why the decision-making algorithms must take into account the existing structural injustice in these institutions. Therefore, any

measure of algorithmic fairness must analyze algorithms with a lens of structural injustice. That is, an algorithm is only fair or just if it considers structural injustice, leaving many of the current metrics by which algorithms are measured incomplete in this way. I will now highlight some of the most prominent algorithmic fairness measures and argue why they are incomplete. To clarify, these algorithmic measures are useful and probably have a place within a more comprehension algorithmic fairness assessment, yet my intention is to point out how they lack accounting for structural injustice.

Measure 1: Counterfactual Equality

One idea for measuring fairness involves "counterfactual equality", which is based on measuring if an individual would have the same outcome had they been in a different demographic group. This is done by selecting groups A and B, and an algorithm is considered fair if people with the same qualifications from different groups end up with the same outcome. In the hiring example, if someone of race X went to NYU and had a 3.7 GPA, they should have the same hiring outcome as someone of race Y who went to NYU and had a 3.7 GPA. Another similar metric is measuring false positives or false negatives and ensuring their equality across demographic groups. For instance, in college admissions, a measure of equalizing false negatives would be ensuring qualified White students have the same chance of rejection as qualified Black students. The idea in these types of fairness measures is that the algorithm is race blind, and does not unfairly act based on one's race.

I argue that this metric fails to account for structural injustice in two ways: intersectionality of demographic groups and a naive race blindness. Regarding the first failure, the intersectionality of demographic groups is a large part of understanding social and racial hierarchies. It is impossible to combat injustice by picking the most obvious groups and

equalizing chances between them, because there are often subgroups and more intersectional complications to demographic divisions. For instance, 2.8% of physicians are Black women, and only 2 Latina women have ever been CEO of a Fortune 500 company (Hinchliffe 2023, Carnevale, Smith, and Quinn 2021). It would be incomplete to achieve counterfactual equality among Black and White applicants, because it misses the additional hierarchies that exist among racial subgroups and other categories such as gender or socioeconomic status. All bias and hiring or jail sentencing discrimination cannot be simplified into two groups, although it is a step in the right direction.

The second, and more substantial problem with a counterfactual assessment of fairness is that it is race-blind. Removing the demographic feature explicitly from the input to the machine is naive and can lead to the same, or worse biases. Due to the correlation of race with many input features such as hometown, socioeconomic status, and education, a machine can guess with decent accuracy what race a person is. As described in *The Ethical Algorithm* by Michael Kearns and Aaron Roth, "No matter what things we [demand] that algorithms ignore in making their decisions, there will always be ways of skirting those preferences by finding and using proxies for the forbidden information."

With the practical infeasibility concern, there is also a moral concern with ignoring racial features. Inputting individuals' details into a machine, excluding race and gender, and assuming there will be a fair output relies on the assumption that unfair outputs were entirely due to human bias. While human bias is a factor, this kind of assumption ignores the existence of structural injustice and acts as if there is no correlation between race and gender and the other features. There needs to be an understanding of the fact that Black men are more likely to drop out of school before being a school dropout can be cause to penalize someone in a prison risk

assessment. Decision making processes such as hiring, admissions, and risk assessments are biased by existing stereotypes, and different traits are, in some part, construed by racial inequalities. The social norm of a "good" applicant for college or for a job aligns with a rich white male, so the actual features prioritized in the decisions could be racially biased even if race is not an explicit feature.

This moral concern is outlined in Charles Mills's *The Racial Contract,* in which he argues that our set of formal and informal social agreements have created a system of racial subordination that places whites over nonwhites. Mills asserts that a raceless approach used in a traditional social contract or conceptualization of justice is dangerous because it "characteristically abstracts away from the things that matter, the actual causal determinants and their requisite theoretical correlates" (Mills 1997). These "causal determinants" from historic and current racial oppression pervade conceptions of favorable and moral traits in humans. Simply removing demographic features from the machine learning model is, therefore, not sufficient to remove structural racism from the model.

