Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness
Ben Green, University of Michigan (bzgreen@umich.edu)

Abstract
Algorithmic fairness provides novel methods for promoting equitable public policy using machine learning. Yet the narrow formulation of algorithmic fairness often provides cover for algorithms that exacerbate oppression, leading critics to call for a more justice-oriented approach. This article takes up these calls and proposes a method for operationalizing a social justice orientation into algorithmic fairness. First, I argue that algorithmic fairness suffers from a significant methodological limitation: it restricts analysis to isolated decision points. Because algorithmic fairness relies on this narrow scope of analysis, it yields a reform strategy that is fundamentally constrained by the “impossibility of fairness” (an incompatibility between mathematical definitions of fairness). Second, in light of these flaws, I draw on theories of substantive equality from law and philosophy to propose an alternative methodology: “substantive algorithmic fairness.” Because substantive algorithmic fairness takes a more expansive scope to fairness, it suggests reform strategies that escape from the impossibility of fairness. These strategies provide a rigorous guide for employing algorithms to alleviate social injustice. In sum, substantive algorithmic fairness presents a new direction for the field of algorithmic fairness: away from formal mathematical models of “fairness” and toward substantive evaluations of how algorithms can (and cannot) promote justice.

1 INTRODUCTION
1.1 Algorithmic Fairness and Its Discontents
Machine learning algorithms have become central components in many efforts to promote equitable public policy. In the face of widespread concerns about discriminatory institutions and decision-making processes, many policymakers and scholars praise algorithms as critical tools for enabling equality-enhancing reforms [7, 33, 50, 79]. To policymakers, policy advocates, and scholars across multiple fields, algorithms overcome the cognitive limits and social biases of human decision-makers, enabling more objective and fair decisions [7, 50, 56, 69, 86]. Thus, for instance, in light of concerns about the biases of judges, many court systems in the United States have adopted pretrial risk assessments as a central component of criminal justice reform efforts [45, 58, 79].

Undergirding these reform efforts is the burgeoning field of algorithmic fairness. Grounded primarily in computer science, algorithmic fairness applies the tools of algorithm design and analysis—in particular, an emphasis on formal mathematical reasoning [48]—to questions of fairness. The central components of algorithmic fairness are developing mathematical definitions of fair decision-making [11], optimizing algorithms for these definitions [34, 49], and auditing algorithms for violations of these definitions [5, 75, 81]. In the context of reform efforts, algorithmic fairness is often employed to determine whether an algorithm is “fair” and, therefore, appropriate to use for decision-making. For instance, in regulation and in settings such as pretrial decision-making and child welfare, debates about whether and how to employ algorithms often hinge on evaluations of algorithmic fairness [5, 16, 29, 33].

Yet even as algorithmic fairness has risen in prominence, critical scholars have raised doubts about its ability to support desirable policy reforms. Efforts to formulate mathematical definitions of fairness overlook the contextual and philosophical meanings of fairness [13, 47, 53, 85]. As a result, algorithmic fairness fails to provide a reliable standard for promoting fairness in practice.
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 [52]. When deployed in practice, algorithms that satisfy fairness standards often reproduce inequities and legitimize unjust institutions [26, 45, 55, 76, 77, 80]. In turn, some scholars have called for rejecting the frame of “fairness” altogether, proposing alternative frames of “justice” [14, 43], “equity” [24], and “reparation” [26].

This article takes up these calls and proposes a method for operationalizing a social justice orientation into algorithmic fairness. To accomplish this goal, it is necessary to grapple with a fundamental technical limitation in algorithmic fairness: the “impossibility of fairness.” This result reveals that it is impossible for an algorithm to satisfy all desirable mathematical definitions of fair decision-making [17, 57]. An algorithm that is fair along one standard will inevitably be unfair along another standard. Although no mathematical definitions of algorithmic fairness fully encapsulate the philosophical notion of fairness [13, 47, 53, 85], each captures a normatively desirable principle.

The impossibility of fairness presents a significant practical limit on efforts to promote fairness using algorithms: all algorithms will violate at least one normatively appealing fairness definition. As one article about algorithmic fairness concludes, “the tradeoff between […] different kinds of fairness has real bite” and means that “total fairness cannot be achieved” [12]. The impossibility of fairness means that any effort to improve decision-making using algorithms—regardless of whether it adheres to the label of fairness, justice, equity, or reparation—will confront an intractable constraint on the ability to achieve the animating normative ideals. Algorithmic justice cannot be achieved merely by selecting a particular fairness definition at the expense of others or by evaluating the tradeoffs between fairness definitions, as some scholars have proposed [26, 56]. Yet given the impossibility of fairness, these appear to be our best paths forward.

The impossibility of fairness clarifies the task of reorienting algorithmic fairness toward justice and equity. Simply put, any attempt to develop more justice-aligned algorithms must find a way to escape from the impossibility of fairness. Thus, in addition to reforms such as shifting the power structures behind algorithmic fairness [14, 24, 26, 89], it is also necessary to reform the methodology of algorithmic fairness.

With the goal of algorithmic justice in mind, this article makes two primary contributions. First, I demonstrate that the problems of algorithmic fairness result from the dominant methodology of the field, which limits analysis to isolated decision-making procedures. Second, in light of these flaws, I propose a new methodology for algorithmic fairness that enables an escape from the impossibility of fairness and that, in turn, is better equipped to alleviate social injustice.

1.2 Article Overview: Methodological Reform

A methodology is “a body of methods, rules, and postulates employed by a discipline” [68]. A methodology provides a systematic language for comprehending and reasoning about the world, shaping how practitioners formulate problems and develop solutions to those problems. Problem formulation has both practical and normative stakes [78]. As philosopher John Dewey writes, “The way in which [a] problem is conceived decides what specific suggestions are entertained and which are dismissed” [27]. A poorly conceived problem “cause[s] subsequent inquiry to be irrelevant or to go astray;” the remedy is to reformulate the problem [27]. Furthermore, 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

---

1 I will provide more detail on these fairness definitions and the impossibility of fairness in Section 2.
them. If they can’t represent the causes of certain injustices, then they can’t help us identify solutions” [2]. In sum, if a methodology fails to account for normatively relevant facts and principles, it will generate problem formulations that yield unhelpful or unjust proposals for reform.

