Algorithmic Fairness from a Non-ideal Perspective

Sina Fazelpour & Zachary C. Lipton

Carnegie Mellon University
sinaf@andrew.cmu.edu, zlipton@cmu.edu

January 9, 2020

Abstract

Inspired by recent breakthroughs in predictive modeling, practitioners in both industry and government have turned to machine learning with hopes of operationalizing predictions to drive automated decisions. Unfortunately, many social desiderata concerning consequential decisions, such as justice or fairness, have no natural formulation within a purely predictive framework. In efforts to mitigate these problems, researchers have proposed a variety of metrics for quantifying deviations from various statistical parities that we might expect to observe in a fair world and offered a variety of algorithms in attempts to satisfy subsets of these parities or to trade off the degree to which they are satisfied against utility. In this paper, we connect this approach to *fair machine learning* to the literature on ideal and non-ideal methodological approaches in political philosophy. The ideal approach requires positing the principles according to which a just world would operate. In the most straightforward application of ideal theory, one supports a proposed policy by arguing that it closes a discrepancy between the real and the perfectly just world. However, by failing to account for the mechanisms by which our non-ideal world arose, the responsibilities of various decision-makers, and the impacts of proposed policies, naive applications of ideal thinking can lead to misguided interventions. In this paper, we demonstrate a connection between the fair machine learning literature and the ideal approach in political philosophy, and argue that the increasingly apparent shortcomings of proposed fair machine learning algorithms reflect broader troubles faced by the ideal approach. We conclude with a critical discussion of the harms of misguided solutions, a reinterpretation of impossibility results, and directions for future research.[1]

1 Introduction

Machine Learning (ML) models play increasingly prominent roles in the allocation of social benefits and burdens in numerous sensitive domains, including hiring, social services, and criminal justice [4, 9, 3, 18]. A growing body of academic research and investigative journalism has focused attention on ethical concerns regarding algorithmic decisions [6, 10, 2], with many scholars warning that in numerous applications, ML-based systems may harm members of already-vulnerable communities [3, 12].

Motivated by this awareness, a new field of technical research addressing fairness in algorithmic decision-making has emerged, with researchers publishing countless papers aspiring to (i) formalize "fairness metrics"—mathematical expressions intended to quantify the extent to which a given algorithmic-based allocation is (un)just; and (2) mitigate "unfairness" as assessed by these metrics via modified data processing

[1] A version of this paper was accepted at the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) 2020.

procedures, objective functions, or learning algorithms [23, 60, 38, 29, 10, 22, 8]. However, progress has been hindered by disagreements over the appropriate conceptualization and formalization of fairness [7, 30, 21, 5].

The persistence of such disagreements raises a fundamental methodological question about the appropriate approach for constructing tools for assessing and mitigating potential injustices of ML-supported allocations. Importantly, any useful methodology must provide normative guidance for how a given agent ought to act in a world plagued by systemic injustices. Broadly speaking, justice requires apportioning benefits and burdens in accordance with each person's rights and deserts—giving individuals "their due" [34, 17]. Beyond this general framing, how can we offer more specific and practical guidance?

Drawing on literature in political philosophy, in Section 2, we distinguish between *ideal* and *non-ideal* methodological approaches to developing such normative prescriptions, and highlight three challenges facing the ideal approach. Then, in Section 3, we argue that most of the current technical approaches for addressing algorithmic injustice are reasonably (and usefully) characterized as small-scale instances of ideal theorizing. Next, in Section 4, we support this argument by demonstrating several ways that current approaches are, to varying extents, plagued by the same types of problems that confront naive applications of ideal theorizing more generally. Finally, drawing on these considerations, in Section 5, we provide a critical discussion of the real-world dangers of this flawed framing, and offer a set of recommendations for future work on algorithmic fairness.

2 Two Methodologies: Ideal *vs.* Non-Ideal

How should one go about developing normative prescriptions that can guide decision-makers who aspire to act justly in an unjust world? A useful distinction in political philosophy is between *ideal* and *non-ideal* modes of theorizing about the relevant normative prescriptions [56, 59, 57]. When adopting the ideal approach, one starts by articulating a conception of an ideally just world under a set of idealized conditions. The conception of the just world serves two functions: (i) it provides decision-makers with a *target* state to aspire towards [57]; and (ii) when suitably specified, it serves as an *evaluative standard* for identifying and assessing current injustices "by the extent of the deviation from perfect justice" [47, p. 216]. According to this perspective, a suitably-specified evaluative standard can provide decision-makers with normative guidance to adopt policies that minimize deviations with respect to some notion of similarity, thus closing the gap between the ideal and reality [1].

