Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness

Ben Green, University of Michigan (bzgreen@umich.edu)

Abstract

The burgeoning field of "algorithmic fairness" provides a novel set of methods for reasoning about the fairness of algorithmic predictions and decisions. Yet even as algorithmic fairness has become a prominent component of efforts to enhance equality in domains such public policy, it also faces significant limitations and critiques. The most fundamental issue is the mathematical result known as the "impossibility of fairness" (an incompatibility between mathematical definitions of fairness). Furthermore, many algorithms that satisfy standards of fairness actually exacerbate oppression. These two issues call into question whether algorithmic fairness can play a productive role in the pursuit of equality. In this paper, I diagnose these issues as the product of algorithmic fairness methodology and propose an alternative path forward for the field. The dominant approach of "formal algorithmic fairness" suffers from a fundamental limitation: it relies on a narrow frame of analysis that is limited to specific decision-making processes, in isolation from the context of those decisions. In light of this shortcoming, I draw on theories of substantive equality from law and philosophy to propose an alternative method: "substantive algorithmic fairness." Substantive algorithmic fairness takes a more expansive scope to analyzing fairness, looking beyond specific decision points to account for social hierarchies and the impacts of decisions facilitated by algorithms. As a result, substantive algorithmic fairness suggests reforms that combat oppression and that provide an escape from the impossibility of fairness. Moreover, substantive algorithmic fairness presents a new direction for the field of algorithmic fairness: away from formal mathematical models of "fairness" and towards substantive evaluations of how algorithms can (and cannot) promote equality.

1 Introduction

1.1 *Algorithmic Fairness and Its Discontents*

The burgeoning field of "algorithmic fairness" provides a novel set of methods for reasoning about the fairness of predictions and decisions made by algorithms. Based primarily in computer science, algorithmic fairness involves applying the tools of algorithm design and analysis—in particular, an emphasis on formal mathematical reasoning (Green and Viljoen, 2020)—to questions of fairness. A central component of algorithmic fairness is developing and comparing mathematical definitions of fair decision-making (Barocas et al., 2019). Other lines of research include optimizing algorithms for various definitions of algorithmic fairness (Feldman et al., 2015; Hardt et al., 2016) and auditing algorithms to test for biases (Angwin et al., 2016; Obermeyer et al., 2019; Raji and Buolamwini, 2019).

In the face of widespread concerns about discriminatory decision-making, algorithmic fairness has become a prominent component of efforts to enhance equality. For instance, government agencies have adopted machine learning algorithms to improve the fairness of decision-making in pretrial adjudication, policing, child welfare, and unemployment. To policymakers, policy advocates, and scholars across multiple fields, algorithms overcome the cognitive limits and social biases of human decision-makers, leading to more objective and fair decisions (Arnold Ventures, 2019; Harris and Paul, 2017; Kleinberg et al., 2019; Miller, 2018; Sunstein, 2019). Furthermore, many proponents see algorithmic fairness as providing a rigorous method for reasoning about what fairness entails and how to achieve it in practice (Kleinberg et al., 2019; Sunstein, 2019). As a result, many policymakers and advocates praise algorithms as central and critical tools for reforming decision-making processes rife with discrimination (Arnold Ventures, 2019; Eubanks, 2018; Harris and Paul, 2017; Porrino, 2017).

Yet alongside its rapid rise in prominence, algorithmic fairness has run up against significant limitations and critiques. The most fundamental issue facing algorithmic fairness is the mathematical result known as the "impossibility of fairness." This result reveals a fundamental tension between two different mathematical definitions of fair decision-making. "Separation" requires that different groups receive the same false positive and false negative rates. "Sufficiency" requires that predictions have the same meaning for different groups. The impossibility of fairness proves that it is impossible for an algorithm to satisfy both separation and sufficiency: an algorithm that is fair along one standard will inevitably be unfair along the other standard (Chouldechova, 2017; Kleinberg et al., 2016).[1]

The impossibility of fairness presents a significant practical limit on efforts to promote fairness using algorithms. While neither separation nor sufficiency fully encapsulates the philosophical notion of fairness (Binns, 2018; Green and Hu, 2018; Jacobs and Wallach, 2021; Selbst et al., 2019), both capture a normatively desirable principle. It is therefore of significant concern if all algorithms are guaranteed to violate at least one of these fairness definitions. For instance, a pretrial risk assessment will either misclassify Black and white defendants as recidivists at different rates (violating separation) or will yield different predictions for Black and white defendants who are equally likely to recidivate (violating sufficiency). As one article about algorithmic fairness concludes, "the tradeoff between [these] two different kinds of fairness has real bite" and means that "total fairness cannot be achieved" (Berk et al., 2018).

Furthermore, many scholars and activists have critiqued algorithmic fairness for enabling narrow and superficial reforms. Algorithmic fairness focuses on bad actors, individual axes of disadvantage, and a limited set of goods, thus "mirroring some of antidiscrimination discourse's most problematic tendencies" as a mechanism for achieving equality (Hoffmann, 2019). When deployed in practice, algorithms that satisfy standards of fairness often therefore reproduce inequities and legitimize unjust institutions and policies (Davis et al., 2021; Green, 2020; Kalluri, 2020; Ochigame, 2020; Ochigame et al., 2018; Powles and Nissenbaum, 2018).

These two issues—the impossibility of fairness and the concerns about narrow reforms—call into question whether algorithms and algorithmic fairness can play productive roles in the pursuit of equality. In fact, some scholars have called for rejecting the frame of "fairness" in favor of "justice" (Bui and Noble, 2020; Green, 2018), "equity" (D'Ignazio and Klein, 2020), and "reparation" (Davis et al., 2021). However, it is not clear what these alternative paths forward entail for the field of algorithmic fairness. Is there a role for algorithmic fairness to promote equality without confronting the impossibility of fairness and without reproducing inequities? What would this field entail?

[1] I will provide more detail on these fairness definitions in Section 2. The only exceptions to the impossibility of fairness involve two unlikely scenarios: when the algorithm perfectly predicts every outcome or when the groups in question exhibit the outcome being predicted at the same "base rate" (Kleinberg et al., 2016).

1.2 Article Overview: Methodological Reform

In order to develop a positive agenda for algorithmic fairness, we must first understand why the current approach falls short, suggesting narrow reforms that face an intractable impossibility of fairness. In light of these limitations, this article interrogates the methodology of algorithmic fairness. A methodology is "a body of methods, rules, and postulates employed by a discipline" (Merriam-Webster, 2021). A methodology provides a language for comprehending and reasoning about the world, shaping how its practitioners conceive of problems and develop solutions to those problems. Algorithmic reasoning embodies a particular methodology grounded in mathematical formalism, with both strengths and weaknesses (Green and Viljoen, 2020).

When attempting to address any problem, it is necessary to possess the proper methodology. As philosopher John Dewey writes, "[t]he way in which [a] problem is conceived decides what specific suggestions are entertained and which are dismissed" (Dewey, 1938). When a problem is badly conceived, it "cause[s] subsequent inquiry to be irrelevant or to go astray;" the remedy is to reformulate the problem (Dewey, 1938). When it comes to combatting injustices, problem formulation has significant normative stakes. As philosopher Elizabeth Anderson describes, "Sound political theories must be capable of representing normatively relevant political facts. If they can't represent certain injustices, then they can't help us identify them. If they can't represent the causes of certain injustices, then they can't help us identify solutions" (Anderson, 2009).

Following Dewey and Anderson, developing a better path forward for algorithmic fairness requires reforming its methodology. I argue that algorithmic fairness suffers from a significant methodological limitation: it relies on a narrow frame of analysis that is restricted to specific decision points, in isolation from the context of those decisions.[2] I call this method "formal algorithmic fairness," as it is closely aligned with formal approaches to equality (which emphasize equal treatment for individuals based on their attributes or behavior at a particular decision point). Because of the narrow way in which formal algorithmic fairness is conceived, it cannot represent many of the injustices associated with algorithmic decision-making. As a result, formal algorithmic fairness suggests a misguided reform strategy: enhance fairness by optimizing decision-making procedures with algorithms. These algorithmic interventions fail to reduce (and often even exacerbate) oppression, despite appearing to be fair. Furthermore, these reforms confront the impossibility of fairness, which significantly limits the extent to which fairness can be achieved at all.

