Abandoning Criminal Risk and Recidivism: On Dangerous Goals in ML Scoring-Decision Systems

Benjamin D. Laufer
San Francisco, CA
ben.laufer@gmail.com

Abstract

Machine Learning research has demonstrated that algorithmic risk assessments can assist (or out-perform) judicial decisions in meeting some number of quantifiable goals. Typically, these goals are phrased in terms of 1) minimizing risk to public safety, 2) clinically treating a defendant to reduce their probability of recidivating or 3) guaranteeing some measure of group fairness across racial (or other group) belonging. This paper situates calls for clinical interventions in a history of incapacitation theory in criminology. It argues that evidence-based metrics used to demonstrate successful clinical intervention or group fairness treat criminal risk as exogenous; out of the control of judicial decision-makers. In reality, risk and recidivism are deeply connected to a history of systemic criminal legal violence, and continue to perpetuate that violence. An abolitionist approach to criminal risk does not celebrate lowering an individual's risk of recidivism, or jailing Black and white Americans at a level proportional to their respective crime rates. Instead, it requires acknowledging that the current measures of criminal risk are fraught, full stop. As an alternative, we suggest focusing on reducing *harm* rather than *criminal risk*, which will always excuse the organized, systemic violence carried out legally by the state.

1 Introduction

1.1 The Impulse to Throw Data/ML at Hard Problems

The advent and preliminary success of machine learning models has led to significant speculation and excitement about applications. These algorithms are well-suited to problems that require prediction of an unknown. In medicine, predictions may enable treatments that are more equipped to help individual patients. In finance, predicting default can enable more profitable lending for banks. In education, predicting student success can inform decisions in curricula, teaching, admissions and more. The extensive set of unsolved problems in society naturally drives our excitement about innovation. Accordingly, our impulse is to consider some of the most serious and impactful applications first.

Theorists, policy makers and practitioners have adopted the view that predictive algorithms can aid vital public institutions that persistently under-perform. Many mention criminal legal decisions like bail as an example of a flawed decision-system that could benefit from ML predictions [17, 16]. Even though there have been countless academics and community members who advocate against the use of algorithmic decisions in criminal legal processes, developers of risk assessment tools were quick to adopt much of the technical language from ML and Data Science research. As such, the 'actuarial impulse' in criminal punishment [15] has received a new surge of optimistic support and enthusiasm.

1.2 Data is used to develop and legitimize criminal legal decision processes

Looking at the way practitioners and developers refer to criminal risk assessments, we can get a sense of how they are justified ideologically. In a practitioner's guide for COMPAS, a widely used algorithm developed by Northpointe, Inc., the company tries to describe the wide embrace of statistics and data:

Resistance AI Workshop. 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

> Statistically based risk/needs assessments have become accepted as established and valid methods for organizing much of the critical information relevant for managing offenders in correctional settings. Many researchers have concluded that objective statistical assessments are, in fact, superior to human judgment. COMPAS is a statistically based risk assessment developed to assess many of the key risk and needs factors in adult correctional populations and to provide information to guide placement decisions. [23, p1-2]

Sweeping statements about algorithms outperforming human predictions are cited as evidence for the adoption of risk assessment tools. Similarly, emphasis on big data and academic research is used to establish the reputation of risk assessment tools. The Public Safety Assessment (PSA) boasts its data-driven development: "Researchers designed the PSA based on the largest, most diverse set of pretrial records ever assembled—750,000 cases from nearly 300 jurisdictions. Based on a comprehensive analysis of the data, researchers identified the nine factors that best predict pretrial risk" [25].

Evidently, people are excited about using algorithms in high-impact fields. But the question remains: are these algorithms being developed and adopted ethically? Innovations in highly impactful domains may feel promising, but can cut dangerous corners, or have harmful unintended consequences. Where some emphasize predictive accuracy or algorithmic simplicity over causal understanding, we see a need to reconcile the complexity of the problem with the narrow scope of a predictive algorithm. To deal with a highly complex and impactful problem like policing, bail or sentencing, researchers are finding themselves needing to simplify and narrow their focus. People are fitting the problem to the model, instead of fitting a model to the problem.

1.3 Contribution

We first review a recent push for applying ML predictions to bail, and a corresponding push-back advocating clinical interventions in ML Fairness and Accountability literature. We contextualize this debate within an older development in criminology from the principle of 'risk' to the more holistic 'risk, needs and responsivity' goals that underpin modern criminal treatment. In making this connection we argue that clinical approaches to judicial decisions are compatible with, and grow out of, the theory of incapacitation and risk principle that justify long prison sentences and tough-on-crime policy.