Non-ideal theory emerged within political philosophy as a result of a number of challenges to ideal modes of theorizing [20, 59]. We focus here on three challenges that motivate the non-ideal approach. A first set of issues arises when we consider the intended role of a conception of an ideally just world as an evaluative lens for diagnosing actual injustices. In the ideal approach, the conceptual framing of perfect justice determines whether some actual procedure or event is identified *as unjust* and if so, how that injustice gets represented [36, 41]. When this conception is impoverished, e.g., by failing to articulate important factors, it can lead to systematic neglect of injustices that were overlooked in constructing the ideal. Moreover, the static nature of ideal standards and the pursuant diagnostic lens can overlook the factors that give rise to injustice in the first place. This is because such standards identify injustices in terms of *the discrepancies* between the actual world and an ideally-just target state. However, the historical origins and dynamics of current injustices and the ongoing social forces that sustain them are typically absent from consideration. By obfuscating these *causal factors*, ideal evaluative standards can distort our understanding of current injustices.

According to a second challenge, employing a conception of an ideally just world as an evaluative standard

is *not sufficient* for deciding how actual injustices should be mitigated [54, 55]. This is because, from the standpoint of an ideal, any discrepancy between our imperfect world and that ideal might be interpreted naively as a cause of an actual injustice, and thus, any policy that aims to directly minimize such a discrepancy might be erroneously argued to be justice-promoting [1, 54]. Yet, the actual world can deviate from an ideal in multiple respects, and the same kind of deviation can have varied and complex causal origins [55]. Moreover, as the fair machine learning literature clearly demonstrates (see Section 5.2), simultaneously eliminating all discrepancies might be impossible. Thus, a coherent approach requires not only a mandate to eliminate discrepancies, but also guidance for determining which discrepancies matter in a given context. Crucially, policies that simply seek to minimize any perceived gap between the ideal and reality without consideration for the underlying causes may not only be ineffective solutions to current injustices, but can potentially exacerbate the problem they purport to address. For example, ideal theorizing has been applied to argue for race-blind policies (against affirmative action) [1]. From the perspective of an ideally just society as a race-blind one, a solution to current injustices "would appear to be to end race-conscious policies" [1, 4], thus blocking efforts devised to address historical racial injustices. Absent considerations of the dynamics by which disparities emerge, it is not clear that in a world where individuals have been racialized and treated differently on account of these perceived categories, race-blind policies are capable of bringing about the ideal [1].

Finally, a third challenge concerns the practical usefulness of the ideal approach *for current decision-makers*, given the type of idealized assumptions under which ideal theorizing proceeds. Consider, for example, the assumption of *strict compliance*, frequently assumed by ideal theorists as a condition under which the conception of an ideally just world can be developed. The condition assumes that nearly all relevant agents comply with what justice demands of them [48, 13]. The condition thus idealizes away situations where some agents fail to act in conformity with their ethical duties (e.g., the duty not to racially discriminate), or are unwilling to do so. The vision of a just world constructed under this assumption fails to answer questions about what we might reasonably expect from a decision-maker in the real world, where others often neglect or avoid their responsibilities [51, 14, 59].

In short, when used as lens for identifying current injustices, ideal modes of theorizing (1) can lead to systematic neglects of some injustices and distort our understanding of other injustices; (2) do not, by themselves, offer sufficient practical guidance about *what should be done*, sometimes leading to misguided mitigation strategies; and finally, (3) do not, by themselves, make clear *who, among decision-makers* is responsible for intervening to right specific injustices. As a result of these challenges to ideal modes of theorizing, a number of researchers in political philosophy have turned to non-ideal modes of theorizing. In contrast to the ideal approach, the non-ideal approach begins by identifying actual injustices that are of concern to decision-makers and that give rise to reasonable complaints on behalf of those affected by their decisions [1, 54]. Non-ideal theorizing can be seen as a trouble-shooting effort towards addressing these actual concerns and complaints. As Sen notes, this trouble-shooting aim distinguishes non-ideal modes of theorizing from ideal approaches that focus "on looking only for the simultaneous fulfilment of the entire cluster of perfectly just societal arrangements" [54, p. 218].

Anderson offers a succinct description of the non-ideal approach towards this trouble-shooting goal and what that approach requires:

> [Non-ideal theorists] ... seek a causal explanation of the problem to determine what can and ought to be done about it, and who should be charged with correcting it. This requires an evaluation of the mechanisms causing the problem, as well as responsibilities of different agents to alter these mechanisms [1, p. 22]

As noted by Anderson, there is still a crucial role for normative ideals within the non-ideal approach. But this role is importantly different from the roles assigned to ideals in the ideal approach [1, 6]. In the ideal approach, normative ideals are *extra-empirical*, in the sense that they set the evaluative standards against which actual practices are assessed, without themselves being subject to empirical evaluation. In contrast, in non-ideal theorizing, normative ideals act as *hypotheses* about potential solutions to identified problems. Viewed in this way, normative ideals are subject to revision in light of their efficacy in addressing the concerns and complaints that arise in practice. In the following sections, we show how the distinction can be put to work in understanding and addressing algorithmic injustice.