In the spirit of Dewey and Anderson, I argue that algorithmic fairness suffers from a significant methodological limitation: it relies on a narrow frame of analysis restricted to specific decision points, in isolation from the context of those decisions.2 I call this method “formal algorithmic fairness,” as it aligns with formal equality (which emphasizes equal treatment for individuals based on their attributes or behavior at a particular decision point). Formal algorithmic fairness represents a systematic approach to problem formulation in which fairness is operationalized in terms of isolated decision-making processes. Because formal algorithmic fairness is conceived so narrowly, it cannot represent many of the injustices associated with algorithmic decision-making. Following this problem formulation, formal algorithmic fairness yields a misguided and techno-centric reform strategy: enhance fairness by optimizing decision-making procedures with algorithms. These algorithmic interventions fail to reduce (and often even exacerbate) oppression and are constrained by the impossibility of fairness. Thus, formal algorithmic fairness leaves reform efforts in a bind: it appears that our only options are to pursue superficially “fair” algorithms or to reject algorithmic reforms, leaving the status quo in place.

In light of these flaws, I propose an alternative approach to algorithmic fairness that enables a more justice-oriented positive agenda for developing and applying algorithms. I call this method “substantive algorithmic fairness,” as it draws on theories of substantive equality from law and philosophy (which emphasize the identification and reduction of social hierarchies). 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. My goal is not to incorporate these considerations into a formal mathematical model, a strategy that would fail to provide the necessary methodological shift [48]. Substantive algorithmic fairness is not a method for creating “substantively fair algorithms.” Instead, following an “algorithmic realist” approach [48], my goal is to develop problem formulations that incorporate modes of reasoning from law and philosophy. Substantive algorithmic fairness is a method for rigorously reasoning about how algorithms should (and should not) be used to promote substantive equality.

Following this problem formulation, substantive algorithmic fairness suggests reforms beyond either implementing a superficially “fair” algorithm or leaving the status quo in place. Substantive algorithmic fairness presents a three-step strategy for promoting substantive equality with algorithms: 1) evaluate the substance of the inequalities in question, 2) determine what reforms can remediate the identified inequalities, and 3) analyze how algorithms can enhance the desired reforms. This strategy reveals that the impossibility of fairness is a misnomer: when an algorithm confronts the incompatibility between fairness definitions, it suggests not that fairness is impossible writ large, but that algorithmic reforms are fundamentally limited in the given context. In other words, the impossibility of fairness indicates that it would be fruitful to pursue a reform strategy that involves more than just algorithmic decision-making. In contrast to the formal algorithmic fairness approach of optimizing decision-making procedures with algorithms, substantive algorithmic fairness calls for a) alleviating social hierarchies and b) reducing the scope

---

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.
and stakes of decisions that exacerbate social hierarchies. These reforms ameliorate the relational and structural sources of oppression that produce the impossibility of fairness, thus escaping this dilemma. In sum, substantive algorithmic fairness presents concrete steps toward a new method for algorithmic fairness: away from formal mathematical models of “fairness” as an attribute of algorithms and toward substantive evaluations of how algorithms can (and cannot) support justice-enhancing reforms.

2 THE IMPOSSIBILITY OF FAIRNESS

In May 2016, journalists at ProPublica reported that a risk assessment algorithm used to judge pretrial defendants in Broward County, Florida was “biased against blacks” [5]. This algorithm, known as COMPAS, was created by the company Northpointe and is used by many court systems across the United States.³ Like other pretrial risk assessments, COMPAS predicts the likelihood that pretrial defendants will recidivate; these predictions are presented to judges to inform their decisions to release or detain each defendant until their trial [45, 58]. 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) [5].

Tech critics responded to ProPublica’s article with outrage about racist algorithms [31, 74]. However, Northpointe and numerous academics defended COMPAS, arguing that ProPublica had focused on the wrong measure of algorithmic fairness [20, 29, 36, 40]. These groups asserted that the proper standard of fairness is not whether false positive (and false negative) rates are the same for each race but whether risk scores imply the same probability of recidivism for each race. COMPAS satisfied this notion of fairness, suggesting that the tool was fair.

This debate about whether COMPAS is fair 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.⁴ Separation expresses the idea that people who exhibit the same outcome should be treated similarly. 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 [5].

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.⁵ Sufficiency expresses the idea that people who are equally likely to exhibit the behavior of interest should be treated similarly. Northpointe and others argued that COMPAS is fair because it satisfies sufficiency: the label of “high risk” signifies a similar probability of recidivism for both Black and white defendants [20, 29, 36, 40]. Sufficiency is the most widely used notion of algorithmic fairness, particularly because machine learning models typically satisfy this principle by default [11].

The COMPAS debate raised a fundamental question for algorithmic fairness: can an algorithm simultaneously satisfy separation and sufficiency? As computer scientists soon discovered, the answer is no: there is an inevitable tension between these definitions of fairness [4, 11, 17, 57]. This result is known as the “impossibility of fairness.” The only exceptions to the impossibility of fairness involve two exceedingly rare scenarios: the algorithm makes predictions with perfect

---

³ COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions. Northpointe has since been renamed Equivant.
⁴ Separation is aligned with fairness criteria such as error rate balance and balance for the positive/negative class.
⁵ Sufficiency is aligned with fairness criteria such as calibration and predictive parity.
accuracy, or all groups exhibit the outcome being predicted at the same “base rate” [57]. Thus, for instance, a pretrial risk assessment will necessarily either misclassify Black and white defendants as recidivists at different rates (violating separation) or yield different predictions for Black and white defendants who are equally likely to recidivate (violating sufficiency).

The impossibility of fairness reflects a harsh and intractable dilemma facing efforts to promote equality using algorithms [12]. Work on algorithmic fairness operates within the constraints posed by this dilemma, accepting that the best we can do is to choose a single fairness definition (at the expense of others) or to rigorously balance the tradeoff between multiple definitions. Yet as I will describe (using pretrial risk assessments as a case study), both of these responses lead to narrow reforms that uphold unjust social conditions and institutions. Developing a positive agenda for algorithmic justice requires finding a way to develop and apply algorithms without confronting the impossibility of fairness.

