The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning

Ben Green [*1] **Lily Hu** [*1]

Abstract

Even as machine learning has expanded into the realm of social decision-making, where concerns of bias and justice often rise above those of efficiency and accuracy, the field has remained committed to standard ML techniques that conceive of fairness in terms of statistical metrics and rely heavily on historical data as accurate and neutral representations of the world. So long as the field conforms to these methods and believes it can optimize systems according to universal notions of fairness, machine learning will be ill-suited to address the fundamentally political and ethical considerations at stake when deploying algorithms in the public sphere. The design and adoption of machine learning tools in high-stakes social contexts should be as much a matter of democratic deliberation as of technical analysis.

1. Introduction

When deployed in many human contexts, machine learning distills social processes into quantifiable features and metrics whose statistical correlations can be harnessed to predict behavior. It should be no surprise then, that in dealing with issues of fairness, the field of has followed a similar framework, adopting various methodologies that reify fairness as social concept into fairness as satisfiable technical criterion. The assumed legitimacy of this substitution relies upon the naïve, and perhaps unrecognized, assumption that guaranteeing technical conceptions of fairness is sufficient to achieve social ideals of fairness. We argue that fair machine learning, by striving to satisfy these purported "definitions" of fairness that live inside a model's limited environment, has committed to an abstraction of fairness that evades critical social and normative analysis and thus fails to correspond to concerns of fairness that carry genuine social, political, and moral weight. So long as the epistemologies and methodologies of fair machine learning continue to operate under this mythic equivalence between technical and social notions of fairness, the field will be ill-equipped to address the demands of justice that originally impelled its development.

2. Theories of Fairness

"Unfairness" is diagnosable in context: we know it when we see it, or at least we think we do. But "fairness" untethered and freestanding is hazy—it requires reference to values, principles, and commitments that themselves typically have substance only in specificity. Nevertheless, the fair machine learning community has appeared to agree on two fundamental assumptions about the nature of "fairness" as a social ideal: 1) There exists a reductive definition of fairness that may be presented in terms of a domain-general procedural or statistical guideline, and 2) This definition can be operationalized such that, so long as the chosen fairness criteria are satisfied, the resulting procedures and outcomes of the system are necessarily fair. Any question regarding the legitimacy of the project of fair machine learning must first confront these assumptions—the field's underlying "theory of fairness." In this section, we consider the two primary technical approaches to achieving fairness in machine learning—the procedural account and the statistical account—and argue that both methodologies, in different ways, enact a narrow vision of fairness in machine learning that is structured and delimited by these unexamined assumptions, and as a result, fail to correspond to substantive notions of fairness in the real world.

2.1. The Procedural Account: Fairness as adherence to a procedural maxim

Since "fairness" appears to have commonly agreed-upon colloquial usage and meaning, one may believe that the concept ought to be definable in the form of a general procedural principle that guides the proper functioning of a system. On this account, faithful accordance with the principle ensures the fairness of the machine learning task and also

[*] Equal contribution [1] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, U.S.A. Correspondence to: Lily Hu <lilyhu@g.harvard.edu>, Ben Green <bgreen@g.harvard.edu>.

Presented at the Machine Learning: The Debates workshop at the 35^{th} International Conference on Machine Learning, Stockholm, Sweden. Copyright 2018 by the author(s).

certifies the outcomes generated by the procedure as fair. As an example, in their classic work on fairness in classification, Dwork et al. present a framework that is grounded in the maxim that "similar people should be treated similarly" (2012). Here, the procedural account attempts to resolve the tension between the general principles of fairness prevailing in ethics and the particular operationalizations of fairness required in engineering by replacing the similarity maxim with a similarity metric. This technical sleight-of-hand asserts that fairness is reducible to constructing mathematical summaries of individuals' attributes that admit comparison between persons on a single standardized scale of similarity.