3 Work on Algorithmic Fairness as Small-scale Ideal Theorizing

In political philosophy, the distinction between ideal and non-ideal approaches typically refers to ways of understanding the demands of justice at large, and offering practical normative guidance to basic societal institutions for complying with these demands. While some researchers are beginning to discuss how the automation of decision making in consequential domains interacts with demands of justice at this large scale, most works on algorithmic fairness have the more restricted aim of assessing and managing various disparities that arise among particular demographic groups in connection with the deployment of ML-supported decision systems in various (often-allocative) settings. Nonetheless, in what follows, we show that the distinction between ideal and non-ideal approaches provides a fruitful lens for formulating strategies for addressing algorithmic injustices, even on this smaller scale (of an individual decision-maker). In this section, we argue that the dominant approach among current efforts towards addressing algorithmic harms can be seen as exercises in *small-scale* ideal theorizing.

3.1 Developing a Fairness Ideal

Works on algorithmic fairness typically begin by outlining a conception of a "fairness ideal". Dwork et al. [10, p. 215], for example, seek to "capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly" (see also Jung et al. [27]). Others envision the fair ideal at the group level. In nearly all cases, the groups of interest are those encompassing categories such as race, ethnic origin, sex, and religion. Following precedent in the United States Civil Rights Act, these groups are typically called *protected classes* or *protected groups* in the technical literature. According to one group-level conception of fairness, fair allocative policies and procedure are those that result in outcomes that impact different protected groups in the same way [60, 18]. In other cases, a fair state is taken to be one in which membership in a protected group is irrelevant or does not make a difference to the allocative procedure [29, 22]. According to another view, a treatment disparity might exist in a fair state, if it is justified by the legitimate aims of the distributive procedure [23, 38]. The endorsed fairness ideals have different provenances: in some cases, authors refer to historical legal cases, such as *Carson v. Bethlehem Steel Corp.* or *Griggs v. Duke Power*, to support their conception of fairness. In other cases, the ideal of fairness is derived from people's intuitive judgments about fair allocation [22, 27]. And less frequently, authors allude to works of political philosophers such as Rawls, which is cited to support the conception of individual fairness in Dwork et al. [10].

3.2 Specifying a Fairness Metric

Next, on the basis of their favored fairness ideal, researchers specify a quantitative evaluative standard—a "fairness metric"—for diagnosing potential allocative injustices and guiding mitigation efforts. Typically,

these fairness metrics take the form of mathematical expressions that quantify how far two among the protected groups are from *parity*. The magnitude of (dis)parity measured by a given fairness metric is taken to denote the degree of divergence from the ideal for which that metric is supposed to be a formal proxy.

Given their generality and abstract nature, fairness ideals do not fully determine the specific shape of fairness metrics. Accordingly, in addition to a fairness ideal, the construction of fairness metrics requires researchers to make further value judgments. For example, the ideal that membership in protected groups should be irrelevant to allocative decisions can be articulated in the language of statistics by requiring the outcome $\hat{Y}$ be independent (probabilistically) of the protected attributes A [18]. However, the same ideal can also be expressed in the language of causality, e.g., by requiring that the average causal effect of protected attributes A on $\hat{Y}$ be negligible [29]. Similarly, one can formalize the qualification that protected attributes can make a difference to outcomes when justified by the legitimate aims of allocative procedures in different ways. In the language of statistics, for example, one can require that while there may be some correlation between $\hat{Y}$ and A, the dependency must be screened off by the target variable, Y [23]. Framed in the language of causality, some attempt to formalize this fairness ideal in terms of a parity between the causal effect of A on $\hat{Y}$ along so-called *legitimate pathways* [38], where what counts as legitimate depends on the specific task and Y. Importantly, despite being motivated by the same ideal, such fairness metrics make different demands from the user and can result in different verdicts about the same case. In general, while statistical metrics can be formulated as functions of the joint distribution $P(Y, \hat{Y}, A, X)$, causal metrics additionally require the acquisition of a causal model that faithfully describes the data-generating processes and for which the desired causal effect is identifiable. Thus in some situations, statistical parity metrics may be estimable from data while the corresponding causal quantities may not be, owing to our limited knowledge of the data-generating process [42].