3 LESSONS FROM EGA LITARIAN THEORY
In order to inform the evolution from the existing method of algorithmic fairness to a more justice-oriented approach, I turn to egalitarian theory. Broadly speaking, “Egalitarian doctrines tend to rest on a background idea that all human persons are equal in fundamental worth or moral status” [6]. Although fairness and equality are complex and contested concepts, both share a central concern with comparing the treatment or conditions across individuals or groups, emphasizing the normative value of some form of parity [6, 42, 71]. Indeed, many definitions of algorithmic fairness make explicit reference to equality [11, 12]. Furthermore, egalitarian thinkers have confronted many questions that overlap with central debates in algorithmic fairness [13]. Egalitarian insights can therefore yield important lessons for algorithmic fairness.

3.1 Formal and Substantive Equality
The first benefit of egalitarianism for algorithmic fairness is that debates between “formal” and “substantive” equality shed light on why the current method of algorithmic fairness leads to injustice and how to develop an alternative approach. Just as algorithmic fairness confronts the limits of narrow formulations of fairness, egalitarian theorists have confronted similar limits of narrow formulations of equality. In response, some egalitarian thinkers have devised more expansive formulations that provide a better guide for ameliorating oppression.

A central tension in egalitarian theory is between “formal” and “substantive” equality. Formal equality asserts, “When two persons have equal status in at least one normatively relevant respect, they must be treated equally with regard in this respect. This is the generally accepted formal equality principle that Aristotle articulated […]: ‘treat like cases as like’” [42]. In practice, formal equality typically refers to a “fair contest” in which everyone is judged according to the same standard, based only on their characteristics at the moment of decision-making [35]. In the United States, disparate treatment law is grounded in notions of formal equality, attempting to ensure that people are not mistreated based on protected attributes such as race or gender.

Despite being widely adopted, formal equality suffers from a methodological limitation. Because formal equality restricts analysis to specific decisions at specific points in time, it cannot account for inequalities in the distribution of attributes across the population. Formal equality is therefore prone to reproducing existing patterns of injustice. For instance, a formal equality approach to college admissions would evaluate all applicants based solely on their academic qualifications (e.g., grades and test scores). As long as applicants with similar qualifications are treated similarly, formal equality would be satisfied, even if these decisions lead to racial
disparities. Thus, 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” [22]. Because of racial inequalities in educational opportunities [32], evaluating all students according to a uniform standard would perpetuate racial hierarchy.

The limits of formal equality have led many scholars to develop an alternative frame: substantive equality. This approach “repudiate[s] the Aristotelian ‘likes alike, unlikes unalike’ approach […] and replaces it with a substantive test of historical disadvantage” [64]. “Its core insight is that inequality, substantively speaking, is always a social relation of rank ordering, typically on a group or categorical basis,” that leads to both material and dignitary inequalities [64]. In other words, “hierarchy identifies the substance of substantive equality” [65]. Following this line of reasoning, substantive equality envisions a world free from social hierarchy [64, 65].

In the United States, disparate impact law is grounded in notions of substantive equality (albeit partially [64, 65]), attempting to ensure that formally neutral rules do not disproportionately burden historically marginalized groups.

Substantive equality provides the methodological capacity to identify and ameliorate social hierarchies. In contrast to formal equality, substantive equality relies on a broad frame of analysis that evaluates decisions in light of social hierarchies. 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” [64]. For instance, substantive equality recognizes that racial disparities in college admissions reflect a pervasive racial hierarchy in educational and other opportunities. It therefore rejects the formal equality approach to college admissions. Rather than aiming to evaluate all students according to a uniform standard, substantive equality calls for policies that acknowledge this racial hierarchy (such as affirmative action) and that aim to redress this hierarchy (such as improving educational resources in minority school districts).

As Section 4 will describe, the current approach to algorithmic fairness—which I call “formal algorithmic fairness”—is grounded in formal equality and shares many of formal equality’s limits. This analysis suggests the need for an alternative approach grounded in substantive equality—“substantive algorithmic fairness”—which I present in Section 5.

3.2 Substantive Approaches to Escaping Equality Dilemmas

The second benefit of egalitarianism for algorithmic fairness is that substantive equality provides conceptual tools for escaping from the impossibility of fairness. Just as algorithmic fairness confronts the impossibility of fairness, egalitarian theorists have confronted similar seemingly unresolvable dilemmas between notions of equality. In response, some egalitarian thinkers have devised reform strategies that break free from these dilemmas.

In order to glean insights about how algorithmic fairness can escape the impossibility of fairness, I turn to three complementary substantive equality approaches for analyzing and escaping from equality dilemmas:

- In developing her theory of “democratic equality,” philosopher Elizabeth Anderson responds to a “dilemma” that arises in luck egalitarianism [3]. On the one hand, not providing aid to the disadvantaged means blaming individuals for their misfortune. On the

---

6 Luck egalitarianism advocates compensating people for inequalities that result from misfortunate but not inequalities that result from choice [3, 6].
other hand, providing special treatment to individuals on account of their inferiority means expressing contempt for the disadvantaged.

- In developing her “social-relations approach” to equality, legal scholar Martha Minow engages with the “dilemma of difference” that arises in legal efforts to deal with differences between individuals [70]. 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.

- In developing his theory of “opportunity pluralism,” legal scholar Joseph Fishkin addresses the “zero-sum struggles” that arise in efforts to promote equal opportunity [35]. On the one hand, judging people for an opportunity based solely on their performance or attributes at a particular moment in time perpetuates inequalities. On the other hand, even approaches that attempt to account for existing inequalities (such as Rawlsian equal opportunity and luck egalitarianism) fail to create a truly level playing field and prompt “extraordinarily contentious” debates.⁷

The equality dilemmas presented by Anderson, Minow, and Fishkin all resemble the impossibility of fairness: efforts to promote equality are impaired by a seemingly inescapable zero-sum tension between compelling-yet-conflicting notions of equality. In the face of these tradeoffs, it appears difficult—if not impossible—to make progress in pursuing equality. As Minow notes, “Dilemmas of difference appear unresolvable” [70]. In turn, “decisionmakers may become paralyzed with inaction” [70]. At best, decision-makers appear to be left with a zero-sum tradeoff between competing notions of equality. Yet 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” [35].