Measure 2: Numerical/Proportional Equality

Another way to determine if an algorithm is fair is by looking at the pure number or percentage of different demographics in the outcome. If there are the same amount of hires from race X and race Y, or a percentage deemed fair. Note that it is not unreasonable to have a percentage requirement of a certain race or gender, as it is done frequently with affirmative action today. This approach is not race blind because it takes into account inequalities and reduces them, which shows more promise than previous fairness measures. In practice, a

machine could separately assess individuals based on their group. This would involve the machine learning weights of features differently for different groups.

I argue that this approach is reasonable, but could be incomplete. One reason for its incompleteness is the intersectionality problem. This argument is the same as the one above, which is that there are interconnected demographic groups that make it hard to reduce to an equality among groups. Ensuring that a certain percentage of college admits or job hires belong to group X is a good start if group X is the historically underrepresented group. However, maybe a subgroup X1 of X is entirely excluded from the machine's selection in a way that is unfair. As argued above, intersectionality is a part of structural injustice that is difficult to account for mathematically.

Another reason that the affirmative action approach is insufficient (although it might be necessary) in taking into account structural injustice is that it is inherently a blackbox, so there is much room for biases. Fair outcomes are not enough to justify fairness, and the process must include fairness throughout as well. It is possible that an affirmative action-like algorithmic process could be subject to problematic biases, if features that exist within one race and not another are given disproportionate weight. It could be conceived that something like playing sports is a statistical advantage in college admissions for White students but not for Black students, and it would be problematic to have vastly different qualifications for college based on one's race. Atoosa Kasirzadeh asserts that "the ethical credibility of the algorithm is moot if the algorithm itself does not address the sources of structural injustice, as opposed to merely addressing its effects" (Kasirzadeh 2022). In theory this affirmative action approach to measuring algorithmic fairness is promising as long as it avoids some of the possible pitfalls and is part of a more comprehensive quantitative analysis.

Finding a Solution

A correct measure of algorithmic fairness must find some way to avoid being race-blind, assess the chosen desirable traits for the metric, and have fair outcomes.

First, for reasons outlined earlier in the paper, it is naive to consider individuals in a race-blind way because that is not enough to eliminate biases. There are correlations between features and a person's race (or other demographic) that make a race-blind approach infeasible. Given features like wealth, hometown, and other statistics that do not explicitly include demographic features, there is a high likelihood that demographic features could be predicted by a machine, making the race-blindness useless. That is, even if it is impossible for the algorithm to explicitly use a person's Whiteness as a benefit in a hiring selection, if their combination of class and school attended has a favorable weight, it essentially gives their Whiteness a benefit. It could also very possibly do worse by discrimination to have a race-blind approach, because then there is no reform done to the current discriminatory practices. For instance, as Black women are greatly underrepresented in engineering roles, a demographic-blind approach will not be able to guarantee that more Black women are hired to work towards fixing the current injustices. An algorithm should be considered more fair if it includes demographic features and uses it to contextualize the applicant and work against current disproportionately. Being race-blind is not just naive, but could be actively harmful.

A second metric by which an algorithm's fairness can be judged is by looking at the features given positive and negative weight. Since a predictive algorithm uses patterns to give different weights to features in individuals, it is important that the features weighted are relevant in the right way to the context of the algorithm. For instance, being a first-generation college

applicant should be weighted positively as it should help an applicant get into college, whereas a low GPA can be weighted negatively in college admissions. It makes sense and is fair to give more credit to students in the college application process that do not have college-educated parents to offset the disadvantage it is, whereas it is also reasonable to consider a high school GPA a predictor for college success. It would not be fair if a first-generation college applicant was admitted automatically with no other considerations, just as it would not be fair to disregard all applicants who did not have a perfect GPA. The weight of a feature is important for both the direction it sways the output (typically the options are either positive or negative) and how important it is in the calculation. An algorithmic fairness metric must qualitatively look at the features being considered and whether these features are reasonably weighted. I realize that there is not much specification here and a weight's reasonableness is best determined on a case-by-case basis, but I do believe there is the possibility for common ground on what fair predictive metrics are and how they should be weighted.