3.3 Promoting Justice by Minimizing Deviations from the Ideal

Finally, current approaches seek to promote fairness (or mitigate unfairness) by modifying ML algorithms to maximize utility subject to a parity constraint expressed in terms of the proposed fairness metric. Such fairness-enforcing modifications can take the form of interventions (i) in the pre-processing stage to produce "fair representations" (e.g., Kamiran and Calders [28]); (ii) in the learning stage to create "fair learning" (e.g., Zafar et al. [60]); or (iii) in the post-processing by adjusting the decision thresholds (e.g., Hardt et al. [23]). Crucially, however, in all cases, the range of solutions to algorithmic harms is limited to an intervention *to the ML algorithm*. Absent from consideration in these approaches is the broader context in which the "certifiably fair" model will be deployed. Recalling Anderson's critique [1, 22] of ideal approaches, neither the mechanisms causing the problem, nor the consequences of algorithmically-guided decisions, nor "the responsibilities of different agents to alter these mechanisms" are captured in any of these approaches.

4 Troubles with Ideal Fairness Metrics

If current works on algorithmic fairness pursue (small-scale) ideal theorizing, then we should expect these works to encounter the same types of challenges as those confronting ideal theorizing more generally. As explained above, according to critics, ideal modes of theorizing can (1) lead to systematic neglects of some injustices; and distort our understanding of other injustices. Such ideal evaluative standards (2) do not offer sufficient practical guidance and can lead to misguided mitigation strategies. What is more, they (3) fail to

delineate the responsibilities of current decision-makers in a world where others fail to comply with their responsibilities. Below, we consider each of these challenges in turn, and show that these same types of worries arise with respect to current works on algorithmic fairness.

4.1 Systematic Neglects of Rights

The identification of injustices in ideal theorizing is constrained by the underlying conceptual framing of normative ideals. If this conceptual framing is not sufficiently rich or comprehensive, we run the risk of overlooking many actual injustices. The ideals of fairness in literature on algorithmic fairness are predominantly expressed in terms of some type of parity among designated protected classes. Is this comprehensive enough to be sensitive to the types of injustices that would lead to legitimate complaints by those affected by ML-based allocations? We believe that the answer is negative. To see why, consider that assessing claims of injustice can require attending to different types of information. As noted by Feinberg [15, 16], in some cases, what is someone's due is determinable only in comparison to what is allocated to others or what would have been allocated to them had they been present. In other cases, an individual's just due is determinable independent of any comparison and solely by reference to how that individual should have been treated in light of her rights and deserts. An allocative procedure can thus result in *comparative* as well as *non-comparative* cases of injustice [15, 17, 37].

Yet, virtually all algorithmic fairness ideals are framed in *comparative terms*. This comparative focus renders these ideals insensitive to legitimate claims of non-comparative injustice. Consider from this perspective, a judge who treats all defendants equally, denying parole to them all regardless of the specifics of their cases. Here the defendants can feel aggrieved because of how they *should have been* treated from the perspective of the standards of retributive justice; the review process was based on legally irrelevant factors, infringing on defendants' rights to due process, and at least in some cases, the punishments were disproportionately harsh, potentially resulting in arbitrary incarceration. Indeed, such sentencing behaviour goes against Articles 9 and 11 of the Universal Declaration of Human Rights, cited throughout various documents concerning ethical design such as the *IEEE Ethically Aligned Design* and the Toronto Declaration [40]. Yet, this and other cases of non-comparative injustice in which an individual's rights and deserts have been ignored escape the purview of current fairness metrics.

The situation is troubling even with respect to *comparative* cases of injustice. This is because, due to their narrow focus, fairness metrics essentially take the set of protected classes to *exhaust* comparison classes that might matter from the perspective of justice and fairness. However, consider a case where the appraisal of an employee's performance is influenced by factors such as their weight or height, despite the irrelevance (in a causal sense) of such characteristics to that job [26, 50]. In this setting and from the perspective of comparative justice, height and weight *are* relevant categories. The complete reliance of such metrics on the particular specification of relevant comparison groups limits their adequacy in this regard. Indeed, unconstrained by these demands of comparative justice, algorithmic-based decisions might result in the creation of new "protected groups".