What makes Anderson, Minow, and Fishkin particularly insightful for algorithmic fairness is that they provide methodological accounts of how to escape from these dilemmas. Each scholar reveals that their dilemma is not intractable. Instead, each dilemma only appears intractable if one analyzes inequality through a narrow lens, which restricts the range of possible remedies. Expanding the frame of analysis clarifies the problems of inequality and yields two reform strategies that escape these equality dilemmas.

3.2.1 The Relational Response

First, a substantive analysis highlights how social hierarchies lead to equality dilemmas. Noting 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 equality of social relations [3]. 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 [3, 70].

Following the 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. 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” [70]. In other words, the primary task of

---

⁷ Each of these scholars are aligned with relational egalitarianism, which asserts that “people should relate to one another as equals or should enjoy the same fundamental status” [6]. Although Fishkin is the least explicitly focused on relationships, his analysis has strong overlaps with relational egalitarianism.
reform should not be providing special treatment to “different” individuals but reducing the extent to which superficial differences lead to significant disparities in status and abilities [70]. Without social hierarchies, real or perceived differences between individuals would not lead to different levels of rights or capacities, which in turn would prevent the 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 (indeed, the notion of inherent ugliness should be contested). Instead, “the injustice lies […] in the social fact that people shun others on account of their appearance” [3]. Oppressive social norms turn a superficial difference between people into one marked by severe disparities in status. This feature of social relations creates a dilemma. 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.

The relational response provides an escape from this dilemma: alter 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 differently based on how they look. Such reforms may be difficult to achieve (at least in the immediate term), thus necessitating more individualized remedies. 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” [3].

3.2.2 The Structural Response

Second, a substantive 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” [35]. These competitions typically hinge on attributes that are unequally distributed across groups, compounding existing disadvantage (i.e., oppressed groups are less qualified to succeed in competitions for beneficial opportunities, such as jobs).

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, “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” [35]. In other words, the primary task of reform should not be helping some disadvantaged individuals receive favorable decisions through special treatment but limiting the extent to which high-stakes decisions hinge on attributes that are unevenly distributed across social groups due to oppression. Without these bottlenecks, decisions would not as strongly magnify existing inequalities, which in turn would lower the 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 high 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. This feature of decision structures creates a dilemma. 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.

The structural response provides an escape from this dilemma: lower the stakes of college admissions decisions. Making college admissions less determinative of future life outcomes would reduce the downstream harms of disparities in early educational opportunities. Achieving this goal requires altering the structure of opportunities to create more paths for people to lead comfortable and fulfilling lives without a college degree. By making inequities in primary and secondary education less consequential, these reforms would reduce the dilemma between treating college applicants similarly or differently based on their academic performance.

The relational and structural responses present two concrete substantive equality approaches for dealing with equality dilemmas. As Section 5 will describe, substantive algorithmic fairness applies these substantive equality strategies to the impossibility of fairness. Following the relational and structural responses enables algorithms to escape the impossibility of fairness and to alleviate social hierarchies.

4 FORMAL ALGORITHMIC FAIRNESS: NAVIGATING THE IMPOSSIBILITY OF FAIRNESS

In this section, I characterize the attributes and limits of the dominant method of algorithmic fairness, which I call “formal algorithmic fairness.” Formal algorithmic fairness provides a computational methodology for responding to concerns about discriminatory decision-making. Akin to formal equality, formal algorithmic fairness limits analysis to the functioning of algorithms at particular decision points. When confronted with concerns about discriminatory decision-making, formal algorithmic fairness formulates the problem in terms only of the inputs and outputs of the decision point in question. All of the major definitions of algorithmic fairness are defined in this manner [11, 12]. As a result, formal algorithmic fairness suffers from many of the same limits as formal equality.

My goal in this section is to demonstrate not just that formal algorithmic fairness suffers from substantive flaws, but also that those flaws are the product of methodological limitations. To elucidate the methodological limits of formal algorithmic fairness, I interrogate its two responses to the impossibility of fairness. These responses reveal how the formulation of formal algorithmic fairness inevitably leads to algorithmic interventions that reproduce injustice and yield a notably constrained reform strategy. Even the best-case scenario within formal algorithmic fairness fails to provide a satisfactory response to the impossibility of fairness. All told, the central problem facing algorithmic fairness is not that we lack the appropriate formal definitions of fairness, that data is often biased, or that we cannot achieve sufficient predictive accuracy. The problem is the method of formal algorithmic fairness itself.

4.1 The Fair Contest Response: Reproducing Inequity

The first formal algorithmic fairness response to the impossibility of fairness is what I call the “fair contest response.” This response defends sufficiency as the proper definition of algorithmic fairness, asserting that fairness entails treating people similarly based solely on each person’s likelihood to exhibit the outcome of interest. On this view, as long as an algorithm satisfies sufficiency, any lack of separation is acceptable—it is the inevitable byproduct of groups
exhibiting the outcome in question at different rates. This response applies the logic of a “fair contest,” aiming to evaluate everyone based only on their characteristics at the moment of decision-making.

Most critiques of ProPublica’s COMPAS report followed the fair contest response, asserting that ProPublica focused on the wrong definition of fairness [20, 29, 36, 40]. These respondents argued that COMPAS is fair because it satisfies sufficiency: 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” [29].

Through a formal equality lens, the fair contest 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 likelihoods. For instance, if a Black and a white defendant are equally likely to be arrested in the future, then they should be given the 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 [51, 56].

However, evaluating the fair contest response through a substantive equality lens demonstrates the limits of this response. First, the fair contest 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 [29]. Instead, past and present discrimination has created social conditions in the US in which Black people are empirically at higher risk to commit crimes [19, 84]. This disparity results from social oppression rather than from differences in inherent criminality [72]. For instance, discriminatory practices such as segregation [83], racial criminalization [15, 72], and severe underfunding of schools [32] all increase crime [59, 63, 82].