But what, exactly, is any claim to similarity based on? Insofar as fair treatment refers to an equality of *some* sort, the word "similar" is capacious enough to account for almost any disagreement people may have about the substantive demands of fairness. Nodding together, we may still find ourselves in conflict: issues of "fairness" cannot be resolved by asking who are similar—this is the task proposed and pursued by Dwork et al.—but engaging in discourse about what features, attributes, and conditions within particular contexts *make* individuals similar. On this matter, the similarity maxim and all other leading procedural accounts of fairness in machine learning, are silent.

By foregrounding a chosen procedure and its associated mathematical properties, the field seeks a conception of fairness that is removed from the social and historical context of the larger system within which that procedure is embedded. Such a sanitized notion of fairness is of course mythic, and despite attempting to avoid engaging with substantive questions of fairness, the practice is nevertheless still founded on undisclosed normative positions about what is or is not fair. Any procedural maxim (either a motivating toy example or a definition within a formal model) is grounded in value-laden assertions about a system's purpose, individuals' entitlements, and the relevant criteria for decision-making. Reliance on these silent normative assumptions is especially misleading because the field employs the language of procedural adherence to project a sense of certainty, objectivity, and stability about judgments of fairness that, in reality, are always under contest.

2.2. The Statistical Account: Fairness as satisfaction of balanced statistical metrics

Under statistical notions of fairness, it is possible to assess a particular tool by appealing to statistical summaries of the outcomes that it issues. Of particular interest are disparities that may arise along protected group lines in the tool's distribution of classifications and errors. The statistical account asserts that a machine learning task can be deemed fair if it satisfies various types of "balance," defined as approximate equality across these outcome-based statistics. For each of these types of statistical balance, the field has attached a label: "equal opportunity" describes a machine learning tool that equalizes true positive rates across groups (Hardt et al., 2016); "disparate mistreatment" refers to equalities of error probabilities (Zafar et al., 2017). Such benchmarks have become ubiquitous in discussions of fair machine learning ever since the ProPublica-Northpointe dispute about the nature of bias within the COMPAS recidivism risk tool was found to boil down to conflicting statistical definitions for fairness (Angwin et al., 2016; Chouldechova, 2017; Dieterich et al., 2016; Kleinberg et al., 2016).

While the field's practice of labeling these particular metrics is benign, reifying fairness as *constituted* by satisfaction of the statistical constraints is mistaken. Reductive commitments to statistical parities of various types limit the realm of justice and fairness to one of merely adjudicating comparative claims of treatment and outcomes across groups, which represent only one relevant criterion for assuring fairness. When issued without deeper analysis into other demands of justice, group-based statistical constraints can elide conversation about what individuals and groups are owed *independently* of how their counterparts are treated. Here, the COMPAS debate demonstrates one consequence of an undue fixation on fairness as statistical parity: by focusing only on parity or imparity in outcomes between black and white defendants as potential sources of fairness or unfairness, researchers tend to neglect other key aspects of the tool's behavior and performance that are relevant to its fair deployment. For one, COMPAS's high misclassification rates across both groups is surely cause for concern in itself, yet such an independent claim to justice is not captured by the statistical framework (Angwin et al., 2016; Dressel & Farid, 2018). Nor does the statistical account permit assessments of whether the purpose that the tool serves is itself a just one, further casting into suspicion any certification of that tool as "fair."

2.3. Fairness: No single criterion or category of criteria

The procedural and statistical accounts capture important considerations of fairness: impartiality of process on the one hand and protection from adverse impact on the other. While each theory is necessary, neither is sufficient on its own. The procedural account's singular focus on designing machine learning that is internally-consistent with a fair guideline blinds us to the various questions of justice that bear on its larger decision-making context and the social system within which the tool is embedded. Similarly, by relying on outcome-based data to determine violations of fairness, the statistical account fails to ensure that non-observational criteria of justice such as individuals' entitlements to fair procedure are respected.