4.2 Distortion of the Harms of Discrimination

From the perspective of current fairness ideals, any divergence from the ideal of parity among protected classes (potentially subject to certain qualifications) is identified as a case of unfairness. Accordingly, the fairness metrics based on these ideals often have the property of being *anonymous* or *symmetric*; whether a distribution of benefits and burdens is fair does not depend on who the affected individuals or groups are. In

certain contexts and for certain purposes, anonymity is a desirable property. Quantitative metrics of *income inequality* are required to be anonymous, for example, because "from an ethical point of view, it does not matter who is earning the income" [49]. Unlike the case of income inequality, however, evaluating fairness claims requires going beyond the observation *that* some disparity exists [24]. We need to know *why* the disparity exists and to understand "the processes that produce or maintain it" [1, 18]. This knowledge is required to determine a coherent course of action, and yet it does not inform any of the mitigation strategies in the standard fair machine learning tool-kits, making them unsuitable for off-the-shelf application.

Consider, for example, the very different mechanisms giving rise to disparities in representation between (white and east Asian) vs (white and black) students in US higher education. In the former case, the disparity (appearing to favor Asian students) emerges despite historical and institutional discrimination. In the latter, the disparity stems from well-documented historical and institutional discrimination. However, both represent violations of demographic parity [44]. A naive ideal approach may suggest that in both cases, the disparity requires alterations in admissions policies to enforce the parity across all groups we might expect in our ideal. A more nuanced non-ideal approach might recognize the differences between these two situations. In the literature on fair ML, approaches that incorporate knowledge of demographic labels are colloquially referred to as "fairness through awareness". However, as demonstrated above, awareness of demographic membership alone is too shallow to distinguish between these two situations. Instead, we require a deeper awareness, not only of demographic membership but of the societal mechanisms that imbue demographic membership with social significance in the given context and that give rise to existing disparities.

While this is especially problematic for statistical metrics that neglect the provenance of the observed data, recently-proposed causal approaches, including those formalizing fairness in terms of average causal effect or the effect of treatment on the treated, are similarly insufficient for capturing when a given disparity is reflective of discrimination, let alone whose discrimination it might reflect or providing guidance as to when the current decision-maker has a responsibility or license to intervene. Importantly, these causal methods typically address the problem of mediation analysis, adopting the perspective of an auditor seeking to explain the mechanisms by which the protected trait influences a model's prediction. Missing however, is a coherent theory for how to relate those mechanisms to the responsibilities of the current decision-maker, or any accounting of the causal mechanisms by which a proposed intervention may impact the social system for better or worse.

4.3 Insufficient Insights and Misguided Mitigation

As noted in the previous section, current mitigation strategies are guided by the idea that justice is promoted by intervening on ML algorithms to minimize disparities detected by a given metric. Insofar as the underlying causes of preexisting disparities and the consequences of proposed policies are ignored, however, these mitigation techniques might have adverse effects. As one example, consider a series of proposed approaches that Lipton et al. [31] denote *disparate learning processes* (DLPs). These techniques are designed to jointly satisfy two parities, blindness and demographic parity (e.g., Zafar et al. [60]). However, as Lipton et al. [31] (2018) show, DLPs are oblivious to the underlying causal mechanisms of potential disparities and in some cases, DLPs achieve parity between protected classes (e.g., genders) by giving weight to the irrelevant proxies, (e.g., hair length). Using real-world data from graduate admissions to a computer science program, they showed that prohibited from considering gender directly, a DLP would pick up on proxies such as the subfield of interest. In order to achieve parity, the DLP must advantage those applicants that appear (based on their non-protected attributes) to be more likely to be women, while disadvantaging those that are more

likely to be men. Thus, the DLP satisfies demographic parity by advantaging those pursuing studies in sub-fields chosen historically by more women (e.g., human-computer interaction) while disadvantaging those pursuing studies that are currently more male-dominated (e.g., machine learning). While the DLP achieves overall demographic parity, women in fields that already have greater parity receive the benefit, while women in those precise fields that most want for diversity would actually be penalized by the DLP.

Stepping back from a myopic view of the statistical problem and these arbitrarily-chosen deviations (the fairness metrics) from an ideal, when we consider the impact of a deployed DLP on a broader system of incentives, it becomes clear that the DLP risks amplifying the very injustices it is intended to address.

In addition to the non-comparative harm of making decisions on irrelevant grounds, the supposed remedy can reinforce social stereotypes, e.g., by incentivizing female applicants towards only those fields where they are already well represented (and away from others). Similarly, in simply seeking to minimize the disparity detected by fairness metrics, current metrics neglect considerations about whether the enforced parity might in fact result in long term harms [32].

4.4 Lack of Practical Guidance

Finally, consider that the type of unjust disparities often faced in a given allocation context correspond to events potentially unfolding over decades. Current approaches to algorithmic fairness seek to address *"is there discrimination?"* but leave open the questions of *"who discriminated?"* and *"what are the responsibilities of the current decision-maker?"* If sensitive features influence education, which in turn influences employment decisions, then to what extent does the causal effect reflect the discrimination of the education system compared to that of the employer? The answer to this question is not straightforward and requires considerations not captured in the entries of confusion matrices. While identifying statistical disparities may be valuable unto itself, e.g., as a first step to indicate particular situations that warrant investigation, it provides little moral or legal guidance to the decision-maker. While the influence of protected attributes on predictions may reflect injustice, providing normative guidance requires identifying not only what would constitute a just world but also what constitute just decisions in the actual world, with its history of injustice.

