"Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform

Ben Green [1]

Abstract

As risk assessments become increasingly recommended and adopted as a tool for criminal justice reform, the technical community and advocates alike must ask the right questions. Thus far, most analyses of risk assessments presume that narrow computational definitions of fairness are sufficient to ensure that the impacts of risk assessments are themselves fair and take for granted that risk assessments are an effective tool for advancing criminal justice reform. This paper interrogates both assumptions to highlight how even "fair" risk assessments can be unfair and hinder efforts to reform the criminal justice system. This analysis suggests several ways in which the field of fair machine learning must expand the considerations and questions that it deems relevant to evaluating and deploying risk assessments.

1. Introduction

Over the span of just two months in 2016 emerged competing stories: first, ProPublica exposed that the COMPAS recidivism risk score algorithm disproportionately falsely labels black criminal defendants as high risk of committing future crimes (Angwin et al., 2016); yet soon after, the Wisconsin Supreme Court defended the use of COMPAS to inform criminal sentencing decisions (Wisconsin Supreme Court, 2016). Ever since, the prevailing framework around debates over machine learning's role in the criminal justice system has been set: belief that better information could help judges make more accurate and unbiased decisions, on the one hand, versus concern that the algorithms are racially biased, on the other.

It is therefore commonly assumed that, with appropriate technical assurances of fairness, risk assessments that inform bail and sentencing decisions can be tailored into neutral tools to improve the criminal justice system. This sentiment was evident in a 2017 statement from several criminal defense organizations, in which they state, "racial bias [...] concerns should not be used to deter the use of pretrial risk assessment, but should instead be used to guide protocols" (Gideon's Promise et al., 2017). Notably, this endorsement and many others describe risk assessments not just as a means to improve predictive accuracy, but also as a way to achieve criminal justice reform (broadly speaking, eliminating or altering policies and practices that have historically led to mass incarceration and racial injustice). For example, Senators Kamala Harris and Rand Paul introduced the Pretrial Integrity and Safety Act of 2017, proposing to replace money bail with risk assessments as a way to increase pretrial release rates to 85% (Harris & Paul, 2017). Similarly, many endorsements of evidence-based sentencing are grounded in the goal of reducing incarceration (Starr, 2014).

Perhaps because the prospect of new technology being able to promote criminal justice reform is so alluring, the widespread support for predictive risk assessments (even "fair" ones) overlooks several important considerations. Analysis generally begins with the question: Would the criminal justice system be improved if judges made more accurate and unbiased predictions about defendants? About this there can be little debate. From there, however, many conclude that machine learning provides the path for criminal justice reform. Yet this belief conflates the desired end (criminal justice reform) with the proposed means to achieve it (risk assessments), leading to an unrealistically sanguine assessment of machine learning's decarceral potential. In particular, the analysis overlooks two essential, intermediate questions: 1) Do current computational notions of fairness account for the fairness issues borne by risk assessments? 2) Are risk assessments an effective strategy for advancing criminal justice reform?

In responding to both questions, this paper will cast significant doubt on the prospects of risk assessments to promote criminal justice reform. This discussion raises challenges that standard conceptions of fair algorithms do not address—nor, in fact, does the field tend to even recognize these issues as relevant considerations under the domain of fair machine learning—and highlights several possible dangers concomitant with attempts to address legal, social, and political issues related to fairness via computation.

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

Presented at the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018), Stockholm, Sweden.

2. Do current computational notions of fairness account for the fairness issues borne by risk assessments?

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

Machine learning is entirely reliant on data and metrics, and is therefore able to incorporate only facts and considerations that are measured quantitatively. Making decisions via machine learning can therefore distort the values inherent to the task at hand by granting undue weight to quantified considerations at the expense of unquantified ones. This concern is especially salient when applying machine learning to social decision-making processes: many aspects of society have been measured only in limited ways, other aspects of society resist quantification, and the data that exists reflects the biases and power dynamics that have led to certain aspects of society being measured at all (Gilliom, 2001; Scott, 1998).

This raises significant challenges when attempting to use machine learning to fairly adjudicate complex decisions. Determining sentences, for example, involves 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 COMPAS 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.

In the case of sentencing, if fairness requires the holistic balancing of several factors, increasing the weight placed on incapacitation will lead to sentences being determined on an unfair basis. Such a change can cause what are, in effect, significant shifts in policy and jurisprudence. But because these shifts emerge as indirect consequences of deploying an algorithm, they are likely to occur 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.