Second, the fair contest response ignores the consequences of the actions that the algorithm informs. When a risk assessment labels a defendant “high risk,” that person is likely to be detained in jail until their trial. This practice of detaining defendants due to their crime risk, known as “preventative detention,” is both controversial and harmful. When the US Supreme Court deemed preventative detention constitutional in 1987, Justice Thurgood Marshall declared the practice “incompatible with the fundamental human rights protected by our Constitution” [88]. Preventative detention has faced continued scrutiny and challenge for undermining the rights of the accused and exacerbating mass incarceration [10, 58]. Pretrial detention imposes severe costs on defendants, including the loss of freedom, an increased likelihood of conviction, and a reduction in future employment [30].

By failing to account for these dimensions of pretrial decision-making, the fair contest response suggests a reform strategy in which even the best-case scenario—a perfectly accurate risk assessment—would perpetuate racial inequity.9 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

---

8 This is true above and beyond racial disparities in arrest and enforcement patterns. 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 [53].

9 Because this risk assessment makes perfect predictions, it would satisfy both sufficiency and separation [57].
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 \[19, 36, 61, 84\], a perfect risk assessment will accurately label a higher proportion of Black defendants as “high risk.”\(^{10}\) 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.

### 4.2 The Formalism Response: Constraining Reform

The second formal algorithmic fairness response to the impossibility of fairness is what I call the “formalism response.” Compared to the fair contest response, the formalism response takes a more measured view of the dilemma. Recognizing that sufficiency reflects a limited notion of fairness, the formalism response does not require strict 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.\(^{11}\)

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 [11, 12, 62]. Algorithms therefore provide “clarity” to help us identify and manage the unavoidable tradeoffs between competing goals [56, 86]. 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” [56].

As with the fair contest response, the formalism response appears appropriate through the lens of formal equality. If our interventions are limited to reforming specific decision-making procedures, then it is desirable to understand the tradeoffs between different fairness metrics. For instance, given an existing population of Black and white defendants, it is beneficial to have clarity on how optimizing a risk assessment for sufficiency will cause the algorithm to violate separation. 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, evaluating the formalism response through a substantive equality lens demonstrates the limits of this response. First, the formalism response leaves us stuck making a zero-sum choice 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 a defendant’s race [36, 66]. Such explicit differential treatment based on race

---

\(^{10}\) After all, if data is collected about an unequal society, then an accurate algorithm trained on that data will reflect those unequal conditions.

\(^{11}\) The formalism response is inclusive of the fair contest response: after considering the tradeoffs, one could determine that an algorithm should be optimized for sufficiency. The formalism response can also account for other tradeoffs, such as the tension between fairness and accuracy.
would be illegal to implement in many instances [20, 51]. Furthermore, although a lack of separation demonstrates that different groups face disparate burdens from mistaken judgments [17, 51], separation does not prevent 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. Although the formalism response provides “clarity” regarding the tradeoffs involved in promoting fairness, this clarity is limited to the narrow scope of specific decision-making procedures. Everything beyond this scope is treated as static and thus irrelevant to evaluations of fairness. For instance, research on fairness in risk assessments explicitly places structural disadvantage and racial disparities outside the scope of algorithms and the responsibility of developers [17, 20, 56]. Following this logic, the formalism response suggests that implementing an algorithm is the only possible (or, at least, pertinent) alternative to the status quo [12, 56, 69]. 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 obscure and entrench the sources of oppression [54, 73]. In the context of pretrial decision-making, implementing a risk assessment legitimizes preventative detention and hinders efforts to promote less carceral alternatives [45].

In fact, the narrow purview of the formalism response is what 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 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, turning a constraint on fair decision-making into a constraint on fairness writ large. When all other aspects of society are treated as static or irrelevant, an algorithm’s behavior comes to represent “total fairness” [12]. Under this assumption, the zero-sum tradeoff between mathematical definitions of fair decision-making represents an inescapable limitation on “total fairness.”

4.3 Recap: The Methodological Limits of Formal Algorithmic Fairness

The flaws of formal algorithmic fairness are methodological. Akin to formal equality, formal algorithmic fairness formulates fairness within the scope of isolated decision points. As a result, formal algorithmic fairness is unable to account for social hierarchies and the harmful policies that act on those hierarchies. In turn, formal algorithmic fairness suggests reforms that often entrench injustice and that are constrained by the impossibility of fairness. Any approach to algorithmic fairness that restricts analysis to decision-making processes will suffer from these flaws. 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 [2]. 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” [27]. Thus, in order to develop a positive agenda for algorithmic justice, it is necessary to develop a new methodology to algorithmic fairness grounded in substantive equality.
5 SUBSTANTIVE ALGORITHMIC FAIRNESS: ESCAPING THE IMPOSSIBILITY OF 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 a formal mathematical model. This approach of “formalist incorporation” may yield some benefits, but would be subject to many of the same limits as formal algorithmic fairness [48]. As with fairness more generally [13, 47, 53, 85], attempting to reduce substantive equality to mathematical definitions is likely to narrow and distort the concept. Substantive algorithmic fairness therefore follows an approach of “algorithmic realism” [48], adopting methods from law and philosophy to reason rigorously about how to promote substantive equality with algorithms.

Substantive equality sheds 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) an oppressed group disproportionately exhibits the attributes deemed “negative” in the given context (e.g., indicators of high crime risk), and b) policy punishes (or restricts benefits to) individuals who exhibit these negative attributes. When these relational and structural factors are present, any attempt to improve decision-making with an algorithm will confront the impossibility of fairness. However, the impossibility of fairness does not mean that it is impossible to enhance equality. Instead, the impossibility of fairness means that algorithms are being used to pursue a misguided reform strategy. 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. If there were no social hierarchies or if consequential decisions did not exacerbate social hierarchies, then the impossibility of fairness would not arise (or, at the very least, would not be so concerning). Following this logic, substantive algorithmic fairness follows the relational and structural reform strategies described in Section 3.2.