5 Discussion

5.1 If not Solutions, then Solutionism?

Even as the mitigation strategies arising from the recent technical literature on fair machine learning fail to offer practical guidance on matters of justice, they have not failed to deliver in the marketplace. From the perspective of stakeholders caught in the tension between (i) the potential profit to be gained from deploying machine learning in socially-consequential domains, and (ii) the increased scrutiny of a public concerned with algorithmic harms, these metrics offer an alluring solution: continue to deploy machine learning systems per the status quo, but use some chosen parity metric to claim a certificate of fairness, seemingly inoculating the actor against claims that they have not taken the moral concerns seriously, and weaponizing the half-baked tools produced by academics in the early stages of formalizing fairness as a shield against criticism.

In socially-consequential settings, requiring caution or even abstention (from applying ML) such as criminal justice and hiring, fair ML offers an apparent academic stamp of approval. Notable recent examples include

the IBM fairness 360 toolkit, which offers fairness metrics and corresponding mitigation strategies as an open-source software service that claims to be able to "examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle" [25]. Using just one parity metric (demographic parity), algorithmic hiring platform Pymetrics, Inc. claims that their system is "proven to be free of gender and ethnic bias" [46].

The literature on fair machine learning bears some responsibility for this state of affairs. In many papers, these fairness-inspired parity metrics are described as *definitions of fairness* and the resulting algorithms that satisfy the parities are claimed axiomatically to be *fair*. While many of these *metrics* are useful diagnostics, potentially alerting practitioners to disparities warranting further investigation, the looseness with definitions creates an opening for stakeholders to claim compliance, even when the problems have not been addressed. Lacking the basic primitives required to make the relevant moral distinctions, when blindly optimized, these metrics are as likely to cause harm as to mitigate it. Thus current methods produced by the fair ML community run the risk of serving as *solutionism* if not as solutions [53].

5.2 Re-interpreting Impossibility Results

An additional benefit of viewing fairness in ML through the lens of non-ideal theorizing in political philosophy is that it gives a new perspective for parsing the numerous impossibility results [30, 7] famously showing that many statistical fairness metrics are irreconcilable, presenting inescapable trade-offs. These results are sometimes misinterpreted as communicating that *fairness is impossible*. However, through the non-ideal lens, these impossibility theorems are simply a frank confirmation of the fact that we do not live in an ideal world. The inputs to statistical fairness metrics include four groups of variables: the covariates X, the group membership A, the label Y, and the classification $\hat{Y}$. The distribution over these variables at a given point in time is the consequence of the complex dynamics of an unjust society constituted of many decision-making agents. Of these, the current decision-maker has control only over their own predictions $\hat{Y}$. That various metrics/parities cannot be satisfied simultaneously merely by setting the values taken by $\hat{Y}$ indicates only that our present decision-maker cannot *through their actions alone* bring about the immediate end to all disparity, even as viewed locally through the variables that their individual decisions concern.

One potential contribution of ML impossibility theorems to philosophy is that they make evident an often-overlooked shortcoming with the ideal approach. These impossibility results make clear that in general, if we start from a non-ideal world, no set of actions (by a single agent) can instantaneously achieve the ideal world in every respect. Moreover, matching the ideal in a particular respect may only be possible at the expense of widening gaps in others. Thus this naive form of an ideal approach appears to be fundamentally under-specified. If matching the ideal in various respects simultaneously is impossible, then we require, in addition to an ideal, a basis for deciding which among competing discrepancies to focus on. In this manner, the impossibility results in fair ML provide a novel lens to approach the philosophical debate about the extent to which normative theorizing on matters of justice can proceed in isolation from empirical socio-historical facts [55, 13].

While characterizing disparities and understanding the fundamental trade-offs among them may be valuable work, this work cannot by itself tell us what to do. The pressing issue in determining how to act justly is not how to optimize a given metric but how to make the determination of what, in a given situation, should be optimized in the first place.

5.3 Towards a Non-Ideal Perspective

Even if the reader finds the case against the ideal approach compelling, there remains a pragmatic question of what precisely a non-ideal approach might look like in practice. To begin, non-ideal theorizing about the demands of justice is a *fact-sensitive* exercise. Offering normative prescriptions to guide actions requires understanding the relevant causal mechanisms that (i) account for present injustices; and (ii) govern the impact of proposed interventions.