2.2. Machine learning's narrow focus entrenches historical discrimination

Any discussion of fairness that considers a decision in isolation of its broader social context is underspecified: what may appear to be fair under a narrow frame of predicting recidivism may be deeply unfair within a broader historical and cultural context. For example, the empirical finding that blacks recidivate at higher rates than whites (which leads to the conflict between calibrated predictions and error rate balance at issue in the COMPAS debate) is the product of historical discrimination. With this in mind, even perfectly accurate 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, narrow considerations of fairness that operate within a broader unfair context perpetuate the harm—one group of people being imprisoned disproportionately due to their race—that the introduction of machine learning into bail and 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 and entrench historical discrimination. For even if machine learning methodology accounts for biases that result from individual instances of prejudice, it is not equipped 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. For instance, following reforms such as text message reminders to appear in court for defendants released pretrial, risk assessments have produced "zombie predictions" that overestimate flight risk because they fail to account for the benefits of the program (Koepke & Robinson, 2018). Thus, even if society were to enact reforms that address current inequalities and reduce recidivism among communities of color, risk assessments may be blind to these new circumstances and continue operating under an assumed world in which blacks recidivate at their current levels. Not only would this compound COMPAS' existing issues by leading it to predict blacks as having inaccurately high recidivism risks (likely leading to longer and more punitive sentences), but also, because of the criminogenic impacts of incarceration (Cullen et al., 2011; DeFina & Hannon, 2010; 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.

2.3. Machine learning algorithms can never be neutral and free from normative influence

Part of the appeal of risk assessments is their supposed ability to make neutral, non-partisan predictions. Indeed, the case for algorithms in the criminal justice system largely relies on their being "objective" and "evidence-based" (Harris & Paul, 2017; Starr, 2014). Risk assessments are therefore typically presented to judges as part of a dossier of information about defendants (e.g., in pre-sentence investigations). No matter how much data and statistics are involved, however, an algorithm can never be truly neutral and free from normative values. Many subjective choices go into developing and implementing risk assessments: choosing what

data and features to use, designing algorithms to predict recidivism rather than outcomes such as rehabilitation, and defining thresholds for labels like "high risk" and actions like "detain." Although these decisions may appear benign, they embed political values and causal assumptions within risk assessments and frame sentencing around the prosecutorial and racialized notion of crime risk (Harcourt, 2015).

The false assumption of algorithmic neutrality is particularly dangerous because the criminal justice system operates on an adversarial process to adjudicate cases. Claims presented by the prosecution or defense are subject to the rigorous scrutiny of cross-examination, which has been described as "the greatest legal engine ever invented for the discovery of truth" (Wigmore, 1905). Presenting claims made by an algorithm as neutral facts removes risk assessments from the adversarial process; they are subject only to interrogation regarding their general scientific validity and the facticity of data about the defendant (Wisconsin Supreme Court, 2016). This leaves criminal defendants subject to claims made about their crime risk without sufficient means to confront or challenge these statements.

3. Are risk assessments an effective strategy for advancing criminal justice reform?

3.1. Machine learning narrows the scope of judgments about fairness

Recognizing that judges have cognitive limitations and personal prejudices, many have promoted risk assessments as a way to improve the accuracy and fairness of judicial decisions regarding bail and sentencing (aided by fair machine learning to ensure that these algorithms do not reproduce the biases that plagued these decisions in the past). This diagnosis of how to make the criminal justice system fairer, although well-intended, is limited by its narrow diagnosis of bias as the result of individuals acting in biased ways. Many forms of discrimination and oppression are produced not by people making biased judgments about other people, but through laws and institutions that systematically benefit one group over another; this is particularly true within the criminal justice system (Alexander, 2012). With their emphasis on improving individual decision-making within the criminal justice system, reform efforts based on risk assessments hazard overlooking structural issues and deeming the system fair because of improvements to limited components of it. Although these endeavors may advance certain goals of justice, their impact is constrained by their individualistic diagnosis of discrimination and reform, in a manner that mirrors the limited impacts of individualistic legal rights. As Mark Tushnet has argued, "progressive victories [of rights] are likely to be short-term only; in the longer run the individualism of rights-rhetoric will stabilize existing social relations rather than transform them" (Tushnet, 1993).

