Harvard Data Science Review • Issue 2.1, Winter 2020

Transparency and Simplicity in Criminal Risk Assessment

Alexandra Chouldechova¹

¹Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America

Published on: Mar 31, 2020

DOI: <https://doi.org/10.1162/99608f92.b9343eec>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

In “[The Age of Secrecy and Unfairness in Recidivism Prediction](#),” Rudin, Wang, and Coker (2020, hereafter, RWC) contend that the current focus on questions of algorithmic fairness is misplaced. Rather, they argue, we must insist first and foremost that the algorithms used in high-stakes settings such as criminal justice are *transparent*, prioritizing transparency over other forms of fairness. The authors make their case by taking us on a deep dive into what can be learned about a proprietary risk assessment tool, COMPAS. Through their methodical statistical detective work and reflections on stories such as that of Mr. Rodríguez, who was denied parole on the basis of an erroneously calculated COMPAS score, they make clear that without transparency “algorithms may not do what we think they do, [...] they may not do what we want,” and we may be left limited in our recourse against it.

I agree that algorithmic transparency is in many ways of fundamental importance. However, to ensure that algorithms ‘do’ what we want them to, we need more than the transparency afforded by having a simple model whose formula is publicly accessible. A key issue here is that algorithms do not do very much. In settings such as criminal justice, secret (or not-so-secret) algorithms commonly influence and inform decisions about individuals, but they do not make them. Recidivism risk, such as may be accurately estimated through the simple models lauded by the authors, is only one part of the decision-making equation. As I will discuss further in this response, a big part of the challenge in assessing an algorithm’s fit-for-purpose is that there is no decision-making equation to speak of in the first place. My use of the term ‘equation’ is thus aspirational at best.

Let us start with what may feel like a deceptively simple question: *Why might we not want to take all the human decision makers currently involved in sentencing and parole decisions and replace them with a simple recidivism prediction model?* As RWC note, human decision-makers are yet another way of providing “nontransparent, potentially biased decisions.” So why not replace them altogether? I ask this question not because I believe this is what RWC are proposing. Rather, these concrete questions—such as the authors’, “is age fair?”—force us to think more critically about what we are seeking to achieve, and how we can tell whether we have achieved it.

In reasoning about this question, I want to begin by distinguishing between three related tasks: sentencing and parole decision-making, risk assessment, and risk prediction. We can think of risk prediction as the task of estimating the likelihood that an offender will commit a crime upon release. Risk assessment, on the other hand, is the task of evaluating whether an individual has various criminogenic risk and/or needs factors, often with the partial aim of condensing some of this information into a single composite risk score. For instance, it may be important to learn about a person’s substance abuse history for the purpose of evaluating drug and alcohol treatment eligibility irrespective of whether substance abuse items feature in the risk prediction model. Risk prediction and risk assessment are just two components of sentencing and parole decisions, whose ultimate payoffs depend strongly on a multitude of other considerations. Replacing human decision makers

with a recidivism risk prediction algorithm would ignore all these other considerations except insofar as they are captured by risk as predicted by the model. This may be somewhat more reasonable in the current pretrial decision-making context, where judges and magistrates are tasked with assessing whether individuals present flight or public safety risks. But sentencing does not map cleanly onto a risk prediction exercise.

To flesh this out further, consider the question raised by the authors in the discussion about the role of current charge information in risk prediction. RWC write:

How much should one weigh the current charges with the COMPAS scores? This is not clear. Because COMPAS is a black box, it is difficult for practitioners to combine the current charge (or any other outside information) with the COMPAS scores. Because current charges are separate, COMPAS scores are not single numbers that represents risk. Instead their interpretation has a large degree of freedom.

I would like to offer an alternative perspective on this issue. First, I agree that it is generally unclear how COMPAS scores should be weighed with current charges. But this is not because COMPAS is opaque, nor because it does not use current charge information in its risk calculation. The COMPAS scores are the result of a risk assessment and risk prediction exercise, and can thus be viewed as single numbers that represent risk. But in some sense that is *all* they can be viewed as. What they are not is single numbers that represent everything relevant to *sentencing*. Making risk the focal point of sentencing decisions would represent a fundamental shift away from the principles of limiting retributivism that underlie much of existing sentencing practice.