Empirical understanding of the problem:

Developing causal models for understanding social dynamics that cause and maintain particular injustices requires extensive domain-knowledge as well as numerous value judgements about the relevance and significance of different aspects of the domain of interest. Choices must be made about what abstractions are reasonable, which simplifying assumptions are justified, and what formalizations are appropriate. Inevitably, these choices, embedded in design and modeling, raise *coupled ethical-epistemic* questions [58, 45]. Consider, for instance, choices that might be made in understanding the causes of racial injustice in a particular allocative domain and a specific social setting. Aside from the challenge of understanding the concept of race [35, 33], research in psychology and sociology shows racial classification and identification to be dynamic categories that are shaped by a variety of socioeconomic factors such as unemployment, incarceration, and poverty [11, 43, 19]. Appreciating the complex and dynamic nature of race and the perception thereof is thus not only of ethical import; it also has important epistemic implications for formal models of racial injustice, as it shapes how "race" as an attribute should be understood and what causal relation it might bear to other factors of interest.

Empirically-informed choice of treatment:

Deployment of predictive models—whether those that simply maximize utility or those that maximize utility subject to some "fairness" constraint—constitutes a social intervention. As mentioned above, most existing approaches to fair ML consist only of modifying the data processing procedures or the objective functions. Crucially, the evaluation of these interventions is *local* and *static*: the evaluation is local insofar as it concerns the impact of the intervention only on that particular predictive model's statistics (i.e., its accuracy and various fairness metrics). The accompanying literature seldom considers the broader impacts of deploying such models in any particular social context. Moreover, the evaluation is typically static, ignoring the longer-term dynamics of proposed policies. When authors have attempted dynamic evaluations, the results have sometimes contraindicated proposed mitigation strategies [32].

In contrast, a non-ideal approach to offering normative guidance should be based on evaluating the situated and system-wide (involving not just the predictive model but also the broader social context, actors, and users) and dynamic (evolving over longer periods) impact of potential fairness-promoting interventions.

Once more, we must face difficult questions and make value judgments. As some authors have noted, for instance, unjust circumstances can naturally arise as a result of seemingly benign initial conditions [52, 39]. To determine how to act, a coherent framework is needed for understanding when is it desirable or permissible for a given decision-maker to intervene. Importantly, we stress that the appropriate judgments simply cannot be made based on the reductive $(X, A, Y\ \hat{Y})$ description upon which most statistical fair ML operates. Developing a coherent non-ideal approach requires (for the foreseeable future) human thought, both to understand the social context and to make the relevant normative judgments.

6 Conclusion

Approaching the issue of algorithmic fairness from a non-ideal perspective requires a broadening of scope beyond parity-constrained predictive models, and considering the wider socio-technological system consisting of human users, who informed by these models, make decisions in particular contexts and towards particular aims. Effectively addressing algorithmic harms demands nothing short of this broader, human-centered perspective, as it enables the formulation of novel and potentially more effective mitigation strategies that are not restricted to simple modifications of existing ML algorithms.

Acknowledgements

Many thanks to David Danks, Maria De-Arteaga, and our reviewers for helpful discussions and comments. Funding was provided by Social Sciences and Humanities Research Council of Canada (No. 756-2019-0289) and the AI Ethics and Governance Fund.

References

[1] Elizabeth Anderson. *The Imperative of Integration*. Princeton University Press, Princeton, 2010.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, 2016.

[3] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *California Law Review*, 2016.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments. *Sociological Methods & Research*, 2018.

[5] Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. In *Fairness, Accountability and Transparency (FAT*)*, 2018.

[6] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Conference on Human Factors in Computing Systems (CHI)*, 2019.

[7] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR*, abs/1610.0, 2016.

[8] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[9] Kate Crawford and Jason Schultz. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, 2014.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Innovations in Theoretical Computer Science (ITCS)*, New York, NY, USA, 2012.

[11] Charles R. Epp, Steven Maynard-Moody, and Donald P. Haider-Markel. *Pulled over: how police stops define race and citizenship*. University of Chicago Press, 2014.

[12] Virginia Eubanks. *Automating inequality: how high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[13] Colin Farrelly. Justice in ideal theory: A refutation. *Political Studies*, 2007.

[14] Joel Feinberg. Duty and Obligation in the Non-Ideal World. *Journal of Philosophy*, 1973. doi: 10.2307/2025007.

[15] Joel Feinberg. Noncomparative justice. *The Philosophical Review*, 1974. doi: 10.2307/2183696.

[16] Joel Feinberg. *Rights, Justice, and the Bounds of Liberty: Essays in Social Philosophy*. Princeton University Press, Princeton, 2014.