Drawing on theories of substantive equality from law and philosophy, I propose an alternative method to formal algorithmic fairness, which I call "substantive algorithmic fairness." Substantive algorithmic fairness expands the frame of analysis beyond isolated decision points, evaluating fairness in light of the social hierarchies represented by data and the impacts of decisions facilitated by algorithms. The goal is not to incorporate these considerations into a formal mathematical model, a strategy which would fail to provide the necessary methodological shift (Green and Viljoen, 2020). Substantive algorithmic fairness is not a method for creating "substantively fair algorithms." Instead, substantive algorithmic fairness embodies an "algorithmic realist" approach (Green and Viljoen, 2020), incorporating modes of reasoning from law and philosophy to inform

[2] By decision points, I refer to the specific moments in which decisions are made about individuals. Examples include decisions about whether to release or detain pretrial defendants and decisions about whether to admit or reject college applicants.

efforts to enhance equality using algorithms. Substantive algorithmic fairness is a method that applies the framework of substantive equality to rigorously reason about how algorithms should (and should not) be used to promote a more just society.

Because of its broader scope of analysis than formal algorithmic fairness, substantive algorithmic fairness suggests more effective paths for pursuing equality. This approach to reform reveals that the impossibility of fairness is not actually intractable: when reforms confront the impossibility of fairness, it suggests not that fairness is impossible, but that an algorithm is being used to achieve the wrong reform. Instead of attempting merely to optimize decision-making procedures with algorithms, substantive algorithmic fairness calls for a) alleviating social hierarchies and b) reducing the scope and stakes of decisions that exacerbate those hierarchies. These reforms ameliorate the relational and structural sources of oppression that produce the impossibility of fairness, thus providing an escape from this dilemma. In sum, substantive algorithmic fairness presents a new direction for the field of algorithmic fairness: away from formal mathematical models of "fairness" and towards substantive evaluations of how algorithms can (and cannot) promote equality.

2 The Impossibility of Fairness

The impossibility of fairness centers on one of the most controversial domains in which algorithms have been used to promote fairer public policy: US pretrial adjudication.[3] In May 2016, investigative journalists at ProPublica reported that a risk assessment algorithm used to judge pretrial defendants in Broward County, Florida was "biased against blacks" (Angwin et al., 2016). This algorithm, known as COMPAS, was created by the company Northpointe and is used by many court systems across the United States.[4] ProPublica found that, among defendants who were not arrested in the two years after being evaluated, Black defendants were 1.9 times more likely than white defendants to be misclassified by COMPAS as "high risk" (i.e., subjected to false positive predictions).

Tech critics responded to ProPublica's article with outrage about racist algorithms (Doctorow, 2016; O'Neil, 2016). However, others challenged ProPublica's methods and conclusions. Northpointe and numerous academics defended COMPAS, arguing that ProPublica had focused on the wrong measure of algorithmic bias (Corbett-Davies et al., 2017; Dieterich et al., 2016; Flores et al., 2016; Gong, 2016). These groups argued that the proper measure of fairness is not whether false positive (and false negative) rates are the same for each race, but whether predicted risk scores imply the same probability of recidivism for each race. COMPAS satisfied this notion of fairness, suggesting that the tool was fair.

[3] I use the setting of pretrial risk assessments as a case study throughout the article, as it is central to debates about algorithmic fairness. After someone is arrested and before they face trial, courts must decide whether to detain the defendant in jail until their trial or release them with a mandate to return for their trial. In light of concerns about the traditional policy of determining release using money (i.e., "cash bail") and the biases of human decision-makers, many jurisdictions have turned to pretrial risk assessments. These algorithms predict the likelihood that pretrial defendants will be arrested in the future or will fail to appear for trial; these predictions are presented to judges to inform their decisions about whether to release or detain each defendant.

[4] COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions. Northpointe has since been renamed Equivant.

This debate about whether COMPAS is biased concerns two distinct definitions of algorithmic fairness. The first is "separation," which is satisfied if all groups subject to an algorithm's predictions experience the same false positive rate and the same false negative rate.[5] Separation expresses the idea that people who exhibit the same outcome should receive the same prediction. ProPublica argued that COMPAS is biased because it violates separation: Black non-recidivists are more likely to be labeled "high risk" than white non-recidivists (Angwin et al., 2016).

The second notion of algorithmic fairness is "sufficiency," which is satisfied if, among those who receive a particular prediction, all groups exhibit the outcome being predicted at the same rate. Sufficiency expresses the idea that individuals who have a similar likelihood to exhibit the particular behavior of interest should be treated similarly.[6] Northpointe and others argued that COMPAS is fair because it satisfies sufficiency: the label of "high risk" signifies a similar probability of rearrest for both Black and white defendants (Corbett-Davies et al., 2017; Dieterich et al., 2016; Flores et al., 2016; Gong, 2016). Sufficiency is the most widely used notion of fairness in both research and practice, particularly as machine learning models typically satisfy this principle by default (Barocas et al., 2019).

These competing claims about fairness raised a fundamental question about algorithmic fairness: is it possible for an algorithm to simultaneously satisfy both separation and sufficiency? As several groups of computer scientists soon discovered, the answer is no: there is an inevitable tension between these definitions of fairness (Angwin and Larson, 2016; Barocas et al., 2019; Chouldechova, 2017; Kleinberg et al., 2016). This result is known as the "impossibility of fairness." The only way an algorithm could satisfy both separation and sufficiency is if a) the algorithm makes predictions with perfect accuracy or b) the groups in question exhibit the outcome of interest at the same "base rate" (Kleinberg et al., 2016). These exceptions are exceedingly rare in practice: algorithms never operate at perfect accuracy on real-world problems, and social stratification means that socially salient groups exhibit many outcomes at different rates.[7]

The impossibility of fairness has been taken to reflect a harsh and intractable dilemma facing any efforts to promote equality using algorithms (Berk et al., 2018). Work on algorithmic fairness operates against the backdrop of this dilemma. As I will describe below, responses to the impossibility of fairness typically take one of two forms. First, many computer scientists argue that we should focus only on sufficiency as the proper definition of fairness. Second, other computer scientists argue that we should use the formalism of algorithmic reasoning to consider the tradeoffs between competing notions of fairness. In order to evaluate these responses to the impossibility of fairness—and to inform more productive responses—I turn now to egalitarian theories that provide suggestions for how to escape from similar seemingly intractable dilemmas between notions of equality.

[5] Separation is aligned with fairness criteria such as error rate balance and balance for the positive/negative class.

[6] Sufficiency is aligned with fairness criteria such as calibration and predictive parity.

[7] This is true along lines of race, gender, class, and more, and for outcomes such as crime, educational attainment, and wealth. As I describe below, these disparities exist above and beyond biased data collection processes and are the result of oppression.

3 Lessons from Egalitarian Theory

3.1 Formal and Substantive Equality

Egalitarianism is a school of thought in political philosophy that emphasizes equality between individuals. Broadly speaking, "Egalitarian doctrines tend to rest on a background idea that all human persons are equal in fundamental worth or moral status" (Arneson, 2013). However, because there are many ways of defining equality, egalitarianism takes numerous, often conflicting, forms (Arneson, 2013).

A central tension in egalitarian theory is between "formal" and "substantive" equality. Formal equality defines equality as equal treatment or equal process at a particular moment in time, with everyone judged according to the same standard (Fishkin, 2014; MacKinnon, 2011). It stipulates that similar people should be treated similarly. In the United States, disparate treatment law is grounded in notions of formal equality, attempting to ensure that people are not mistreated on the basis of protected attributes such as race or gender. Formal equality aligns with foundational principles of equality and presents a clear standard for decision-makers. However, although a formal approach may be sufficient in an equitable society, it "would make no sense at all in a society in which identifiable groups had actually been treated differently historically and in which the effects of this difference in treatment continued into the present" (Crenshaw, 1988). For instance, because of racial inequalities in educational opportunities (EdBuild, 2019), evaluating all students for college admissions according to the same standard would perpetuate racial injustice.

Substantive equality takes a more expansive scope to analyzing equality. Rather than striving to treat each person similarly, substantive equality accounts for disadvantage and aims to redress marginalization (Fredman, 2016). Legal scholar Catharine MacKinnon describes substantive equality as opposition to social hierarchies, which she defines as "social relation[s] of rank ordering, typically on a group or categorical basis," that lead to both material and dignitary inequalities (MacKinnon, 2011). In other words, "hierarchy identifies the substance of substantive equality" (MacKinnon, 2016). In the United States, disparate impact law is grounded in notions of substantive equality (albeit partially (MacKinnon, 2011; MacKinnon, 2016)), attempting to ensure that formally neutral rules do not disproportionately burden protected groups. More broadly, substantive equality envisions a world free from social hierarchy (MacKinnon, 2011; MacKinnon, 2016).