5.1 The Substantive Algorithmic Fairness Approach to Reform

As with formal algorithmic fairness, the starting point for reform in substantive algorithmic fairness 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 central 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 [64]. When faced with disparities in data, substantive algorithmic fairness asks: do these disparities reflect social conditions of hierarchy? Similarly, when faced with particular decision points, substantive algorithmic fairness 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—then it is necessary to pursue reforms through substantive algorithmic fairness.¹²

The second step is to consider what types of reforms can remediate the identified substantive inequalities. Substantive algorithmic fairness draws on the reforms proposed by Anderson [3], Minow [70], and Fishkin [35] for promoting equality without becoming trapped by intractable 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 structure of decisions to avoid or lower the stakes of decisions that exacerbate social hierarchies. Because these reforms target the relational and structural factors that produce equality dilemmas, 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 remedy all social problems. Algorithmic interventions should be considered through an “agnostic approach” that prioritizes the reform agenda identified in the second step, without assuming any necessary or particular role for algorithms [48]. This approach requires decentering technology when studying injustice and remaining attentive to the broader structural forces of marginalization [38]. 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 [1], particularly when deployed in conjunction with policy and governance reforms [44].

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 pretrial reform. Formal algorithmic fairness suggests that the appropriate pretrial reform strategy is to make release/detain decisions 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 algorithms.

When pursuing pretrial reform through 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). 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 is to consider what reforms could appropriately address the substantive inequalities identified in the first step. Here, we can compare risk assessments with the relational and structural responses. The relational response suggests altering the relationships that define “risk” and shape its unequal distribution across the population. By differentiating people based on

---

¹² 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 broad deliberation that includes the communities in question.
risk levels, risk assessments treat 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 a socially salient category. The relational response thus suggests aiming to reduce the crime risk of Black communities by alleviating criminogenic conditions of disadvantage. For instance, public policies that extend access to education [63], welfare [87], and affordable housing [28] 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 and the effects of this association. This entails not merely challenging stereotypes that link Blackness with crime, but also decriminalizing behaviors that were previously criminalized to subjugate minorities [15, 72].

The structural response suggests altering the structure of decisions to reduce the harmful consequences associated with being high risk to recidivate. By informing preventative detention decisions, risk assessments uphold the notion that the appropriate response to high-risk defendants is incarceration. Instead, we should reform policy 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 recidivism. If fewer people were subjected to decisions in which liberty and well-being depend on exhibiting low levels of crime risk, racial disparities in the distribution of risk would be less consequential. Most directly, such an approach could entail abolishing (or drastically reducing the scope of) pretrial detention, such that fewer people would 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 is to consider the potential role for algorithms in advancing relational and structural reforms. In some cases, this analysis will provide arguments against the use of certain algorithms for reform. For instance, because pretrial risk assessments naturalize racial disparities in risk that are the product of oppression and legitimize preventative detention, these algorithms conflict with the relational and structural responses. In other cases, however, this analysis will reveal new, 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 education [60], welfare [25], and affordable housing [91]. 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 [9, 66], 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 [5, 46], enable a systemic view of how the criminal justice system exacerbates racial inequalities [23, 39], and empower communities advocating for criminal justice reform [8, 21]—all of which would help to inform and enable structural responses.

In sum, substantive algorithmic fairness demonstrates how an expansive analysis of social conditions and institutions can lead to rigorous theories of social change, and how 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 suggest algorithmic interventions that escape the impossibility of fairness and that ameliorate oppression.
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 \[18\] and college admissions \[37\].

5.3 The Path Forward
Substantive algorithmic fairness offers a new direction for algorithmic fairness. It shifts the field’s concern away from formal mathematical models of “fair” decision-making and toward substantive evaluations of how algorithms can (and cannot) combat social hierarchies. In doing so, substantive algorithmic fairness brings the field in line with recent calls for algorithmic “justice” \[14, 43\], “equity” \[24\], and “reparation” \[26\]. Although there remains a role for formal evaluations of algorithmic decision-making, the field’s work should also focus on studying how algorithms can be incorporated into broader reform efforts to promote equality. Substantive algorithmic fairness thus requires different 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 in a principled manner, rather than a mandatory checklist. It cannot be reduced to an optimization problem. This lack of explicit prescription is not so much a limit 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 reforms that align with formal algorithmic fairness. There similarly is no straightforward mechanism for determining how to best incorporate algorithms into reform efforts. Future work is necessary to better understand the appropriate roles for algorithms in reform efforts, the conditions that facilitate effective algorithmic reforms, and how to allocate authority over algorithmic reforms.

Nonetheless, although it lacks a straightforward prescription for achieving reform, 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 only alternative to implementing fallible and biased algorithms is to fall back on even more fallible and biased human decision-makers \[12, 56, 69\]. 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” \[41\], “real
utopias” [90], and prison abolition [67], 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 [14, 45, 80]. 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 also leads to notions of algorithmic fairness that are inattentive to the lived realities of oppressed groups [89]. 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 promoting equality. It is therefore essential to consider whether algorithmic fairness possesses suitable conceptual and practical tools to guide reform in public policy and other domains. 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 for enhancing 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 productively guide efforts to pursue equality, we must alter its methodology to encompass more comprehensive conceptual and practical tools.

Substantive algorithmic fairness provides an alternative methodology that incorporates social hierarchies and the structure of decisions into the scope of analysis. As a result, substantive algorithmic fairness offers an escape from the impossibility of fairness and suggests new roles for algorithms in combating 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” [14, 43], “equity” [24], and “reparation” [26]. Although this reorientation shifts 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 for social change.

Although substantive algorithmic fairness does not yield a precise roadmap for reform, it presents a compass for linking incremental algorithmic reforms with visions of substantive equality. Substantive algorithmic fairness reveals that reformers do not face a binary choice 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 provides a method to elucidate the range of possible reforms, evaluate which reforms can best advance substantive equality, and consider how algorithms can 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 constrains opportunities for reform and 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

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson, “Roles for computing in social change,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020, Association for Computing Machinery: Barcelona, Spain. p. 252–260.

[2] Elizabeth Anderson, “Toward a Non-Ideal, Relational Methodology for Political Philosophy: Comments on Schwartzman’s Challenging Liberalism.” Hypatia, 2009. 24(4): p. 130-145.