Second, we use a very simple toy statistical score-decision system to show that risk is not external to criminal legal decision-making, and risk-based scoring systems can pass typical 'fairness' metrics without acting in a way that remotely resembles fair institutional treatment.

Finally, we discuss how current methods developed under the auspices of 'Fairness and Accountability in Machine Learning' can legitimize and perpetuate an understanding of risk as innate to the individual and exogenous to the state. We conclude that the impulse to set measurable goals in CSS can dangerously simplify complex socio-technical systems and in doing so can harm communities most vulnerable to organized violence and organized abandonment [12].

2 A Literature that Cannot Decide its Goal

Recent literature on criminal legal decision-making focuses either on enhanced prediction capabilities or on more sensitive and individually-catered clinical treatment.

2.1 Risk as a Prediction of the Future

Some leading researchers have encouraged predictive accuracy above meaningful causal inference in these 'high-impact' fields [17]. Others suggest simple linear regressions and rounding techniques for complex decisions, even when such models do not carefully treat underlying relationships between variables [16]. As a result, numerous papers demonstrate that algorithmic approaches to bail decision can 'out-perform' unassisted judges on simple metrics like recidivism rates [19].

Risk assessment developers and researchers who 'validate' these assessments similarly have used statistical tests to establish a tool's predictive accuracy and fair treatment of defendants. The Risk Principle, a term used widely in criminology literature and research on risk assessment algorithms, is the idea that defendants should be treated in a way that is commensurate with their level of risk to public safety [21, 1]. While some have called into question whether criminal legal treatment should ever be based off of a prediction of a future crime [15], it is notable how foundational the risk principle is to decisions made in policing, pre-trial detention and bail, sentencing, and parole.

2.2 Risk as a Condition to be Treated

Departing from the predictive view of algorithmic risk assessment, a few authors have begun to emphasize alternative goals that may sacrifice predictive accuracy. These include decision explainability and transparency [4], group fairness [20, 18], and re-imagining criminal decisions as clinical [5].

The clinical view is encapsulated well in Barabas et al. [5], which challenged the conventional treatment of risk assessment as a prdictive policy problem. The paper instead draws attention to the treatment effect of criminal intervention: "If machine learning is operationalized merely in the service of predicting individual future crime, then it becomes difficult to break cycles of criminalization that are driven by the iatrogenic effects of the criminal justice system itself" [5, p1]. Their central claim is that treating risk-assessment as a prediction policy problem does not actually answer the question of how we may be able to *lower* risk in the future: "Predictive risk assessments offer little guidance on how to effectively intervene to lower risk" [5, p10].

This clinical approach to criminal legal decision-making pushes back convincingly on the idea that risk is an immutable characteristic that can be objectively predicted. However, it still clings to the idea that criminal risk is a legitimate characteristic that can be measured and demonstrably reduced.

Clinical Approaches to Criminality are Not New

While identifying an important flaw in solely using predictive algorithms in criminal policy, clinical treatment of criminal policy is not a new idea.

After statistical tools to predict criminal risk were introduced as early as 1928 [6], tough-on-crime policies were often justified because criminals were seen as a risk to public safety [14]. The underlying 'risk principle' was a theory in criminology that legal responses should be commensurate with the level of risk that a defendant holds - an anticipation of future crime is a legitimate reason to cater a length of sentence or bail decision.

A reform-minded push in criminology in the 1970s and 1980s, formalized first in 1990, advocated for a 'risk-needs-responsivity' framework that focused not only on static risk factors but on dynamic statuses (employment, housing, etc) and on the specific responsiveness that a legal punishment might have on the defendant [2]. Interestingly, the paper frames risk, needs and responsivity as guiding principles for psychological *classifications* related to rehabilitation. From its inception, the idea of criminal legal decisions as a matter of clinical treatment relied on having some meaningful measures that can be shown to improve after intervention.

The influence of risk-needs-responsivity is evident in criminology. Risk assessment developers and researchers have adopted language like "treatment" [1, 13, 21] and "dosage" [22]. These medical terms' use imply that there exists a fundamental, measurable level of risk that can be reduced by interventions.

Therefore, criminal punishment-as-treatment is still compatible with (and premised on) the risk principle. Theories that emphasize clinical treatment over predictions still rely on the designation of *criminality* to determine interventions. Barabas et al. [5] advocate what they see as an alternative to predictive algorithmic risk assessment: "Rather than using machine learning for prediction, these methods could be used to identify features that are highly predictive of recidivism, in order to inform hypotheses on interventions (and their timing) that can then be tested using causal inference". Indeed, treatment methods share assumptions with theories of selective incapacitation that trace back to the 1980s - that criminal policy should anticipate future crimes, and intervene in a way that protects society from potentially dangerous people [14]. These methods assume that outcome variables - namely, recidivism - are objectively and equitably distributed. They assume that the designation of criminal action itself is in some way fundamental, and that police officers, juries, judges, prison guards and parole officers do not influence the *labelling* of people as criminals in problematic ways.