Combining both procedural and statistical considerations would yield a more holistic view of fairness in machine learning, but a deeper flaw lies in the meta-methodology of pursuing a "solution" to fairness that is limited to satisfaction of technical definitions without referring to a broader analysis of social and moral context. The field's quest for a fairness that may be encapsulated by general guidelines or metrics leaves little room for engagement with questions about a task's purposes and obligations, an individual's luck and desert, and a system's social and economic conditions—all morally salient considerations that can be brought to bear only through an honest and thorough analysis of the demands of fairness and justice.

The view of fairness as a metric can also encourage a conflation of fairness as a mathematical property and fairness as a broader social ideal. This confusion has played a large part in the field's interpretations of Kleinberg et al. and Chouldechova's "impossibility results," which exposed the fundamental limitations of simultaneously satisfying multiple common statistical outcome-based notions of fairness (Chouldechova, 2017; Kleinberg et al., 2016). For one, the language of "impossibility of fairness" is exemplary of the community's current susceptibility to the reification fallacy: What is strictly impossible here is the perfect balance of three specific group-based statistical measures. Labeling a particular incompatibility of statistics as an impossibility of fairness generally is mistaking the map for the territory.

But more important and more pernicious than the community's misnomer of the finding lies in its potential interpretation as offering, by way of mathematical proof, a path out of grappling with the ethical obligations of our technologies. The framing of "impossible" can be seen as inviting the community to view unfairness with resigned inevitably, under the view that the pursuit of fairness and social justice more broadly is a fundamentally arbitrary venture—one in which, without a clearly optimal solution, all outcomes become equally legitimate and grappling with moral considerations is ultimately a matter of relativist opinion. Highlighting the irreconcilability of various "fair" statistical constraints validates ProPublica's analysis of COMPAS while simultaneously absolving Northpointe. Instead, rather than reveal that there are no *right* answers, Kleinberg et al. and Chouldechova show that there are no *easy* answers. The community has correctly recognized that fairness is a fundamentally hard problem, but misdiagnoses why. Fair machine learning is hard not because of statistical or computational challenges, but because striving for fairness is ultimately a process of continual social negotiation and adjudication between competing needs and visions of the good.

3. Machine Learning Methodologies

In addition to evaluating formal methodologies for achieving fairness, the field must also interrogate whether and how its fundamental practices are equipped to make fair decisions.

3.1. Machine learning's reliance on data and metrics can distort deliberative processes

Although there is nothing inherently unfair about utilizing data and metrics to make decisions, there is a danger that relying only on these types of information will distort the values inherent to the task at hand by granting undue weight to considerations and values that are quantified at the expense of those that are not. This concern is especially salient when considering the application of machine learning in social decision-making processes, since many aspects of society have been measured only in limited ways and in many cases resist quantification.

In practice, reliance on quantitative data makes machine learning prone to limiting and reweighting the many considerations that factor into complex decisions in unexpected and potentially undesirable ways. Determining sentences within the criminal justice system, for example, requires balancing several goals: incapacitating offenders from committing further crimes, deterring others from committing similar crimes in the future, rehabilitating offenders, and delivering just punishment. But only one of these factors—incapacitation, via recidivism—has been rigorously measured in a manner conducive to machine learning. Thus, while introducing the COMPAS risk tool into judicial decision-making may provide judges with better assessments of recidivism risk, it may also have the unintended consequence of framing sentences around recidivism risk in a manner that leads judges to place greater emphasis on incapacitation as a goal of sentencing.

This example highlights a significant challenge of using machine learning to fairly adjudicate complex decisions. If, in the case of sentencing, fairness requires the delicate balancing of several societal goals, incorporating a tool that privileges incapacitation will lead to changes in sentencing that, in effect, represent significant shifts in policy and jurisprudence. Because these shifts emerge as unintended consequences of deploying an algorithm, they are likely to take hold with neither formal review nor public discussion. In this manner, algorithms have the potential to distort the values underlying laws and policies that, in principle, society has collectively determined to be fair, and to do so without proper democratic input.