[3] Elizabeth S. Anderson, “What is the Point of Equality?” Ethics, 1999. 109(2): p. 287-337.

[4] Julia Angwin and Jeff Larson, “Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say.” ProPublica, 2016. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, “Machine Bias.” ProPublica, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[6] Richard Arneson, “Egalitarianism.” The Stanford Encyclopedia of Philosophy, 2013. https://plato.stanford.edu/entries/egalitarianism/.

[7] Arnold Ventures, “Statement of Principles on Pretrial Justice and Use of Pretrial Risk Assessment.” 2019. https://craftmediabucket.s3.amazonaws.com/uploads/Arnold-Ventures-Statement-of-Principles-on-Pretrial-Justice.pdf.

[8] Mariam Asad, “Prefigurative Design as a Method for Research Justice.” Proceedings of the ACM on Human-Computer Interaction, 2019. 3(CSCW).

[9] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain, “Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment,” in Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018. Proceedings of Machine Learning Research: PMLR.

[10] Shima Baradaran, “Restoring the Presumption of Innocence.” Ohio State Law Journal, 2011. 72: p. 723-776.

[11] Solon Barocas, Moritz Hardt, and Arvind Narayanan, Fairness and Machine Learning. 2019: fairmlbook.org.

[12] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, “Fairness in Criminal Justice Risk Assessments: The State of the Art.” Sociological Methods & Research, 2018: p. 1-42.

[13] Reuben Binns, “Fairness in Machine Learning: Lessons from Political Philosophy,” in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, A. Friedler Sorelle and Wilson Christo, Editors. 2018, PMLR. p. 149–159.

[14] Matthew Le Bui and Safiya Umoja Noble, “We’re Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness,” in The Oxford Handbook of Ethics of AI, Markus D. Dubber, Frank Pasquale, and Sunit Das, Editors. 2020, Oxford University Press.
[15] Paul Butler, *Chokehold: Policing Black Men*. 2017: The New Press.
[16] California Legislature, “AB-13 Public contracts: automated decision systems.” 2021.
[17] Alexandra Chouldechova, “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data*, 2017. 5(2): p. 153-163.
[18] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan, “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*, A. Friedler Sorelle and Wilson Christo, Editors. 2018, PMLR: Proceedings of Machine Learning Research. p. 134--148.
[19] Alexia Cooper and Erica L. Smith, “Homicide Trends in the United States, 1980-2008.” U.S. Department of Justice, Bureau of Justice Statistics, 2011. https://www.bjs.gov/content/pub/pdf/htus8008.pdf.
[20] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, “Algorithmic Decision Making and the Cost of Fairness,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017. Halifax, NS, Canada: ACM.
[21] Sasha Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*. 2020: MIT Press.
[22] Kimberlé Williams Crenshaw, “Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law.” *Harvard Law Review*, 1988. 101(7): p. 1331-1387.
[23] Andrew Manuel Crespo, “Systemic Facts: Toward Institutional Awareness in Criminal Courts.” *Harvard Law Review*, 2015. 129: p. 2049-2117.
[24] Catherine D'Ignazio and Lauren F. Klein, *Data Feminism*. 2020: MIT Press.
[25] DataSF, “Keeping Moms and Babies in Nutrition Program.” 2018. https://datasf.org/showcase/datascience/keeping-moms-and-babies-in-nutrition-program/.
[26] Jenny L. Davis, Apryl Williams, and Michael W. Yang, “Algorithmic reparation.” *Big Data & Society*, 2021. 8(2).
[27] John Dewey, *Logic: The Theory of Inquiry*. 1938: Henry Holt and Company.
[28] Rebecca Diamond and Tim McQuade, “Who Wants Affordable Housing in Their Backyard? An Equilibrium Analysis of Low-Income Property Development.” *Journal of Political Economy*, 2019. 127(3): p. 1063-1117.
[29] William Dieterich, Christina Mendoza, and Tim Brennan, “COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.” *Northpointe Inc. Research Department*, 2016.
[30] Will Dobbie, Jacob Goldin, and Crystal S. Yang, “The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges.” *American Economic Review*, 2018. 108(2): p. 201-40.
[31] Cory Doctorow, “Algorithmic risk-assessment: hiding racism behind "empirical" black boxes.” Boing Boing, 2016. https://boingboing.net/2016/05/24/algorithmic-risk-assessment-h.html.
[32] EdBuild, “$23 Billion.” 2019. https://edbuild.org/content/23-billion/full-report.pdf.
[33] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. 2018: St. Martin's Press.

[34] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, “Certifying and Removing Disparate Impact,” in 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015. Sydney, NSW, Australia: ACM.

[35] Joseph Fishkin, *Bottlenecks: A New Theory of Equal Opportunity*. 2014: Oxford University Press.

[36] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp, “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*, 2016. 80: p. 38-46.

[37] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian, “The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making.” *Communications of the ACM*, 2021. 64(4): p. 136–143.

[38] Seeta Peña Gangadharan and Jędrzej Niklas, “Decentering technology in discourse on discrimination.” *Information, Communication & Society*, 2019. 22(7): p. 882-899.

[39] Sharad Goel, Justin M. Rao, and Ravi Shroff, “Precinct or prejudice? Understanding racial disparities in New York City’s stop-and-frisk policy.” *The Annals of Applied Statistics*, 2016. 10(1): p. 365-394.

[40] Abe Gong, “Ethics for powerful algorithms (1 of 4).” Medium, 2016. https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84.

[41] Andre Gorz, *Strategy for Labor*. 1967: Beacon Press.

[42] Stefan Gosepath, “Equality.” *The Stanford Encyclopedia of Philosophy*, 2021.

[43] Ben Green, “Putting the J(ustice) in FAT.” Berkman Klein Center Collection - Medium, 2018. https://medium.com/berkman-klein-center/putting-the-j-ustice-in-fat-28da2b8eae6d.

[44] Ben Green, *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future*. 2019: MIT Press.

[45] Ben Green, “The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.

[46] Ben Green and Yiling Chen, “Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments,” in Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

[47] Ben Green and Lily Hu, “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. 2018.