The third consideration I will discuss for a measure of algorithmic fairness is in looking at the output of the algorithm. My previous considerations were ways to measure an algorithm's fairness throughout its process, but the outcome of the algorithm is just as important. Even if it seems that the feature weights and algorithmic process is just, if there is an outcome where, for example, a company hires exclusively male employees, there is an injustice. A reasonably proportionate and diverse output of individuals, with some basis in the diversity of the input individuals, is a necessary aspect of a machine learning algorithm's fairness. In looking at the output, some relevant factors to judge include counterfactual fairness, demographic proportionality, and representation of minority groups.

The crucial part of a correct algorithmic fairness measure is that it considers a multitude of aspects of fairness, and not any single one is sufficient. It is possible that what I have outlined above is not sufficient, and there are more necessary considerations. It is also possible that there is no correct measure of algorithmic fairness because the use of the machine makes fairness impossible. I think this is not likely the case if human qualitative judgment is also involved, unless we assume that it is also impossible for humans to make fair decisions. With the combination of quantitative metrics and human assessment of algorithms that takes into account structural injustice, it is likely that there is an ideal way in which algorithmic fairness can be measured and used to improve decision-making algorithms. Although humans can use algorithms and their technology as tools to guide decisions, "the substitution of technology for human is not a perfect substitution" (Martin 2019).

Ideal and Nonideal Theory: Concluding Thoughts

While I am not sure exactly what better algorithms or more apt measures of algorithmic fairness look like, I know it must take into account the state of structural injustice of the institution it makes predictions about. This means that the current state of algorithmic fairness is too ideal, and needs to become less idealized. The distinction between ideal and nonideal theory is crucial in philosophy, as it provides a way to clarify and critique what might be incomplete about a policy. Ideal theory is best used when looking at big picture ways society should run. This includes what a fair setup and justification of government looks like, what an education system should look like, and more large-scale, "end goals" of institutions. The benefit of nonideal theory is to take into account issues in today's world and acknowledge that unless we consider these, we will perpetuate these issues and fail under naivete. Affirmative action is a

good example of a policy justified by nonideal theory that plays into what correct predictive algorithms look like. If our world was racially just and fair, there would be no need to counteract injustice and unfairness using affirmative action. However, if we do not take into account the racial injustice that exists, we will perpetuate the inequalities and systems of white supremacy in higher education. As Charles Mills explains in a critique of Rawls's use of ideal theory, a nonideal approach is necessary to correct current injustices while an ideal theory is less applicable to practical use. He argues that "what is required is the nonideal ideal that starts from the reality of these injustices and then seeks some fair means of correcting for them…" (Mills 2009). If we can agree that there are real systemic injustices that exist and decision-making algorithms should correct for these, then there needs to be a nonideal theoretical approach.

The predictive algorithms discussed in this paper require something similar to affirmative action, to ensure fair outcomes within an unjust society. What has been shown is that there are clear structural injustices in the social realms covered by these algorithms, making them far from ideal. Objective algorithms work in an ideal sense, because they make predictions based on features shown to correlate to certain outcomes. These algorithms, however, fail to consider the reasons for this correlation. Often this is socioeconomic disparities, in the case that attending private school makes a student more likely to attend an elite university, or systemic racism, in the case of common features associated with recidivism. The field of algorithmic justice needs to understand that algorithms ought to handle the world in a nonideal way, and appreciate the underlying injustices as it works towards creating more just outcomes.

The beauty of machine learning is that it has the power to eliminate implicit bias in a way human decisions cannot. If done right and with all the correct considerations, machine learning

could work towards diminishing injustice and as technology becomes increasingly advanced, the potential seems increasingly endless.

Bibliography

Alexander, Michelle, author. *The New Jim Crow : Mass Incarceration in the Age of Colorblindness*. Tenth Anniversary edition. New York : The New Press, 2020.

Angwin, Jilia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 23, 2016.