Formal and substantive equality derive from distinct methods for analyzing equality, leading to divergent agendas for policy reform. The formal approach to equality relies on a narrow frame of analysis that focuses on specific decisions at specific points in time. When confronted with instances of inequality, the formal approach considers only the outcomes within the scope of a single decision point. As a result, formal equality prescribes that people should be treated similarly based solely on their qualifications for the given decision. In contrast, the substantive approach to equality requires a broader frame of analysis that considers social hierarchies and institutional structures. When confronted with instances of inequality, "[a] substantive equality approach […] begins by asking, what is the substance of this particular inequality, and are these facts an instance of that substance?", emphasizing that "it is the hierarchy itself that defines the core inequality problem" (MacKinnon, 2011). As a result, substantive equality suggests reforms that account for injustices that exist beyond particular decision points. By taking a more expansive view that accounts for social hierarchies, substantive equality is better equipped to promote equal societies in the face of existing social hierarchies.

3.2 Substantive Approaches to Escaping Intractable Equality Dilemmas

A particular benefit of substantive equality is that it provides conceptual tools for escaping from seemingly intractable dilemmas between competing notions of fairness. Just as algorithmic fairness confronts the impossibility of fairness, egalitarian theorists have confronted seemingly unresolvable tensions between notions of equality. In order to inform our understanding of the impossibility of fairness, I turn to three complementary egalitarian approaches for analyzing and escaping such dilemmas: philosopher Elizabeth Anderson's theory of "democratic equality" (Anderson, 1999), legal scholar Martha Minow's "social-relations approach" to managing social differences (Minow, 1991), and legal scholar Joseph Fishkin's theory of "opportunity pluralism" (Fishkin, 2014).[8]

Anderson, Minow, and Fishkin each confront a seemingly intractable dilemma that arises within common approaches to enhancing equality.

- Anderson responds to a "dilemma" that arises in luck egalitarianism (a view that advocates compensating people for inequalities that result from misfortunate but not inequalities that result from choice) (Anderson, 1999). On the one hand, not providing aid to the disadvantaged means blaming individuals for their misfortune. On the other hand, providing special treatment to individuals on account of their inferiority means expressing contempt for the disadvantaged.
- Minow engages with the "dilemma of difference" that arises in legal efforts to deal with differences between individuals (Minow, 1991). On the one hand, giving similar treatment to everyone regardless of their circumstances can "freeze in place the past consequences of differences." On the other hand, giving special treatment to those deemed "different" risks further entrenching and stigmatizing that difference.
- Fishkin addresses the "zero-sum struggles" that arise in efforts to promote equal opportunity (Fishkin, 2014). The "formal equal opportunity" approach of judging people for an opportunity based on their performance or attributes at a particular moment in time perpetuates inequalities in different groups' development opportunities and life chances. Yet because it is impossible to create a truly level playing field, even approaches that attempt to account for existing inequalities (such as Rawlsian equal opportunity and luck egalitarianism) fall short and prompt "extraordinarily contentious" debates.

The equality dilemmas presented by Anderson, Minow, and Fishkin all resemble the impossibility of fairness: efforts to promote equality are impaired by a seemingly inevitable zero-sum tradeoff between two competing notions of fairness. However, each scholar reveals that their dilemma is not intractable. Instead, each dilemma only appears that way due to a narrow approach to reasoning about equality. Expanding the frame of analysis clarifies the problems of inequality and suggests two paths to escape these equality dilemmas.

[8] Each of these approaches are aligned with relational egalitarianism, which asserts that "people should relate to one another as equals or should enjoy the same fundamental status" (Arneson, 2013). Although Fishkin is the least explicitly focused on relationships, his analysis has strong overlaps with relational egalitarianism.

First, an expanded analysis highlights how social hierarchies lead to equality dilemmas. Noting that that the goal of egalitarianism is “to end oppression, which by definition is socially imposed,” Anderson expands the analysis of equality from distributions (of both tangible and intangible goods) to social relations (Anderson, 1999). From this perspective, the problem of inequality is not merely that some people have more of a particular good than others. A broader problem is that society imposes disadvantages on individuals who lack certain attributes or abilities (Anderson, 1999; Minow, 1991). Without social hierarchies, real or perceived differences between individuals would not lead to different levels of rights or capacities, which in turn would prevent a dilemma between treating everyone the same and providing special treatment. For instance, the injustice faced by someone who is stigmatized because of their physical appearance is not that they are inherently ugly. Instead, “the injustice lies […] in the social fact that people shun others on account of their appearance” (Anderson, 1999). Oppressive social norms turn a superficial difference between people into one marked by severe disparities in status. As a result, treating everyone the same would leave “ugly” individuals in a subordinate position. However, a remedy such as subsidizing plastic surgery for “ugly” individuals would uphold oppressive beauty norms even if it provides aid for some people.

Following this reorientation toward relationships, the first approach to escaping equality dilemmas is what I call the “relational response”: reform institutions and social norms to reduce social hierarchies. While ignoring the context of social differences “make[s] the difference dilemma seem intractable,” questioning social arrangements makes the dilemma “less paralyzing” (Minow, 1991). Recognizing social categories as relational (rather than intrinsic to individuals) and social arrangements as political and mutable (rather than neutral and static) introduces reforms that “escape or transcend the dilemmas of difference” (Minow, 1991). In other words, the primary task of reform should not be providing special treatment to “different” individuals, but reducing the extent to which superficial differences lead to meaningful disparities in status and abilities (Minow, 1991). In the case of someone who is stigmatized because of their appearance, Anderson suggests altering social norms so that no one is shunned or treated as a second-class citizen due to their appearance. If one’s appearance has no relationship to their social status, appearance ceases to be a normatively relevant category, such that there is no dilemma between treating people similarly or different based on how they look. Such reforms may be difficult to achieve (at least in the immediate term), thus necessitating a more individualized remedy such as plastic surgery. Nonetheless, this approach “lets us see how injustices may be better remedied by changing social norms and the structure of public goods than by redistributing resources” (Anderson, 1999).

Second, an expanded analysis highlights how the structure of decisions exacerbates social hierarchies and raises the stakes of equality dilemmas. Fishkin expands the focus from individual competitions to the broader structure of opportunities. From this perspective, the problem of inequality is not merely that groups face vastly different development opportunities, making it impossible to create fair contests between all individuals. A broader problem is that opportunities are structured around a small number of “zero-sum, high-stakes competitions,” which Fishkin calls “bottlenecks” (Fishkin, 2014). These competitions typically hinge on attributes that are unequally distributed across social groups, compounding existing disadvantage (i.e., oppressed groups are less qualified to succeed in competitions for beneficial opportunities, such as jobs). Without these bottlenecks, decisions would not as strongly magnify existing inequalities, which in turn would lower of stakes of the dilemma between treating everyone the same and providing special treatment. For instance, debates about admission to elite US colleges and universities are contentious not only because of inequities in educational resources, but also because admission provides a rare pathway to a high level of social status and material comfort. The significance of college admissions decisions makes disparities in primary and secondary education particularly consequential for determining future life outcomes. As a result, evaluating all students according to the same standard would entrench inequalities in primary and secondary education. However, attempts to promote equality through affirmative action are inevitably zero-sum and leave the bottleneck in place.

Following this reorientation toward the structure of decisions, the second approach to escaping equality dilemmas is what I call the "structural response": reduce the scope and stakes of decisions that exacerbate social hierarchies. Fishkin suggests that, "Instead of taking the structure of opportunities as essentially given and focusing on questions of how to prepare and select individuals for the slots within that structure in a fair way, [we should] renovate the structure [of opportunities] itself" (Fishkin, 2014). Rearranging the structure of opportunities can diminish the significance of bottlenecks, limiting the extent to which high-stakes decisions hinge on attributes that are unevenly distributed across social groups. In the case of college admissions, the structural response suggests lowering the stakes of college admissions decisions. If college admissions are less determinative of future life outcomes, the injustices that result from disparities in early educational opportunities would be reduced, making the dilemma between treating students similarly or different based on their academic performance less troubling. There are two primary strategies for achieving this goal. First, help people through the bottleneck, making college admission more accessible for disadvantaged individuals. Second, and more importantly, help people around the bottleneck, creating more paths for people to lead comfortable and fulfilling lives without a college degree. Providing alternative pathways outside of existing bottlenecks would make inequities in primary and secondary education less consequential for determining future life outcomes, thus mitigating the downstream harms of this disparity.