3.2. Machine learning narrows judgments about fairness and entrenches historical discrimination

Recognizing that people are subject to cognitive limitations and personal prejudices, machine learning has been promoted as a useful tool that can improve the accuracy and fairness of human decision-making. The field of fair machine learning strives to ensure that these algorithms do not reproduce the biases that plague human decision-making. But this diagnosis of how to make decisions fairer, although well-intended, is limited by its focus only on societal bias that arises due to the behavior of individual actors. Many forms of discrimination and inequality are produced not by individual people making biased judgments about other people, but through laws and institutions that systematically benefit one group over another. As a result, much of the field misunderstands and vastly understates the extent of the problem of bias and unfairness in society and, hence, in data about society.

A thoughtful assessment of whether a decision is fair requires multiple scopes of analysis; taking on a single perspective is never sufficient to deem a decision fair. For example, while the debate about COMPAS has been framed in terms of competing statistical notions of fairness, it can also be analyzed as a debate about the different lenses through which one ought to analyze fairness. Within the narrow interest of predicting recidivism, the algorithm satisfies one "definition" of fairness (calibration). But a discussion of fairness in isolation of its broader social context is underspecified: what may appear to be fair under the narrow frame of predicting recidivism may be deeply unfair within a broader historical and cultural context. After all, the empirical finding that blacks recidivate at higher rates than whites (which leads to the conflict between calibrated predictions and error rate balance) is the product of centuries-long discrimination whose recent history includes segregation, police brutality, and severe underfunding of social resources. With this in mind, even accurate and calibrated predictions of recidivism extend the legacy of historical discrimination by punishing blacks for having been subjected to such criminogenic circumstances in the first place. In other words, narrowly tailored considerations of fairness that operate within a broader unfair context can perpetuate the harm—one group of people being imprisoned disproportionately due to their race—that the introduction of machine learning into sentencing was intended to ameliorate.

Machine learning's inability to incorporate social and historical context into broader perspectives of fairness has the potential to hinder social change in two ways. First, in locating the problem of societal bias and discrimination at the site of singular decision points, machine learning turns our attention toward improving individual actors' judgments at the expense of interrogating systemic discrimination: without paying proper attention to broader contexts of injustice, we run the risk of overlooking systemic issues and deeming social structures fair simply because we have improved one component of them.

The second danger is that machine learning algorithms will act to entrench historical discrimination in society's decision-making. Machine learning operates by detecting historical correlations between features and outcomes, and applying those correlations to new data under the assumption that those same correlations will apply. But this methodology, even if it accounts for biases that result from individual instances of prejudice, is not suited to recognize changing social circumstances. Instead, it is conditioned on existing social circumstances under the assumption that the correlations indicative of certain outcomes in the training data will continue to apply in the future. In the case of COMPAS, for example, conditioning on the past prevents the algorithm from adapting to new circumstances that may arise due to social changes. That is, even if society enacted reforms that reduce recidivism among communities of color, the algorithm would be blind to these changes and would issue inaccurately high recidivism risks (likely leading to longer and more punitive sentences) to black defendants. Moreover, because of the criminogenic effects of incarceration (Cullen et al., 2011; Vieraitis et al., 2007), such predictions could in fact impede efforts to reduce recidivism—thus perpetuating the cycle of recidivism and incarceration that is rooted in racial injustice.