2.3 Risk as a benchmark for establishing fairness

Research in fairness and algorithmic accountability has directed significant attention to the the possibility of machine bias. Perhaps the most famous instance is the ProPublica-Northpointe debate [3, 10]. A ProPublica article in 2016 alleged that COMPAS is biased towards Black defendants, and the algorithm's developers at Northpointe published a response that defended the algorithm. Further studies found that the two conflicting findings used different and largely incompatible definitions of fairness [18, 8, 7]. Dieterich et al. [10] was considering whether the algorithm had different accuracy levels across race groups, and Angwin et al. [3] was considering whether the false positive and false negative misclassification rates were different across race groups.

Both studies, however, used some kind of 2-year rearrest rate as their ‘source of truth’ for recidivism. Numerous analyses have pointed out that this outcome variable itself can be unfair, so predicting the outcome in an unbiased way can still lead to unfair or unjust treatment [11, 18]. Generally, this paper aims to demonstrate that the underlying concept of criminal risk has problems. These problems concern the very foundation of using risk assessment tools, and are not aided by incremental improvements to some measure of bias or accuracy.

3 Faux Exogeneity in Socio-Technical Scoring

When the designation of criminality is historically or currently biased, a ‘fair’ determination of risk will still exhibit those biases. In this section we aim to show formally that when the criminal label is racist, so too are ‘perfectly unbiased, perfectly predictive’ assessments of future criminal risk.

Suppose we have some outcome variable Y that is a measure of recidivism. We have a criminal legal system has a scoring system that takes information X about a defendant and predicts $\hat{Y} = P(Y|X)$, the risk of recidivism. The population can be split based on racial identity group g, where $g \notin X$.

Calibration is a measure of fairness that aims to guarantee that race does not mediate the score’s accuracy for defendants at a certain risk level. So, if a judge looks at a score, the defendant’s race will not tell the judge any additional information: $P(Y|\hat{Y}) = P(Y|\hat{Y}, g)$.

Equal Misclassification Rates aims to guarantee that defendants are misclassified as high-risk or low-risk at equal rates across race group. $FPR_{\hat{Y},g} = FPR_{\hat{Y}}$ and $FNR_{\hat{Y},g} = FNR_{\hat{Y}}$.

Findings by Kleinberg et al. [18] and others find that the metrics above are only satisfied in certain unrealistic cases, including a ‘perfect prediction’ algorithm. When the score is perfectly indicative of the underlying value (in this case, recidivism), the accuracy is 100% independent of race or group belonging (which establishes calibration) and misclassification rates are always 0% and thus equal across groups.

Now suppose that Y is somehow a function of race group g, or $Y = f(g)$. It is not hard to understand how the label Y may be racially mediated. If police officers are more likely to search a Black man, then Black men will be more likely to be convicted of crimes generally, and therefore recidivism Y would depend on group designation.

Even when the system allocates criminal labels in a racist way, a perfect prediction algorithm (even if it doesn’t model race explicitly) would pass the above tests for algorithmic fairness. The point here is that formal algorithmic fairness can be achieved for any scoring system, *no matter what quantity is being predicted or how the score is being used.*

4 Discussion

This paper begins to look holistically and how risk and recidivism are invoked in criminal legal research. More broadly, the inquiry can help expose the different ways that fairness and accountability studies in ML may actually help to consolidate power through algorithmic decision-making. Routine audits and fairness guarantees in scoring-decision systems are important, but can ignore the fact that the same agent might be behind the scoring, labelling, and decisions.

Criminality is not something innate to an individual; it is not something that can be deduced from somebody’s history or upbringing or genetics. It is instead a concept that is fully fabricated by the dominant social forces, and can, quite possibly, be utterly dismantled [9].

Instead of guaranteeing some particular fairness metric to justify jailing Black and white Americans at rates proportional to their crime rates, or to reduce recidivism by caging defendants for certain periods of time, we should think more deeply and critically about our goals (see [24]). Criminalized behavior is not the only type of violence that faces our society - the fully legal system of mass incarceration, policing, and war are all violent and all cause harm. If we take aim at these highly organized, systemic, legally protected forms of violence, we may more effectively serve people. Machines are very good at learning when there is a quantifiable, well-defined goal. It is only through getting these goals right that machines can help people.