In sum, substantive equality draws attention to the two sources of equality dilemmas: social hierarchies and decisions that magnify the stakes of those hierarchies. If there were no social hierarchies, or if consequential decisions did not exacerbate social hierarchies, then equality dilemmas would not arise (or, at the very least, would not be so concerning). By making these relational and structural factors legible, substantive equality introduces agendas for reform that provide an escape from these seemingly intractable dilemmas. Rather than focus only on altering the distribution of a particular good or opportunity, it is necessary to ameliorate the underlying social hierarchies and to reform the policy structures that magnify the downstream consequences of these hierarchies. These reforms are not beholden to equality dilemmas. In fact, to the extent that they are successful, these reforms reduce the extent to which equality dilemmas arise at all. The following two sections will consider the implications of these lessons for algorithmic fairness.

4 Formal Algorithmic Fairness: Navigating the Impossibility of Fairness

In this section, I return to the impossibility of fairness, characterizing the attributes and limits of the dominant approach to algorithmic fairness, which I call "formal algorithmic fairness." Akin to formal approaches to equality, formal algorithmic fairness is an approach to algorithmic fairness in which analysis is limited to the functioning of algorithms at particular decision points. In other words, when confronted with concerns about a discriminatory decision-making process, formal algorithmic fairness considers only the inputs and outputs of the decision point in question. Following this logic, formal algorithmic fairness suggests a reform strategy of improving the fairness of specific decision-making procedures using algorithms. Although this approach is common, efforts to achieve fairness in this manner entrench inequality and are unavoidably impaired by the impossibility of fairness.

In order to elucidate the limits of formal algorithmic fairness, I interrogate this method's two responses to the impossibility of fairness. These responses accept the impossibility of fairness, attempting to navigate the tradeoffs imposed by this dilemma. Interrogating these responses through the lens of substantive equality reveals how the narrow frame of formal algorithmic fairness fundamentally restricts its ability to redress social hierarchy. Responses that look appealing within the frame of formal algorithmic fairness leave reforms stuck within the impossibility of fairness and often actually reproduce injustice. In other words, the central problem facing algorithmic fairness is not that data is often biased or that we lack the appropriate technical tools. The problem is formal algorithmic fairness itself.

4.1 The Formal Equality Response: Reproducing Inequity

The first response to the impossibility of fairness is what I call the "formal equality response." This response defends sufficiency as the proper instantiation of algorithmic fairness.[9] In other words, an algorithm is fair as long as it satisfies sufficiency. On this view, as long as sufficiency is satisfied, any lack of separation is acceptable—the inevitable byproduct of groups exhibiting the outcome in question at different rates. This response adheres to the logic of formal equality, asserting that fairness entails treating people the same based solely on each person's likelihood to exhibit a given outcome.

Most critiques of ProPublica's COMPAS report followed the formal equality response, asserting that ProPublica focused on the wrong definition of fairness (Corbett-Davies et al., 2017; Dieterich et al., 2016; Flores et al., 2016; Gong, 2016). These respondents argued that COMPAS is fair because it satisfies sufficiency (i.e., each COMPAS score implies a similar likelihood of being arrested for both Black and white defendants). COMPAS produces a higher false positive rate for Black defendants simply because Black defendants are more likely to recidivate, not because COMPAS is racially biased. Most notably, Northpointe emphasized that the violation of separation presented by ProPublica "does *not* show evidence of bias, but rather is a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores" (Dieterich et al., 2016). Another critique similarly described, "Given the higher observed recidivism rates for Black defendants, […] it is nothing short of logical that these defendants evidence higher COMPAS scores (after all, isn't that precisely what the COMPAS is measuring?)" (Flores et al., 2016).

Within a formal frame of analysis, the formal equality response seems appropriate. If the goal of a decision-making procedure is to differentiate people based on their likelihood for a given outcome, then it seems fair to make decisions based on those probabilities. For instance, if a Black and a white defendant are equally likely to be arrested in the future, then they should be given the

[9] Recall that sufficiency expresses the idea that individuals who have a similar likelihood to exhibit a particular behavior should be treated similarly. In contrast, separation expresses the idea that people who exhibit the same outcome should receive the same prediction.

same risk label. Under this logic, the best way to advance algorithmic fairness is to increase prediction accuracy and thereby ensure that decisions are based on accurate judgments about each individual (Hellman, 2020; Kleinberg et al., 2019). In this sense, the formal equality response aligns with a common formal strategy for promoting equality: seek more accurate evaluations of individual merit or behavior.[10]

However, considering the formal equality response through a substantive lens demonstrates the limits of this response. First, the formal equality response fails to consider whether group differences in outcome rates reflect social hierarchy. In the case of risk assessments, Black and white defendants do not just "happen to have different distributions of scores," as adherents of sufficiency assert (Dieterich et al., 2016). Instead, past and present discrimination has created social conditions in the US in which Black people are empirically at higher risk to commit crimes, above and beyond racial disparities in arrest and enforcement patterns (Cooper and Smith, 2011; Sampson et al., 2005).[11] This disparity results from social oppression rather than from differences in inherent criminality (Muhammad, 2011). For instance, discriminatory practices such as redlining and segregation (Rothstein, 2017), racial criminalization (Muhammad, 2011), and severe underfunding of schools (EdBuild, 2019) all increase crime (Krivo et al., 2009; Lochner and Moretti, 2004; Rose and Clear, 1998).

Second, the formal equality response ignores the consequences of the actions taken in response to an algorithm's advice. When a risk assessment labels a defendant "high risk," that person is likely to be detained in jail until their trial. This policy of detaining defendants due to their crime risk, known as "preventative detention," is both controversial and harmful. When the United States Supreme Court deemed preventative detention constitutional in 1987, Justice Thurgood Marshall declared the practice "incompatible with the fundamental human rights protected by our Constitution" (U.S. Supreme Court, 1987). Preventative detention has faced continued scrutiny for undermining the rights of the accused and exacerbating mass incarceration (Baradaran, 2011; Koepke and Robinson, 2018). Pretrial detention imposes severe costs on defendants, including the loss of freedom, an increased likelihood of conviction, and a reduction in future employment (Dobbie et al., 2018).

By failing to account for these dimensions of pretrial decision-making, the formal equality response suggests a reform strategy in which even the best-case scenario—a perfectly accurate risk assessment—would perpetuate racial inequity.[12] The central injustice of risk assessments is not that flawed data might lead an algorithm to make erroneous predictions of someone's crime risk, but that racial stratification makes Black defendants higher risk than white ones and that the consequences of being deemed high risk include the loss of liberty. Because Black defendants recidivate at higher rates than white defendants (Cooper and Smith, 2011; Flores et al., 2016; Larson et al., 2016; Sampson et al., 2005), a perfect risk assessment will correctly label a higher proportion of Black defendants as "high risk." In other words, if data is collected about an unequal

[10] Minow calls this strategy the "equal rights" approach (Minow, 1991) and Fishkin calls it the "formal-plus" approach (Fishkin, 2014).

[11] Measurement bias is typically present in crime datasets but is not the only source of racial disparities in crime rates. For a broader discussion of the relationship between measurement and algorithmic fairness, see (Jacobs and Wallach, 2021).

[12] Because this risk assessment makes perfect predictions, it would satisfy both sufficiency and separation (Kleinberg et al., 2016).

society, then an accurate algorithm trained on that data will reproduce those unequal conditions. To the extent that these predictions direct pretrial decisions, this risk assessment would lead to a higher pretrial detention rate for Black defendants than white defendants, in effect punishing Black communities for having been subjected to criminogenic circumstances. Thus, although a perfect risk assessment may help some Black defendants who are low risk but could be stereotyped as high risk, it would also naturalize the fact that many Black defendants actually are high risk and become incarcerated as a result.