[48] Ben Green and Salomé Viljoen, “Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.
[49] Moritz Hardt, Eric Price, and Nati Srebro, “Equality of Opportunity in Supervised Learning,” in 30th Conference on Neural Information Processing Systems (NIPS 2016). 2016. Barcelona, Spain.

[50] Kamala Harris and Rand Paul, “Pretrial Integrity and Safety Act of 2017.” 115th Congress, 2017.

[51] Deborah Hellman, “Measuring Algorithmic Fairness.” Virginia Law Review, 2020. 106(4): p. 811-866.

[52] Anna Lauren Hoffmann, “Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse.” Information, Communication & Society, 2019. 22(7): p. 900-915.

[53] Abigail Z. Jacobs and Hanna Wallach, “Measurement and Fairness.” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021: p. 375–385.

[54] Jonathan Kahn, Race on the Brain: What Implicit Bias Gets Wrong about the Struggle for Racial Justice. 2017: Columbia University Press.

[55] Pratyusha Kalluri, “Don’t ask if artificial intelligence is good or fair, ask how it shifts power.” Nature, 2020. 583: p. 169.

[56] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein, “Discrimination in the Age of Algorithms.” Journal of Legal Analysis, 2019. 10: p. 113-174.

[57] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, “Inherent trade-offs in the fair determination of risk scores.” arXiv preprint arXiv:1609.05807, 2016.

[58] John Logan Koepke and David G. Robinson, “Danger Ahead: Risk Assessment and the Future of Bail Reform.” Washington Law Review, 2018. 93: p. 1725-1807.

[59] Lauren J. Krivo, Ruth D. Peterson, and Danielle C. Kuhl, “Segregation, Racial Structure, and Neighborhood Violent Crime.” American Journal of Sociology, 2009. 114(6): p. 1765-1802.

[60] Himabindu Lakkaraju, et al., “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. 2015, ACM: Sydney, NSW, Australia. p. 1909–1918.

[61] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, “How We Analyzed the COMPAS Recidivism Algorithm.” ProPublica, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[62] Katrina Ligett, “FAccT 2021 Keynote: In Praise of Flawed Mathematical Models.” 2021. https://www.youtube.com/watch?v=gZrZwF3XDBw.

[63] Lance Lochner and Enrico Moretti, “The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports.” American Economic Review, 2004. 94(1): p. 155-189.

[64] Catharine A. MacKinnon, “Substantive Equality: A Perspective.” Minnesota Law Review, 2011. 96: p. 1.

[65] Catharine A. MacKinnon, “Substantive equality revisited: A reply to Sandra Fredman.” International Journal of Constitutional Law, 2016. 14(3): p. 739-746.

[66] Sandra G. Mayson, “Bias In, Bias Out.” Yale Law Journal, 2019. 128(8): p. 2218-2300.
Allegra M. McLeod, “Prison Abolition and Grounded Justice.” *UCLA Law Review*, 2015. **62**: p. 1156-1239.

Merriam-Webster, “Methodology.” 2021. https://www.merriam-webster.com/dictionary/methodology.

Alex P. Miller, “Want Less-Biased Decisions? Use Algorithms.” *Harvard Business Review*, 2018. https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms.

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

Martha Minow, “Equality vs. Equity.” *American Journal of Law and Equality*, 2021. **1**: p. 167-193.

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

Naomi Murakawa, *The First Civil Right: How Liberals Built Prison America*. 2014: Oxford University Press.

Cathy O’Neil, “ProPublica report: recidivism risk models are racist.” mathbabe, 2016. https://mathbabe.org/2016/05/24/propublica-report-recidivism-risk-models-are-racist/.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations.” *Science*, 2019. **366**(6464): p. 447-453.

Rodrigo Ochigame, “The Long History of Algorithmic Fairness.” Phenomenal World, 2020. https://phenomenalworld.org/analysis/long-history-algorithmic-fairness.

Rodrigo Ochigame, Chelsea Barabas, Karthik Dinakar, Madars Virza, and Joichi Ito, “Beyond Legitimation: Rethinking Fairness, Interpretability, and Accuracy in Machine Learning,” in *Machine Learning: The Debates workshop* at the **35th International Conference on Machine Learning**. 2018.

Samir Passi and Solon Barocas, “Problem Formulation and Fairness,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019.

Christopher S. Porrino, “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.” 2017. https://www.nj.gov/lps/dcj/agguide/directives/ag-directive-2016-6_v3-0.pdf.

Julia Powles and Helen Nissenbaum, “The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence.” OneZero, 2018. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.

Inioluwa Deborah Raji and Joy Buolamwini, “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*. 2019. Honolulu, HI, USA: ACM.

Dina R. Rose and Todd R. Clear, “Incarceration, Social Capital, and Crime: Implications for Social Disorganization Theory.” *Criminology*, 1998. **36**(3): p. 441-480.

Richard Rothstein, *The Color of Law: A Forgotten History of How Our Government Segregated America*. 2017: Liveright Publishing Corporation.
[84] Robert J. Sampson, Jeffrey D. Morenoff, and Stephen Raudenbush, “Social Anatomy of Racial and Ethnic Disparities in Violence.” American Journal of Public Health, 2005. 95(2): p. 224-232.

[85] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi, “Fairness and Abstraction in Sociotechnical Systems,” in Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. Atlanta, GA, USA: ACM.

[86] Cass R. Sunstein, “Algorithms, Correcting Biases.” Social Research, 2019. 86(2): p. 499-511.

[87] Cody Tuttle, “Snapping Back: Food Stamp Bans and Criminal Recidivism.” American Economic Journal: Economic Policy, 2019. 11(2): p. 301-27.

[88] U.S. Supreme Court, “United States v. Salerno.” 481 U.S. 739, 1987.

[89] Sarah Myers West, “Redistribution and Rekognition.” Catalyst: Feminism, Theory, Technoscience, 2020. 6(2).

[90] Erik Olin Wright, Envisioning Real Utopias. 2010: Verso.

[91] Teng Ye, et al., “Using machine learning to help vulnerable tenants in New York City,” in Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies. 2019, ACM: Accra, Ghana. p. 248–258.