3.2. Technocratic reforms sanitize rather than alter the criminal justice system

The impacts of risk assessments are further limited by the focus on making the existing criminal justice system more fair rather than on substantively changing the system. By providing a veneer of neutrality and fairness, risk assessments may sanitize—and, hence, justify and perpetuate—the criminal justice system in its current state. Again, the limitations of risk assessments as a tool for reform mirror those of individual rights: building on the work of Tushnet, Paul Butler writes, "procedural rights may be especially prone to legitimate the status quo, because 'fair' process masks unjust substantive outcomes and makes those outcomes seem more legitimate" (Butler, 2012). Although technocratic reforms can have value, they must follow rather than precede systemic reforms—especially where systemic reforms are both necessary and possible. There is currently broad support across the political spectrum for criminal justice reform: for example, 71% of voters in New York State support ending pretrial jail for misdemeanors and non-violent felonies (FWD.us, 2018) and 87% nationwide support removing mandatory minimums for nonviolent offenses (Blizzard, 2018). Reformers would therefore be better served building momentum to abolish practices like pretrial detention altogether, rather than justifying its existence by making it "fair." Doing so sanitizes the current state of the criminal justice system, and may in turn distract from or reduce political will for alternative approaches to reducing recidivism, such as food stamps and prisoner education programs (Davis et al., 2013; Tuttle, 2018). In effect, reforms centered on risk assessments appear to concede that the criminal justice system is largely immutable and that the only appropriate response to someone with a predicted high risk of recidivism is locking them in jail or prison. Criminal justice reform efforts must dismantle, rather than accept, such notions.

3.3. The impacts of risk assessments are brittle and subject to political capture

Because technocratic reforms tie political ends (in this case, decreasing discrimination and incarceration) with technical means (risk assessments), achievement of the political goal rests on a particular use of the tool. Yet just because algorithms *could* help create a more informed and fair criminal justice system does not mean that they inevitably *will*; existing social structures and power dynamics (and the biases therein) will shape their social impacts. Although risk assessments may reduce biases and incarceration under circumstances that support these goals, they can also be manipulated to achieve the opposite ends. In New Jersey, for example, after some defendants accused of certain gun charges were released before trial and went on to commit further crimes, the State Attorney General's office pressured the courts to automatically detain every defendant arrested for

those same gun charges, regardless of his or her risk score (Schuppe, 2017). Similarly, criminal justice practitioners "strategically exercise their discretion when filling out risk assessments [...] to control the final score" based on their own clinical judgment (Hannah-Moffat, 2015). Susceptibility to such manipulation is inherent to risk assessments: setting parameters such as cutoffs that determine who receives a label of "high risk" or who is released before trial is ultimately a political exercise that involves making normative judgments about the tradeoffs between reducing incarceration and reducing crime. Even if a threshold is set at the outset to promote high levels of pretrial release, it can always be changed at a later date to reduce pretrial release. In addition to defeating the original purpose of the tool, such a change is likely to pass without sufficient political scrutiny or public discussion because it could be framed as a technical tweak rather than a significant policy change. Thus, for those who strive for criminal justice reform, putting faith in risk assessments is akin to putting many eggs in a flimsy basket. If risk assessments are not implemented as desired—or if the predictions simply indicate high levels of risk such that incarceration cannot be sufficiently reduced under existing parameters—reformers will face the unenviable task of recalibrating their position, either revealing their stance on risk assessments to be highly contingent on a particular implementation or abandoning risk assessments altogether.

4. Conclusion

It is clear, as others have similarly argued, that the field of fair machine learning must expand the considerations and questions that it deems relevant to assessing and applying risk assessments (Barabas et al., 2018). Algorithms used in contexts like bail and sentencing have emerged out of the computer terminal—where they can be fully defined in terms of their technical specifications—and into a sociotechnical environment in which they interact with judges, power dynamics, and laws. It is in this context that criminal justice algorithms must be assessed for bias, discrimination, and other social impacts. Computer scientists developing or otherwise studying risk assessments must abandon naïve notions of neutrality and recognize themselves as participating in normative and political constructions of society. While interdisciplinary collaborations are essential, equally important is coming to terms with the political nature of the field's work. Computer scientists may disagree about what goals and values to support (and the field need not dogmatically enforce a single position), but it is necessary to surface such debates and view them as an integral component of the fair machine learning research process.

To the extent that the field continues to pursue risk assessments as a tool for criminal justice reform (although this paper describes several reasons to be skeptical of such a cause), it must take steps to do so more responsibly and thoughtfully. Efforts to collect more detailed and unbiased data about society are essential. If only some of the considerations behind a decision are well-measured, then machine learning algorithms will be unable to capture the full set of principles that are meant to underlie that decision. And if risk assessments are trained on data from a location or time period with different conditions than the ones in which they are deployed, they may embed the discrimination from one context in the decisions of another. It is also necessary to develop machine learning models that can address a wider range of social questions and situations. Already there is some promising work along these lines, such as efforts to develop algorithms that adapt to changing conditions (Lipton et al., 2018) and to deploy machine learning to assess the structural conditions of crime and the criminal justice system (Crespo, 2015; Green et al., 2017).