This substantive analysis also sheds light on why the impossibility of fairness presents such a troubling dilemma. The issue is not merely that different groups receive different distributions of outcomes. Instead, the issue arises through the combination of social hierarchy and decisions that exacerbate the hierarchy. Without the oppression that creates group disparities in outcome rates, the impossibility of fairness would not arise, making it possible to satisfy both separation and sufficiency. Furthermore, without policies that harm individuals who express high probabilities for an undesired outcome (or low probabilities for a desired outcome), group differences in outcome rates would not lead to further injustice. In sum, the impossibility of fairness appears particularly stark and distressing when an oppressed group disproportionately exhibits the attributes associated with receiving high-stakes, negative decisions (i.e., receiving punishment or not receiving benefit). Thus, when a pretrial risk assessment satisfies sufficiency (as most do), the result is that Black defendants disproportionately receive the additional injustices of higher pretrial detention rates.

4.2 The Formalism Response: Constraining Reform

The second response to the impossibility of fairness that arises out of formal algorithmic fairness is what I call the "formalism response." Compared to the formal equality response, the formalism response takes a more measured view of the dilemma. Recognizing that sufficiency is an imperfect definition of fairness, the formalism response does not require adherence to this measure. Instead, the formalism response focuses on analyzing the tradeoffs between notions of fairness. In particular, the formalism response suggests using the explicit mathematical formalization required by algorithms to rigorously consider the tradeoffs between separation and sufficiency in any given context.[13]

Under the formalism response, the formalism of algorithms provides a reality check by revealing the difficult tradeoffs between notions of fairness that might otherwise remain opaque and unarticulated (Barocas et al., 2019; Berk et al., 2018; Ligett, 2021). Algorithms therefore provide "clarity" to help us identify and manage the unavoidable tradeoffs between competing goals (Kleinberg et al., 2019; Sunstein, 2019). In fact, proponents of this view argue that algorithms can "be a positive force for social justice" because they "let us precisely *quantify tradeoffs* among society's different goals" and "force us to make more explicit judgments about underlying principles" (Kleinberg et al., 2019).

As with the formal equality response, the formalism response appears appropriate within a formal frame of analysis limited to specific decision points. If our interventions are limited to reforming

[13] The formalism response is inclusive of the formal equality response: it is possible, after considering the tradeoffs, to determine that an algorithm should be optimized for sufficiency. The formalism response can also account for other tradeoffs, such as the tension between a given fairness metric and overall accuracy.

specific decision-making procedures, then it is desirable to understand the practical tradeoffs presented by different fairness metrics rather than leave these tradeoffs obscured. For instance, given an existing population of Black and white defendants, it is beneficial to have clarity on how tuning a risk assessment for one notion of fairness will cause the algorithm to violate another notion of fairness. Under this logic, the best way to advance algorithmic fairness is to precisely balance sufficiency and separation based on the particular context at hand.

However, considering the formalism response through a substantive lens demonstrates the limits of this response. First, the formalism response leaves us stuck making zero-sum choices between two highly limited notions of fairness. Although separation may appear to be a desirable alternative to sufficiency, separation also fails to account for subordination. In the case of risk assessments, separation entails having different thresholds for Black and white defendants (e.g., a higher risk threshold for labeling Black defendants "high risk"). This practice would seem to obviate the point of using algorithmic risk predictions at all, as risk scores would have different meanings based on the defendant in question (Flores et al., 2016; Mayson, 2019). Such explicit differential treatment based on race would be illegal to implement in many instances (Corbett-Davies et al., 2017; Hellman, 2020). Furthermore, although a lack of separation demonstrates that different groups face disparate burdens from mistaken judgments (Chouldechova, 2017; Hellman, 2020), separation cannot capture the injustices associated with accurate predictions. As demonstrated by the perfect pretrial risk assessment described in Section 4.1, an algorithm can satisfy separation while still reproducing racial hierarchy.

Second, the formalism response suggests a reform strategy that is incredibly constrained and largely ineffective. By restricting analysis to isolated decision points, the formalism response provides "clarity" regarding the tradeoffs involved in promoting fairness, but only within the narrow scope of specific decision-making procedures. Everything beyond this scope is treated as static and thus not a relevant site for scrutiny or reform. For instance, research on fairness in risk assessments explicitly places structural disadvantage and existing racial disparities outside the scope of algorithms and the responsibility of their designers (Chouldechova, 2017; Corbett-Davies et al., 2017; Kleinberg et al., 2019). As a result, arguments following the formalism response assume that implementing an algorithm is the only possible (or, at least, pertinent) alternative to the status quo (Berk et al., 2018; Kleinberg et al., 2019; Miller, 2018). This leads to the conclusion that the most appropriate path for reform is to improve specific decision-making processes using algorithms. Despite the prominence of this approach, it is fundamentally limited: egalitarian goals can rarely be achieved by reforming only the mechanisms of specific decision points. Reforms that aim to remedy structural oppression by targeting decision-making procedures often have the perverse effect of obscuring and entrenching the sources of oppression (Kahn, 2017; Murakawa, 2014). In the context of pretrial decision-making, implementing a risk assessment legitimizes preventative detention and hinders efforts to promote less carceral alternatives (Green, 2020).

In fact, the narrow purview of the formalism response makes the impossibility of fairness appear to be such a troubling and intractable dilemma. Simply put, it is only because analysis is restricted to decision-making procedures that that the tension between fairness definitions is interpreted as a fundamental "impossibility of fairness." Mathematical proofs demonstrate that it is impossible to satisfy all mathematical definitions of fairness when making decisions about individuals in an unequal society. What is strictly "impossible" is simultaneously achieving two different mathematical notions of fair decision-making. By limiting analysis to isolated decision points, however, formal algorithmic fairness magnifies the stakes of this mathematical incompatibility. When all other aspects of society are treated as static or irrelevant, an algorithm's behavior comes to represent "total fairness" (Berk et al., 2018). Under this assumption, the zero-sum tradeoff between mathematical definitions of fair decision-making represents an inescapable limitation on "total fairness." In other words, when the scope of analysis is restricted to specific decision-making processes, a constraint on fair decision-making becomes a constraint on fairness writ large. Within formal algorithmic fairness, the impossibility of fairness represents an intractable constraint on the ability to promote equality.

5 Substantive Algorithmic Fairness: Escaping the Impossibility of Fairness

The limits of formal algorithmic fairness suggest the need for a new approach to algorithmic fairness. Because it restricts analysis to isolated decision points, formal algorithmic fairness is unable to account for social hierarchies and the harmful policies that act on those hierarchies. In Anderson's terms, formal algorithmic fairness fails to "represent the causes of certain injustices" and therefore "can't help us identify solutions" that adequately address those injustices (Anderson, 2009). Reforms that look appealing within the frame of formal algorithmic fairness typically uphold oppression and are ensnared by the impossibility of fairness. In Dewey's terms, the issues with "what specific suggestions are entertained and which are dismissed" under formal algorithmic fairness are due to "[t]he way in which the problem is conceived" (Dewey, 1938). In order to develop a positive agenda for fairness-enhancing algorithmic reforms, it is necessary to reconceive the methodology of algorithmic fairness.

As an alternative to formal algorithmic fairness, I propose a method of "substantive algorithmic fairness." Drawing on substantive equality, substantive algorithmic fairness is an approach to algorithmic fairness in which the scope of analysis encompasses the social hierarchies and institutional structures that surround particular decision points. The goal is not to incorporate substantive concerns into the language of formal mathematical modeling. This approach of "formalist incorporation" may yield some benefits, but would be subject to many of the same limits as formal algorithmic fairness (Green and Viljoen, 2020). As with fairness more generally (Binns, 2018; Green and Hu, 2018; Jacobs and Wallach, 2021; Selbst et al., 2019), attempting to reduce substantive equality to mathematical definitions and metrics is likely to narrow and distort the concept.

Substantive algorithmic fairness therefore follows an approach of "algorithmic realism" (Green and Viljoen, 2020), adopting methods from law and philosophy for reasoning about what substantive equality entails and how to promote it with algorithms. The tools of substantive equality provide two particular benefits for reasoning about algorithmic fairness. First, substantive equality emphasizes the need to identify and redress social hierarchies, which form the "substance" of inequality. Second, when facing social hierarchies, substantive equality suggests strategies for reform that avoid falling into intractable, zero-sum dilemmas between competing notions of fairness.