4. Conclusion

Fair machine learning as a field suffers from a significant lack of clarity about the types of problems that reside within its purview. The field's totalizing language of "fairness" stands in tension with its few attempts at critical engagement with the social and political contexts within which its tools are deployed—in fact, the adoption of such language to encompass what is actually a broad set of ethical, social, and political concerns is itself indicative of this lack of engagement. Underlying this approach is the belief that it is possible (and desirable) to optimize existing systems according to abstract universal notions of fairness without participating in political and social deliberation. Yet the field's desire for objectivity and neutrality is misguided (Porter, 1996). As philosopher Roberto Unger writes, neutrality is an "illusory and ultimately idolatrous goal" because "no set of practices and institutions can be neutral among conceptions of the good" (1987). In other words, even the most seemingly self-evident aspects of fairness can reflect a particular worldview that ought to be examined rather than taken for granted. By attempting to deal in universal frameworks for fairness, the field obscures the normative judgments that underlie its practice, allowing such judgments to pass without proper critical assessment. For the field to responsibly deal with issues of fairness, it must surface and interrogate its background assumptions and principles as well as engage more deeply with existing scholarship from other disciplines and the people whose lives will be affected by its algorithmic tools.

There are already some promising avenues of machine learning research that are pushing beyond the field's dominant methodology. Recent work has considered the feedback effects that may follow machine learning predictions, paying special attention to how even algorithms that satisfy standard fairness constraints can compound inequality (Hu & Chen, 2018; Liu et al., 2018). One recent manuscript develops a machine learning model that can adapt to "label shift," i.e., changes in outcome rates (Lipton et al., 2018). Other papers deploy the tools of machine learning to assess structural social conditions, producing analyses of topics such as gun violence (Green et al., 2017) and police behavior (Goel et al., 2016; Voigt et al., 2017).

Lastly, there is a deep need for the field to pursue expanded data collection efforts. Machine learning research tends to rely on existing datasets, and fairness is no exception: many papers are centered on the COMPAS dataset released by ProPublica or the Adult income dataset from the UCI ML repository. Relying on a small number of datasets constrains the field's ability to assess the full scope of fair machine learning questions and how they apply in different contexts.

Fair machine learning must recognize itself as participating in a normative construction of the world. The responsibilities of such a role require new methods that, rather than imposing a traditional machine learning paradigm to solve social challenges, are adapted to constructively engage in the ceaseless project of collectively building a fairer world.

References

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica, May*, 23, 2016.

Chouldechova, Alexandra. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Cullen, Francis T, Jonson, Cheryl Lero, and Nagin, Daniel S. Prisons do not reduce recidivism: The high cost of ignoring science. *The Prison Journal*, 91(3_suppl):48S–65S, 2011.

Dieterich, William, Mendoza, Christina, and Brennan, Tim. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

Dressel, Julia and Farid, Hany. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1): eaao5580, 2018.

Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Goel, Sharad, Rao, Justin M, Shroff, Ravi, et al. Precinct or prejudice? Understanding racial disparities in New York City's Stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394, 2016.

Green, Ben, Horel, Thibaut, and Papachristos, Andrew V. Modeling contagion through social networks to explain and predict gunshot violence in Chicago, 2006 to 2014. *JAMA internal medicine*, 177(3):326–333, 2017.

Hardt, Moritz, Price, Eric, Srebro, Nati, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hu, Lily and Chen, Yiling. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pp. 1389–1398, 2018.

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

Lipton, Zachary C, Wang, Yu-Xiang, and Smola, Alex. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*, 2018.

Liu, Lydia T, Dean, Sarah, Rolf, Esther, Simchowitz, Max, and Hardt, Moritz. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

Porter, Theodore M. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press, 1996.

Unger, Roberto Mangabeira. *False necessity: Anti-necessitarian social theory in the service of radical democracy*. CUP Archive, 1987.

Vieraitis, Lynne M, Kovandzic, Tomislav V, and Marvell, Thomas B. The criminogenic effects of imprisonment: Evidence from state panel data, 1974–2002. *Criminology & Public Policy*, 6(3):589–622, 2007.

Voigt, Rob, Camp, Nicholas P, Prabhakaran, Vinodkumar, Hamilton, William L, Hetey, Rebecca C, Griffiths, Camilla M, Jurgens, David, Jurafsky, Dan, and Eberhardt, Jennifer L. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, pp. 201702413, 2017.

Zafar, Muhammad Bilal, Valera, Isabel, Gomez Rodriguez, Manuel, and Gummadi, Krishna P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017.