Computer scientists who support criminal justice reform ought to proceed thoughtfully, ensuring that their efforts are driven by clear alignment with the goals of justice rather than a zeitgeist of technological solutionism. Although it is possible for risk assessments to reduce incarceration and bias, they come with no guarantee of doing so and have the potential to hinder more systemic reforms in the long run. It is not enough to have good intentions—computer scientists must critically assess the likely impacts and downstream consequences of risk assessments based on the values and incentives of the system in which they are embedded. At the very least, reformers must push for additional safeguards to be implemented alongside the adoption of risk assessments. Given that machine learning algorithms are political in ways that have received insufficient attention, it is imperative to increase the transparency and democratic governance of risk assessments. With this in mind, reformers should compel courts to treat predictive algorithms like COMPAS as a form of expert testimony on behalf of the state (who typically purchases and manages the software). Extending cross-examination to risk assessments would protect defendants by ensuring that appropriate scrutiny is paid both to the methods used and the subjectivity of framing decisions around a particular prediction (Roth, 2016).

Many of the issues and questions raised in this paper fall outside the strict bounds of fair machine learning, yet are deeply tied to questions of fairness in machine learning, broadly conceived. As the field increasingly looks toward social domains like the criminal justice system, it can no longer take for granted that fairness in its myriad and conflicting meanings can be reduced to a single mathematical definition that exists in the abstract, away from social, political, and historical context. Interdisciplinary and critical analyses such as those described here are essential if the field is to develop a synthesis of computation with law and policy that enhances not only fairness, but justice.

Acknowledgements

Thank you to Lily Hu and two anonymous reviewers for constructive feedback on earlier drafts of this paper.

References

Alexander, Michelle. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2012.

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. Machine bias. *ProPublica*, 2016.

Barabas, Chelsea, Dinakar, Karthik, Virza, Joichi Ito, and Zittrain, Jonathan. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.

Blizzard, Robert. Key findings from a national survey of 800 registered voters January 11-14, 2018. Report, Justice Action Network, 2018. URL http://www.justiceactionnetwork.org/wp-content/uploads/2018/01/JAN-Poll-PPT-Jan25.2018.pdf.

Butler, Paul D. Poor people lose: Gideon and the critique of rights. *Yale Law Journal*, 122:2176–2204, 2012.

Crespo, Andrew Manuel. Systemic facts: Toward institutional awareness in criminal courts. *Harvard Law Review*, 129:2049–2117, 2015.

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.

Davis, Lois M., Bozick, Robert, Steele, Jennifer L., Saunders, Jessica, and Miles, Jeremy N.V. *Evaluating the Effectiveness of Correctional Education: A Meta-Analysis of Programs That Provide Education to Incarcerated Adults*. RAND Corporation, 2013.

DeFina, Robert and Hannon, Lance. For incapacitation, there is no time like the present: The lagged effects of prisoner reentry on property and violent crime rates. *Social Science Research*, 39(6):1004–1014, 2010.

FWD.us. Broad, bipartisan support for bold pre-trial reforms in New York State. 2018. URL https://www.fwd.us/wp-content/uploads/2018/03/NYCJR-poll-memo-Final.pdf.

Gideon's Promise, The National Legal Aid and Defenders Association, The National Association for Public Defense, and The National Association of Criminal Defense Lawyers. Joint statement in support of the use of pretrial risk assessment instruments. 2017. URL http://www.publicdefenders.us/files/Defenders%20Statement%20on%20Pretrial%20RAI%20May%202017.pdf.

Gilliom, John. *Overseers of the Poor: Surveillance, Resistance, and the Limits of Privacy*. University of Chicago Press, 2001.

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.

Hannah-Moffat, Kelly. The uncertainties of risk assessment: Partiality, transparency, and just decisions. *Federal Sentencing Reporter*, 27(4):244–247, 2015.

Harcourt, Bernard E. Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, 27(4): 237–243, 2015.

Harris, Kamala and Paul, Rand. Pretrial integrity and safety act of 2017. *115th Congress*, 2017.

Koepke, John Logan and Robinson, David G. Danger ahead: Risk assessment and the future of bail reform. *Washington Law Review, Forthcoming*, 2018.

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.

Roth, Andrea. Machine testimony. *Yale Law Journal*, 126: 1972–2053, 2016.

Schuppe, Jon. Post bail. *NBC News*, 2017.

Scott, James C. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, 1998.

Starr, Sonja B. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, 66(4):803–872, 2014.

Tushnet, Mark. The critique of rights. *SMU Law Review*, 47:23–34, 1993.

Tuttle, Cody. Snapping back: Food stamp bans and criminal recidivism. 2018.

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.

Wigmore, John Henry. *A Treatise on the System of Evidence in Trials at Common Law*. 1905.

Wisconsin Supreme Court. *State v. Loomis*. 881 Wis. N.W.2d 749, 2016.