These substantive principles shed light on how to escape from the impossibility of fairness. Debates and consternation about the impossibility of fairness are most extreme when making decisions in which a) relevant attributes are unevenly distributed across groups due to oppression and b) the oppressed group disproportionately exhibits the attributes that lead to receiving negative decisions (i.e., receiving punishment or not receiving benefit). Without these relational and structural factors, the impossibility of fairness would not arise. When these factors are present, however, any attempt to improve decision-making with an algorithm will confront the impossibility of fairness. Yet this does not mean that it is impossible to achieve more significant improvements in equality. Instead, the impossibility of fairness means that algorithms are being used to pursue a misguided reform strategy. As Fishkin notes, "If […] zero-sum tradeoffs are the primary tools of equal opportunity policy, then trench warfare is a certainty, and any successes will be incremental" (Fishkin, 2014). The proper response to the impossibility of fairness is not to tinker within the contours of this intractable dilemma, but to reform the relational and structural factors that produce the dilemma. Following this logic, substantive algorithmic fairness suggests roles for algorithms to alleviate oppression without being constrained by the impossibility of fairness.

5.1 The Substantive Algorithmic Fairness Approach to Reform

As with formal algorithmic fairness, the starting point for reform is concern about discrimination or inequality within a particular decision-making process. Drawing on the substantive equality approaches introduced in Section 3, substantive algorithmic fairness presents a three-step strategy for promoting equality in such scenarios. Each step can be boiled down to a single driving question. 1) What is the substance of the inequalities in question? 2) What types of reforms can remediate the identified substantive inequalities? 3) What roles, if any, can algorithms play to enhance or facilitate the identified reforms?

The first step is to consider the substance of the inequalities in question. This entails looking for conditions of hierarchy and questioning how social and institutional arrangements reinforce those conditions (MacKinnon, 2011). When faced with disparities in data, a substantive approach asks: do these disparities reflect social conditions of hierarchy? Similarly, when faced with particular decision points, a substantive approach asks: do these decisions (and the interventions that they facilitate) exacerbate social hierarchies? If the answers to these questions are no, then formal algorithmic fairness presents an appropriate path forward. However, if the answers to these questions are yes—as they often will be when confronting inequalities in high-stakes decisions—it suggests that reforms should not be limited to the decision-making process alone.[14] Such efforts would confront the impossibility of fairness, leading to the multitude of issues described in Section 4.

The second step is to consider what types of reforms are equipped to remediate the identified substantive inequalities. A substantive analysis will often reveal that concerns about unfair decision-making are caused by disparities between groups that are the product of oppression and by decisions that magnify the impacts these disparities. In these settings, substantive algorithmic fairness draws on the reforms proposed by Anderson (Anderson, 1999), Minow (Minow, 1991), and Fishkin (Fishkin, 2014) for promoting equality without becoming trapped by intractable equality dilemmas. The first approach is the relational response: reform the relationships that create and sustain social hierarchies. The second approach is the structural response: reshape the

[14] Of course, even answering these questions represents a political and potentially contested task. Substantive equality provides conceptual tools for making these judgments. Answers should also be informed by engagement with the communities in question.

structure of decisions to avoid or lower the stakes of decisions that act on social hierarchies. Thus, substantive algorithmic suggests searching for ways to a) ameliorate the underlying conditions of social hierarchy and b) reduce the scope and harms of decisions that hinge on attributes which are unequally distributed due to social hierarchy. Because these reforms consider a scope beyond isolated decision points, they are not subject to the impossibility of fairness.

Finally, the third step is to analyze whether and how algorithms can enhance or facilitate the reforms identified in the second step. In considering the potential role for algorithms, computer scientists should be wary of technological determinism and the assumption that algorithms can be applied to remedy all social problems. Algorithmic interventions should be considered through an "agnostic approach" that prioritizes the reforms identified in the second step, without assuming any necessary or particular role for algorithms (Green and Viljoen, 2020). This approach requires decentering technology when studying injustice and remaining attentive to the broader structural forces of marginalization (Gangadharan and Niklas, 2019). Although these practices will often reveal that algorithms are unnecessary or even detrimental tools for reform, they can also prompt new approaches for developing and applying algorithms to combat oppression. Algorithms can play productive roles in support of broader efforts for social change (Abebe et al., 2020), particularly when deployed in conjunction with policy and governance reforms (Green, 2019).

5.2 Example: The Substantive Algorithmic Fairness Approach to Pretrial Reform

We can see the benefits of substantive algorithmic fairness by considering how it applies in the context of reforming pretrial decision-making. As described above, formal algorithmic fairness suggests that the appropriate response to injustice within pretrial decision-making is to make release/detain determinations using algorithmic predictions of risk. Despite the support for pretrial risk assessments among many engineers and policymakers, this approach upholds racial injustice and leaves decision-making caught within the impossibility of fairness. In contrast, substantive algorithmic fairness suggests paths for pretrial reform that more robustly challenge the injustices associated with pretrial decision-making and that provide an escape from the impossibility of fairness. Although this approach highlights the limits of pretrial risk assessments, it also suggests new paths for reform and new roles for algorithm.

When pursuing pretrial reform under substantive algorithmic fairness, the first step is to consider the substance of inequalities that manifest in pretrial decision-making. As described in Section 4.1, the disparity in recidivism rates across Black and white defendants reflects conditions of racial hierarchy. This disparity cannot be attributed to chance or to natural group differences (nor is it solely the result of measurement bias, although measurement bias is often present). Furthermore, preventative detention exacerbates this hierarchy by depriving high-risk defendants of rights and subjecting them to a range of negative outcomes.

The second step of pretrial reform under substantive algorithmic fairness is to consider what reforms could appropriately address the substantive inequalities identified in the first step. The substantive analysis in step one reveals that the central problems of pretrial decision-making are the unequal distribution of risk across the population and the harmful policy responses to high levels of risk. Any effort to reform the pretrial decision-making process alone (e.g., with a risk assessment) will entrench these conditions and will confront the impossibility of fairness. Instead, we should apply the relational and structural responses derived from Anderson, Minow, and Fishkin.

The relational response suggests altering the relationships that define “risk” and shape its unequal distribution across the population. The formal algorithmic fairness approach of attempting to differentiate people by risk levels treats risk as an intrinsic and neutral attribute of individuals, naturalizing group differences in risk that are the product of oppression. Instead, we should interrogate “the social arrangements that make those traits seem to matter” (Minow, 1991). The relational response thus suggests attempting to reduce the crime risk of Black communities by alleviating criminogenic conditions of disadvantage. For instance, public policies that extend access to education (Lochner and Moretti, 2004), welfare (Tuttle, 2019), and affordable housing (Diamond and McQuade, 2019) all reduce crime, and therefore could reduce the racial disparity in crime risk. The relational response also suggests combatting the association of Blackness with criminality (Butler, 2017; Muhammad, 2011) and the effects of this association. This effort entails not merely challenging stereotypes that link Blackness with crime, but also decriminalizing behaviors that have been criminalized in the past to subjugate minorities.

The structural response suggests altering the structure of decisions to reduce the harmful consequences that are associated with being high risk for recidivism. The formal algorithmic fairness approach of informing preventative detention decisions upholds the notion that the appropriate response to high-risk defendants is to incarcerate them. Instead, we should “renovate the structure” of decisions (Fishkin, 2014) to ensure that being high risk no longer prompts such severe punishment. The structural response thus suggests attempting to minimize the scope and harms of decisions that determine one’s freedom and opportunities based on their risk of future crime. When fewer people are subjected to decisions in which liberty and well-being depend on having low levels of crime risk, existing racial disparities in the distribution of risk become less consequential. Most directly, such an approach could entail abolishing (or drastically reducing the scope of) pretrial detention, such that fewer people would stand to be incarcerated, regardless of their risk level. Reforms could also aim to decrease the downstream damages of pretrial detention; for instance, reducing the effects of pretrial detention on increased conviction and diminished future employment, would reduce the harms associated with being high risk. Another reform along these lines would be to shift from responding to risk with punishment to responding with social or material support, such that the consequence of being high risk is to receive aid rather than incarceration.

The third step of pretrial reform under substantive algorithmic fairness is to consider the potential role for algorithms in advancing relational and structural reforms. Because pretrial risk assessments naturalize racial disparities in risk that are the product of oppression and legitimize the policy of detaining defendants due to their risk levels, these algorithms conflict with the relational and structural responses. However, these issues do not rule out fruitful roles for algorithms in pretrial reform. Following the relational response, algorithms could be used to reduce the crime risk of disadvantaged groups by improving access to resources such as education (Lakkaraju et al., 2015), welfare (DataSF, 2018), and affordable housing (Ye et al., 2019). Efforts to mitigate the biases in policing practices and crime data would reduce the perceived recidivism risk of Black defendants. Following the structural response, algorithms could be used to reduce the harms of the racial disparity in recidivism risk. For instance, algorithms can be used to target supportive (rather than punitive) responses to risk (Barabas et al., 2018; Mayson, 2019), thus mitigating rather than compounding the injustices behind the high recidivism risk of Black defendants. More broadly, algorithms could be used to audit the design and implementation of risk assessments (Angwin et al., 2016; Green and Chen, 2019), enable a systemic view of how the criminal justice system exacerbates racial inequalities (Crespo, 2015; Goel et al., 2016), and empower communities advocating for criminal justice reform (Asad, 2019; Costanza-Chock, 2020)—all of which would help to inform and enable structural responses.