Barshay, Jill. "College Completion Rates Are up for All Americans, but Racial Gaps Persist." *KQED*, The Hechinger Report, February 20, 2023.

Bauer, William A., and Veljko Dubljević. "AI Assistants and the Paradox of Internal Automaticity." *Neuroethics* 13, no. 3 (October 1, 2020): 303–10. https://doi.org/10.1007/s12152-019-09423-6.

Carnevale, Anthony, Nicole Smith, and Michael Quinn. "Mission Not Accomplished: Unequal Opportunities and Outcomes for Black and Latinx Engineers." *Georgetown University Center on Education and the Workforce*, 2021.

Castro, Clinton, David O'Brien, and Ben Schwan. "Egalitarian Machine Learning." *Res Publica* 29, no. 2 (June 1, 2023): 237–64. https://doi.org/10.1007/s11158-022-09561-4.

Cowls, Josh, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. "The AI Gambit: Leveraging Artificial Intelligence to Combat Climate Change—Opportunities, Challenges, and Recommendations." *AI & SOCIETY* 38, no. 1 (February 1, 2023): 283–307. https://doi.org/10.1007/s00146-021-01294-x.

Dillon, Jonathan. "Ivy-Plus Schools Could Be Perpetuating Economic Inequality." *SSTI*, September 21, 2023.

Drage, Eleanor, and Kerry Mackereth. "Does AI Debias Recruitment? Race, Gender, and AI's 'Eradication of Difference.'" *Philosophy & Technology* 35, no. 4 (October 10, 2022): 89. https://doi.org/10.1007/s13347-022-00543-1.

Himmelreich, Johannes, Désirée Lim, Justin B Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M Hudson, Anton Korinek, et al. *AI and Structural Injustice*. *The Oxford Handbook of AI Governance /*. 1st ed. New York : Oxford University Press, 2022.

Hinchliffe, Emma. "Women CEOs Run 10.4% of Fortune 500 Companies. A Quarter of the 52 Leaders Became CEO in the Last Year." *Fortune*, June 5, 2023, Features: Fortune 500 edition.

Howard, Jacqueline. "Only 5.7% of US Doctors Are Black, and Experts Warn the Shortage Harms Public Health." *CNN*, February 21, 2023.

Hubbert, Jessica. "70+ Women In Technology Statistics (2024)." *Exploding Topics*, September 14, 2023.

Kasirzadeh, Atoosa. "Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy." In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 349–56. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3514094.3534188.

Kearns, Michael, and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. USA: Oxford University Press, Inc., 2019.

Lin, Ann Chih, and David R. Harris. "The Colors of Poverty: Why Racial & Ethnic Disparities Persist." University of Michigan, January 2009. Russell Sage Foundation.

Martin, Kirsten. "Ethical Implications and Accountability of Algorithms." *Journal of Business Ethics* 160, no. 4 (December 1, 2019): 835–50. https://doi.org/10.1007/s10551-018-3921-3.

Mills, Charles W. "Rawls on Race/Race in Rawls." *The Southern Journal of Philosophy* 47, no. S1 (2009): 161–84. https://doi.org/10.1111/j.2041-6962.2009.tb00147.x.

———. *The Racial Contract*. Cornell University Press, 1997. http://www.jstor.org/stable/10.7591/j.ctt5hh1wj.

Rafanelli, Lucia M. “Justice, Injustice, and Artificial Intelligence: Lessons from Political Theory and Philosophy.” *Big Data & Society* 9, no. 1 (2022): 20539517221080676. https://doi.org/10.1177/20539517221080676.

Reber, Sarah, and Ember Smith. “College Enrollment Gaps: How Academic Preparation Influences Opportunity.” *Brookings*, January 23, 2023.

Sangiovanni, Andrea. “Structural Injustice and Individual Responsibility.” *Journal of Social Philosophy* 49, no. 3 (September 1, 2018): 461–83. https://doi.org/10.1111/josp.12250.

Young, Iris Marion. *Responsibility for Justice.* Cary : Oxford University Press, Incorporated, 2011.