The root issue I would point to here is thus not COMPAS, its complexity, or opacity, but rather the murky role that risk plays in sentencing. It is not that complexity and opacity are not problematic—I agree with the authors that they are. It is just that the same fundamental problem would persist even if we had a simple risk prediction model. While sentencing guidelines are generally very clear on how charge information should be taken into account, how risk should be used is hardly ever made explicit. Under existing sentencing practices, current charge information is decision-relevant *whether or not* it is predictive of risk. The proposed final draft of the new Model Penal Code, for instance, stipulates that sentences should fall “within a range of severity proportionate to the gravity of offenses, the harms done to crime victims, and the blameworthiness of offenders,” and within that range they should be set “to achieve offender rehabilitation, general deterrence, [and] incapacitation of dangerous offenders” (American Law Institute, 2017). Current charge information is relevant from a retributive standpoint in setting an appropriate sentence severity range. Within that range, there is space for risk and needs assessment to help in determining a specific sentence, and in identifying higher-risk offenders whose risk may be mitigated through more intensive correctional services.

Whereas the current charge may be relevant to the decision but not to risk prediction, other factors may be relevant only to assessing risk, or both to assessing risk and to other decision criteria. An offender’s sex may be relevant in predicting risk—as determined, for instance, in *Wisconsin v. Loomis*—but would not in general be a

permissible factor in determining a sentence beyond its role in risk. Factors such as prior criminal history, on the other hand, are not only predictive of risk, but, for better or worse, may also influence perceptions of blameworthiness. Such factors may thus enter into sentencing decisions twice: once in determining an offense-appropriate range, and separately in assessing risk for setting a sentence within that range. We may be concerned that this type of double-counting may result in sentences that are too heavily tied to criminal history. But if we are, then we may need to rethink our conceptualization of what we want to get out of a risk prediction tool. That is what I'd like to turn to next.

Just as sentencing decisions depend on many factors other than risk, a risk assessment tool's fit-for-purpose depends on more than its simplicity and the accuracy of its risk predictions. In "Statistical Modeling: The Two Cultures," Leo Breiman introduces the term *Rashomon Effect* to refer to the fact that for any given task "there is often a multitude of different [algorithms] giving about the same minimum error rate" (Breiman, 2001). While this multitude of models has comparable accuracy, they may differ significantly in terms of their features and structure, and also properties such as simplicity, fairness, interpretability, trustedness, and influence on decision-making. Some of these properties appear to be aligned with one another, but in many cases they are not. For instance, we know from work such as Kleinberg and Mullainathan (2019) that, at least theoretically, simplicity can be at odds with equity, and simple models can incentivize decision makers to statistically discriminate based on group membership. To the extent that these theoretical conclusions are borne out in practice in a given setting, this presents clear trade-offs between simplicity, fairness, and behavioral incentives. While accuracy often retains pride of place in modeling considerations, simplicity may be subordinate to other desirable properties when they are found to be at tension.

Breiman's observation tells us that even after we've near-optimized for accuracy, we have a large landscape of possible algorithms between which to choose. If these algorithms differed only in terms of their structure, it would make sense to decide between them on the basis of structural properties such as model simplicity and the features used. That is, when two models that differ in their complexity lead to the same predictions, we may have a strong preference for the simpler of the two. However, as we discuss in Chouldechova and G'Sell (2017), for problems that have a high irreducible error, such as recidivism prediction, comparably accurate models can—and do—disagree on a large fraction of their classifications. This can be the case even when both classify the same fraction of cases as positive (or 'high risk'). For instance, a two-variable regression model involving age and the priors count achieves an AUC similar to that of COMPAS; yet when one looks at whether an individual is estimated to be in the top 25% of risk by the models, one finds they disagree on 1 in 8 of their classifications. A model that uses priors count alone achieves an AUC that is just a few points lower than COMPAS, but that disagrees with COMPAS on 1 in 4 cases. Of course, these are just two models. One can consider many others. The point is that, when classifications influence life-changing decisions about individuals, these differences in individual-level classification may begin to matter. We may find that for any complex model with desirable properties, there is a simple model that enjoys all the same properties. But when we do not, it is reasonable to question whether simplicity should take precedence over competing desiderata.