Substantive algorithmic fairness demonstrates how an expansive analysis of social conditions and institutions can lead to rigorous theories of social change, and how, in turn, those theories of change can inform work on algorithms. Because the substantive responses target social relations and the structure of decisions (rather than isolated decision-making procedures), they enable algorithmic interventions that that escape the harsh tradeoffs imposed by the impossibility of fairness. Substantive algorithmic fairness thus presents significant benefits for reforming pretrial decision-making—the policy area in which the impossibility of fairness has produced the most consternation. These benefits could accrue similarly in other areas in which the impossibility of fairness has been interpreted as a significant and intractable barrier on reform, such as child welfare (Chouldechova et al., 2018) and college admissions (Friedler et al., 2021).

5.3 The Path Forward

Substantive algorithmic fairness presents a new direction for algorithmic fairness. It shifts the field's concern away from formal mathematical models of "fairness" and toward substantive evaluations of how algorithms can (and cannot) combat social hierarchies. In doing so, substantive algorithmic fairness brings the field back in line with the demands for social justice that gave rise to algorithmic fairness in the first place (i.e., concerns among academics, journalists, policymakers, and the public about algorithms that discriminate and entrench inequality). Although there remains a role for formal evaluations of algorithmic decision-making, the field's work should primarily focus on studying how algorithms can be incorporated into broader efforts to promote equality. Substantive algorithmic fairness thus requires new modes of research and training. Researchers must engage deeply not only with scholars from outside the computational sciences, but also with communities directly advocating for reform. Similarly, training in algorithmic fairness must move beyond mathematical methods to also provide rigorous training in social change and sociotechnical systems.

Substantive algorithmic fairness does not provide a precise roadmap for reform. It presents a sequence of questions, with conceptual tools for answering those questions, rather than a comprehensive or mandatory checklist. It cannot be reduced to an optimization problem. This lack of explicit prescription is not so much a flaw of substantive algorithmic fairness as an inescapable reality of pursuing social and political reform. There is no single or straightforward path for how to achieve change. The hardest political questions often revolve around which reforms to pursue in any specific situation, among many potential paths forward. Making these judgments requires contextual assessments of feasibility and impact as well as engagement with affected communities. In some settings, particularly where substantive concerns about social hierarchy and unjust policies are less severe, this analysis may even suggest a role for reforms that align with formal algorithmic fairness. There similarly is no straightforward mechanism for determining what roles algorithms should play in reform efforts. Future work is necessary to better understand what roles algorithms can and cannot play in reform efforts, and under what conditions.

Nonetheless, substantive algorithmic fairness provides a compass to help computer scientists and others reason rigorously and practically about the appropriate roles for algorithms in efforts to combat inequity. Debates about algorithmic reforms often feature a binary contest between algorithmic reforms and the status quo, with proponents for algorithms arguing that the alternative to implementing algorithms is to fall back on even more fallible and biased human decision-makers (Berk et al., 2018; Kleinberg et al., 2019; Miller, 2018). Although often well-intentioned, reforms following this logic typically reinforce social hierarchies (Davis et al., 2021; Green, 2020; Green, 2021; Hoffmann, 2019; Ochigame, 2020; Ochigame et al., 2018; Powles and Nissenbaum, 2018). Substantive algorithmic fairness demonstrates that reformers need not choose between implementing a superficially "fair" algorithm and leaving the status quo in place. Although substantive algorithmic fairness begins with a broad (some might say utopian) vision of substantive equality, it presents multiple strategies for pursuing this goal, which in turn suggest many specific potential algorithmic interventions. The reforms suggested by substantive algorithmic fairness are all incremental: none will create a substantively equal society on their own. Each reform, however, moves society one step closer to substantive equality. In this sense, substantive algorithmic fairness takes after political theories of "non-reformist reforms" (Gorz, 1967), "real utopias" (Wright, 2010), and prison abolition (McLeod, 2015), all of which present strategies for linking short-term, piecemeal reforms with long-term, radical agendas for social justice.

Of course, efforts to achieve substantive algorithmic fairness in practice face a variety of barriers. Many governments and technology companies benefit from and promote formal algorithmic fairness, as it allows them to embrace "fairness" without making significant political or economic concessions (Bui and Noble, 2020; Green, 2020; Powles and Nissenbaum, 2018). Efforts to achieve the reforms suggested by substantive algorithmic fairness will often confront these forces opposed to structural change. The exclusion of women and minorities from algorithm development leads to notions of algorithmic fairness that are inattentive to the lived realities of oppressed groups (West, 2020). Furthermore, institutional barriers and incentives hinder the necessary types of interdisciplinary research and training. Thus, as with all efforts to achieve substantive equality, substantive algorithmic fairness requires ongoing political struggle to achieve the conditions for reform.

6 Conclusion

Algorithmic fairness provides an increasingly prominent toolkit for theorizing about what fairness entails and how to achieve it. It is therefore essential to consider whether algorithmic fairness provides suitable conceptual and practical tools for reforming public policy and enhancing equality. If algorithmic fairness methodology cannot comprehensively recognize and represent the nature of injustices, it will fail to identify effective paths for remediating those injustices.

The current methodology of formal algorithmic fairness is poorly equipped to guide strategies for enhancing social equality. Because it restricts analysis to isolated decision points, formal algorithmic fairness cannot account for social hierarchies and the impacts of decisions informed by algorithms. As a result, formal algorithmic fairness suggests reforms that are impeded by the impossibility of fairness and that uphold social hierarchies. Before algorithmic fairness can provide a productive guide for efforts to achieve a more equal society, we must reformulate its methodology so that it encompasses more expansive conceptual and practical tools.

Substantive algorithmic fairness provides an alternative methodology that incorporates social hierarchies and the structure of decisions into the scope of fairness. As a result, substantive algorithmic fairness provides an escape from the impossibility of fairness and suggests new roles for algorithms in combatting oppression. In doing so, substantive algorithmic fairness provides a new orientation for algorithmic fairness, demonstrating how to act on recent calls to shift the field's emphasis from "fairness" to "justice" (Bui and Noble, 2020; Green, 2018) and to "equity" (D'Ignazio and Klein, 2020). Although this reorientation involves a shift away from formal mathematical models and interventions such as pretrial risk assessments, it also prompts a new positive agenda for how to develop and apply algorithms in the service of social change.

Although substantive algorithmic fairness does not present a precise roadmap for reform, it provides a compass with which to link incremental algorithmic reforms with ideal visions of substantive equality. Substantive algorithmic fairness reveals that reformers need not face a false dichotomy between implementing an algorithm and doing nothing. Instead, there are many potential reforms to consider—all of them, in some form, incremental—and many potential roles for algorithms to enable or supplement those reforms. Substantive algorithmic fairness helps to elucidate the range of possible reforms, evaluate which reforms can best advance equality, and consider what roles algorithms can play to support those reforms.

No single reform—algorithmic or otherwise—can create a utopian society. However, algorithmic fairness researchers need not restrict themselves to a formal algorithmic fairness methodology that often reinforces oppression. By starting from substantive accounts of social hierarchy and social change, the field of algorithmic fairness can stitch together incremental algorithmic reforms that collectively build a more egalitarian society.

7 References

Abebe R, Barocas S, Kleinberg J, et al. (2020) Roles for computing in social change. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 252–260.

Anderson E (2009) Toward a Non-Ideal, Relational Methodology for Political Philosophy: Comments on Schwartzman's *Challenging Liberalism*. *Hypatia* 24: 130-145.

Anderson ES (1999) What is the Point of Equality? *Ethics* 109: 287-337.

Angwin J and Larson J (2016) Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. *ProPublica*.

Angwin J, Larson J, Mattu S, et al. (2016) Machine Bias. *ProPublica*.

Arneson R (2013) Egalitarianism. *The Stanford Encyclopedia of Philosophy*.