[17] Fred Feldman. *Distributive Justice*. Oxford University Press, Oxford, 2016.

[18] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Knowledge Discovery and Data Mining (KDD)*, New York, NY, USA, 2015.

[19] Jonathan B. Freeman, Andrew M. Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. Looking the Part: Social Status Cues Shape Race Perception. *PLoS ONE*, 2011.

[20] William A Galston. Realism in political theory. *European Journal of Political Theory*, 2010. doi: 10.1177/1474885110374001.

[21] Bruce Glymour and Jonathan Herington. Measuring the Biases That Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms. In *Conference on Fairness, Accountability, and Transparency (FAT*)*, New York, NY, USA, 2019.

[22] Nina Grgić-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In Sheila A McIlraith and Kilian Q Weinberger, editors, *Association for the Advancement of Artificial Intelligence (AAI)*, 2018.

[23] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[24] Deborah Hellman. *When is discrimination wrong?* Harvard University Press, Cambridge, 2008.

[25] IBM. Ai fairness 360 open source toolkit, 2019. URL https://aif360.mybluemix.net/.

[26] Timothy A. Judge and Daniel M. Cable. The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model. *Journal of Applied Psychology*, 2004.

[27] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and Enforcing Subjective Individual Fairness. *arXiv*, 2019.

[28] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 2012.

[29] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[30] Jon M Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

[31] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. Does Mitigating ML's Impact Disparity Require Treatment Disparity? In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[32] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning (ICML)*, 2018.

[33] Ron Mallon. 'Race': Normative, Not Metaphysical or Semantic. *Ethics*, 2006.

[34] David Miller. Justice. In Edward N Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, 2017.

[35] Charles W. Mills. *Blackness visible: essays on philosophy and race*. Cornell University Press, 1998.

[36] Charles Wade Mills. "Ideal Theory" as Ideology. *Hypatia*, 2005.

[37] Phillip Montague. Comparative and Non-Comparative Justice. *The Philosophical Quarterly*, 1980.

[38] Razieh Nabi and Ilya Shpitser. Fair Inference on Outcomes. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[39] Cailin O'Connor. *The Origins of Unfairness*. Oxford University Press, Oxford, 2019.

[40] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. Technical report, IEEE, 2017.

[41] Carole Pateman and Charles Wade Mills. *Contract and Domination*. Polity Press, Cambridge, 2007.

[42] Judea Pearl. *Causality*. Cambridge university press, 2009.

[43] Andrew M Penner and Aliya Saperstein. How social status shapes race. *Proceedings of the National Academy of Sciences (PNAS)*, 2008.

[44] Nancy S. Petersen and Melvin R. Novick. An Evaluation of Some Models for Culture-Fair Selection. *Journal of Educational Measurement*, 1976.

[45] Robert N. Proctor and Londa Schiebinger. *Agnotology: The Making and Unmaking of Ignorance*. Stanford University Press, 2008.

[46] Pymetrics, Inc. Matching talent to opportunity, 2019. URL https://www.pymetrics.com/employers/.

[47] John Rawls. *A Theory of Justice*. Harvard University Press, 1999.

[48] John Rawls. *Justice as Fairness: A Restatement*. Harvard University Press, Cambridge, 2001.

[49] Debraj. Ray. *Development economics*. Princeton University Press, Princeton, 1998.

[50] Cort W. Rudolph, Charles L. Wells, Marcus D. Weller, and Boris B. Baltes. A meta-analysis of empirical studies of weight-based bias in the workplace. *Journal of Vocational Behavior*, 2009.

[51] Tamar Schapiro. Compliance, Complicity, and the Nature of Nonideal Conditions. *Journal of Philosophy*, 2003.

[52] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1971.

[53] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Fairness, Accountability, and Transparency (FAT*)*, 2019.

[54] Amartya Sen. What Do We Want from a Theory of Justice? *Journal of Philosophy*, 2006.

[55] Amartya Sen. *The Idea of Justice*. Harvard University Press, Cambridge, 2009.

[56] A. John Simmons. Ideal and Nonideal Theory. *Philosophy & Public Affairs*, 2010.

[57] Zofia Stemplowska and Adam Swift. Ideal and Nonideal Theory. In *The Oxford Handbook of Political Philosophy*. Oxford University Press, 2012.

[58] Nancy Tuana. Leading with ethics, aiming for policy: new opportunities for philosophy of science. *Synthese*, 2010.

[59] Laura Valentini. Ideal vs. Non-ideal Theory: A Conceptual Map. *Philosophy Compass*, 2012.

[60] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *World Wide Web (WWW)*, 2017.