Broader Impact

My hope is that this inquiry exposes some of the ways that ML research can miss the point when it comes to people's lived experiences. The discussion and analysis were specifically about the criminal legal space; however, many of the findings are relevant to the use of high-impact ML algorithms in many fields. In credit and medicine, for instance, risk determinations are premised on historical access to resources (e.g. capital or medical attention), so when future triage decisions are made, risk-based decisions will always exhibit the effects of historical decisions. None of these systems should treat risk as exogenous or innate and should instead have the goal of *minimizing harm*.

More broadly, I want this line of inquiry to contribute to a reckoning about 'Fairness and Accountability' studies in machine learning. In the few years since this field came into popularity, we already see an invasion of corporate interests pushing these concepts and studies in a decidedly not radical direction. Auditing ML scoring algorithms to make sure that score allocation is commensurate with 'base rates' of different groups *is not* making AI/ML work for people. It does more to exonerate practitioners using those algorithms to make decisions the way they always have. It is out of a deep belief that these fields are *not* going in the right direction that I submit this paper. It is also the reason that I am greatly excited about 'Resistance AI' as a field and community with some proximity and sway to NeurIPS.

Finally, because I truly have no idea what broader impact this or any research may have on people, I want to use this venue to directly promote some mutual aid and organizational efforts that can directly and more surely redirect resources into the hands of people who need it.

- For the Gworls https://www.artsbusinesscollaborative.org/asp-products/for-the-gworls-rent-and-gender-affirming-surgery-fund/
- Survived and punished NY/Abolitionist Mutual Aid Fund https://www.paypal.com/pools/c/8npG1wcuiJ
- Artist Relief Tree fund https://artistrelieftree.com/#donate
- Organizers Sex Worker's Relief Fund https://gofundme.com/f/z6w8v5
- Violence Intervention Program (VIP Mujeres) https://violence-intervention-program.networkforgood.com/projects/12549-we-need-your-help-to-broaden-our-impact
- Bay Area Workers Support (BAWS) https://bayareaworkerssupport.org/take-action

Acknowledgments

I'd like to acknowledge Miklos Racz, my undergraduate research advisor who has been helping me pursue and build on my research after graduating.

I'd like to acknowledge my friends, family, colleagues, and role models who have provided me with all the skills and access necessary to submit to a venue like this one.

I'd like to acknowledge all fellow LGBTQ folks in STEM and more broadly all the people, especially BIPOC, who are under-represented and over-impacted by AI, ML and other 'advances' in science and technology.

References

[1] D. A. Andrews and C. Dowden. Risk principle of case classification in correctional treatment: A meta-analytic investigation. *International journal of offender therapy and comparative criminology*, 50(1):88–100, 2006.

[2] D. A. Andrews, J. Bonta, and R. D. Hoge. Classification for effective rehabilitation: Rediscovering psychology. *Criminal justice and Behavior*, 17(1):19–52, 1990.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 23, 2016.

[4] S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[5] C. Barabas, K. Dinakar, J. Ito, M. Virza, and J. Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *arXiv preprint arXiv:1712.08238*, 2017.

[6] E. W. Burgess. Is prediction feasible in social work-an inquiry based upon a sociological study of parole records. *Soc. F.*, 7:533, 1928.

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

[8] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[9] A. Y. Davis. *Are prisons obsolete?* Seven Stories Press, 2011.

[10] W. Dieterich, C. Mendoza, and T. Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

[11] A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[12] R. W. Gilmore. Organized abandonment and organized violence: Devolution and the police. URL `https://vimeo.com/146450686`.

[13] A. Gordon and T. Nicholaichuk. Applying the risk principle to sex offender treatment. In *Forum on Corrections Research*, volume 8, pages 36–38, 1996.

[14] P. W. Greenwood, A. F. Abrahamse, et al. *Selective incapacitation*. Rand Santa Monica, CA, 1982.

[15] B. E. Harcourt. *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press, 2008.

[16] J. Jung, C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein. Simple rules for complex decisions. 2017.

[17] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

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

[19] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2017.

[20] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the age of algorithms. Technical report, National Bureau of Economic Research, 2019.

[21] C. T. Lowenkamp, E. J. Latessa, and A. M. Holsinger. The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency*, 52(1):77–93, 2006.

[22] M. Makarios, K. G. Sperber, and E. J. Latessa. Treatment dosage and the risk principle: A refinement and extension. *Journal of Offender Rehabilitation*, 53(5):334–350, 2014.

[23] Northpointe. Practitioner's guide to compas core, 2015.

[24] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

[25] Laura and John Arnold Foundation. The psa: Factors and formulas, 2019. URL `https://www.psapretrial.org/about/factors`.