Arnold Ventures (2019) Statement of Principles on Pretrial Justice and Use of Pretrial Risk Assessment.

Asad M (2019) Prefigurative Design as a Method for Research Justice. *Proceedings of the ACM on Human-Computer Interaction* 3.

Barabas C, Virza M, Dinakar K, et al. (2018) Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 62--76.

Baradaran S (2011) Restoring the Presumption of Innocence. *Ohio State Law Journal* 72: 723-776.

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

Berk R, Heidari H, Jabbari S, et al. (2018) Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*: 1-42.

Binns R (2018) Fairness in Machine Learning: Lessons from Political Philosophy. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 149--159.

Bui ML and Noble SU (2020) We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness. In: Dubber MD, Pasquale F and Das S (eds) *The Oxford Handbook of Ethics of AI.* Oxford University Press.

Butler P (2017) *Chokehold: Policing Black Men*: The New Press.

Chouldechova A (2017) Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5: 153-163.

Chouldechova A, Benavides-Prado D, Fialko O, et al. (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 134--148.

Cooper A and Smith EL (2011) Homicide Trends in the United States, 1980-2008. *U.S. Department of Justice, Bureau of Justice Statistics*.

Corbett-Davies S, Pierson E, Feller A, et al. (2017) Algorithmic Decision Making and the Cost of Fairness. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797-806.

Costanza-Chock S (2020) *Design Justice: Community-Led Practices to Build the Worlds We Need*: MIT Press.

Crenshaw KW (1988) Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law. *Harvard Law Review* 101: 1331-1387.

Crespo AM (2015) Systemic Facts: Toward Institutional Awareness in Criminal Courts. *Harvard Law Review* 129: 2049-2117.

D'Ignazio C and Klein LF (2020) *Data Feminism*: MIT Press.

DataSF (2018) Keeping Moms and Babies in Nutrition Program.

Davis JL, Williams A and Yang MW (2021) Algorithmic reparation. *Big Data & Society* 8.

Dewey J (1938) *Logic: The Theory of Inquiry*: Henry Holt and Company.

Diamond R and McQuade T (2019) Who Wants Affordable Housing in Their Backyard? An Equilibrium Analysis of Low-Income Property Development. *Journal of Political Economy* 127: 1063-1117.

Dieterich W, Mendoza C and Brennan T (2016) COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpoint Inc. Research Department*.

Dobbie W, Goldin J and Yang CS (2018) The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. *American Economic Review* 108: 201-240.

Doctorow C (2016) Algorithmic risk-assessment: hiding racism behind "empirical" black boxes. *Boing Boing*.

EdBuild (2019) $23 Billion.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*: St. Martin's Press.

Feldman M, Friedler SA, Moeller J, et al. (2015) Certifying and Removing Disparate Impact. In: *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259-268.

Fishkin J (2014) *Bottlenecks: A New Theory of Equal Opportunity*: Oxford University Press.

Flores AW, Bechtel K and Lowenkamp CT (2016) False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.". *Federal Probation* 80: 38-46.

Fredman S (2016) Substantive equality revisited. *International Journal of Constitutional Law* 14: 712-738.

Friedler SA, Scheidegger C and Venkatasubramanian S (2021) The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making. *Communications of the ACM* 64: 136–143.

Gangadharan SP and Niklas J (2019) Decentering technology in discourse on discrimination. *Information, Communication & Society* 22: 882-899.

Goel S, Rao JM and Shroff R (2016) Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics* 10: 365-394.

Gong A (2016) Ethics for powerful algorithms (1 of 4). *Medium*.

Gorz A (1967) *Strategy for Labor*: Beacon Press.

Green B (2018) Putting the J(ustice) in FAT. *Berkman Klein Center Collection - Medium*.

Green B (2019) *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future*: MIT Press.

Green B (2020) The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 594–606.

Green B (2021) Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing*.

Green B and Chen Y (2019) Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 90–99.

Green B and Hu L (2018) The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In: *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning*.

Green B and Viljoen S (2020) Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 19–31.

Hardt M, Price E and Srebro N (2016) Equality of Opportunity in Supervised Learning. In: *30th Conference on Neural Information Processing Systems (NIPS 2016)*. 3315-3323.

Harris K and Paul R (2017) Pretrial Integrity and Safety Act of 2017. *115th Congress*.

Hellman D (2020) Measuring Algorithmic Fairness. *Virginia Law Review* 106: 811-866.

Hoffmann AL (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22: 900-915.

Jacobs AZ and Wallach H (2021) Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 375–385.

Kahn J (2017) *Race on the Brain: What Implicit Bias Gets Wrong about the Struggle for Racial Justice*: Columbia University Press.

Kalluri P (2020) Don’t ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583: 169.

Kleinberg J, Ludwig J, Mullainathan S, et al. (2019) Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10: 113-174.

Kleinberg J, Mullainathan S and Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Koepke JL and Robinson DG (2018) Danger Ahead: Risk Assessment and the Future of Bail Reform. *Washington Law Review* 93: 1725-1807.

Krivo LJ, Peterson RD and Kuhl DC (2009) Segregation, Racial Structure, and Neighborhood Violent Crime. *American Journal of Sociology* 114: 1765-1802.

Lakkaraju H, Aguiar E, Shan C, et al. (2015) A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1909–1918.

Larson J, Mattu S, Kirchner L, et al. (2016) How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*.

Ligett K (2021) FAccT 2021 Keynote: In Praise of Flawed Mathematical Models.

Lochner L and Moretti E (2004) The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *American Economic Review* 94: 155-189.

MacKinnon CA (2011) Substantive Equality: A Perspective. *Minnesota Law Review* 96: 1.

MacKinnon CA (2016) Substantive equality revisited: A reply to Sandra Fredman. *International Journal of Constitutional Law* 14: 739-746.

Mayson SG (2019) Bias In, Bias Out. *Yale Law Journal* 128: 2218-2300.

McLeod AM (2015) Prison Abolition and Grounded Justice. *UCLA Law Review* 62: 1156-1239.

Merriam-Webster (2021) Methodology.

Miller AP (2018) Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review*.

Minow M (1991) *Making All the Difference: Inclusion, Exclusion, and American Law*: Cornell University Press.

Muhammad KG (2011) *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America*: Harvard University Press.

Murakawa N (2014) *The First Civil Right: How Liberals Built Prison America*: Oxford University Press.

O’Neil C (2016) ProPublica report: recidivism risk models are racist. *mathbabe*.

Obermeyer Z, Powers B, Vogeli C, et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366: 447-453.

Ochigame R (2020) The Long History of Algorithmic Fairness. *Phenomenal World*.

Ochigame R, Barabas C, Dinakar K, et al. (2018) Beyond Legitimation: Rethinking Fairness, Interpretability, and Accuracy in Machine Learning. In: *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning*.

Porrino CS (2017) Attorney General Law Enforcement Directive 2016-6 v3.0: Modification of Directive Establishing Interim Policies, Practices, and Procedures to Implement Criminal Justice Reform Pursuant to P.L. 2015, c. 31.

Powles J and Nissenbaum H (2018) The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence. *OneZero*.

Raji ID and Buolamwini J (2019) Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 429-435.

Rose DR and Clear TR (1998) Incarceration, Social Capital, and Crime: Implications for Social Disorganization Theory. *Criminology* 36: 441-480.

Rothstein R (2017) *The Color of Law: A Forgotten History of How Our Government Segregated America*: Liveright Publishing Corporation.

Sampson RJ, Morenoff JD and Raudenbush S (2005) Social Anatomy of Racial and Ethnic Disparities in Violence. *American Journal of Public Health* 95: 224-232.

Selbst AD, Boyd D, Friedler SA, et al. (2019) Fairness and Abstraction in Sociotechnical Systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 59-68.

Sunstein CR (2019) Algorithms, Correcting Biases. *Social Research* 86: 499-511.

Tuttle C (2019) Snapping Back: Food Stamp Bans and Criminal Recidivism. *American Economic Journal: Economic Policy* 11: 301-327.

U.S. Supreme Court (1987) United States v. Salerno. *481 U.S. 739*.

West SM (2020) Redistribution and Rekognition. *Catalyst: Feminism, Theory, Technoscience* 6.

Wright EO (2010) *Envisioning Real Utopias*: Verso.

Ye T, Johnson R, Fu S, et al. (2019) Using machine learning to help vulnerable tenants in New York City. In: *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, 248–258.