Lastly, I want to say a few words about transparency beyond simplicity, and why the focus on fairness is, at least in my view, not all that misplaced. Meaningful algorithmic transparency can entail many things, including but not limited to: access to the training algorithm, the trained algorithm, the training data and its provenance, a justification of the prediction target, a clear articulation of the purpose for which the algorithm is designed and how it is intended to be used, and a reproducible validation study of the algorithm for the proposed use case. Each of these forms of transparency enables us to deliberate on the algorithm and its use in a given context. Having a simple model makes it easier to communicate the details of the trained algorithm, but it does not buy us much if those models predict the wrong outcome, are trained on nonrepresentative data, or are applied out of context.

As the authors demonstrate in their investigation of COMPAS, even without transparency we may be able to go a long way in establishing that an algorithm does not behave as we would like. In recent years we have seen several influential studies of algorithmic bias in limited transparency settings. I will discuss just two here: The “Gender Shades” study of intersectional bias in commercial gender classification tools by Buolamwini and Gebru (2018), and the more recent work by Obermeyer, Powers, Vogeli, and Mullainathan (2019) on racial bias in an algorithm used nationwide to make health care case management decisions.

Buolamwini and Gebru (2018) evaluated several widely used commercial facial analysis systems on how well they perform gender classification. Their analysis revealed that the systems performed significantly better on images of men and light-skinned individuals than they did on women and dark-skinned individuals. In a follow-up study, Raji and Buolamwini (2019) report on the impact the Gender Shades work has had on the commercial tools targeted by their original black-box audit. The authors find that companies whose products were explicitly targeted had significant post-audit reductions in their systems’ racial and gender performance disparities. Without transparency into the training data and the models used it is impossible to definitively say what accounts for the observed improvement. However, the authors note that statements published by IBM and Microsoft “[imply] heavily the claim that data collection and diversification efforts play an important role in improving model performance across intersectional subgroups” (p.5). This type of bias—bias stemming from nonrepresentative data—is a common source of unfairness in algorithmic decision-making, but one that may be no easier to detect with a simple model than a black-box one.

In another study, Obermeyer et al. (2019) investigated a commercial algorithm widely used in the United States to determine which patients should be targeted with ‘high-risk care management’ programs. Patients selected into these programs benefit from additional resources devoted to ensuring that they receive high-quality, well-coordinated care. The researchers find that, across a range of different health markers, black patients are substantially less healthy than white patients who receive the same algorithmic score. That is, to the extent that the scores predict adverse health outcomes, the algorithm vastly underestimates risk for black patients relative to white patients. But why? It is not because the algorithm explicitly includes race—it does not. Rather, the issue is with the outcome the algorithm is designed to predict. The algorithm uses patient health costs as a

proxy for health needs. But because “[b]lack patients generate lesser medical expenses, conditional on health,” an algorithm that predicts costs will underestimate the health care needs of black patients relative to white patients. This is a failing of problem formulation. Predicting a proxy that is a biased measure of a target outcome is yet another common source of unfairness in algorithmic decision-making, but again one that may be no easier to detect with a simple model than a black-box one.

Where does this all leave us? First, I believe there is a tremendous amount of important technical and empirical work remaining to be done in the pursuit of algorithmic fairness. This work would be made easier if we had a system of laws and regulations that provided for meaningful transparency into the algorithms, the data on which they are trained, the outcome they aim to predict, and the use cases for which they have been validated. But it would by no means be made easy.

A large part of the difficulty resides in the problem formulation task. What is it that we are trying to achieve, and how do we map some part of that on to a data-driven task, be it prediction or something else? Even if we land on what we believe to be the right basic formulation, we may find the models produced rely too much on certain factors, or too frequently commit errors on cases where it is important to have high accuracy, or are otherwise imperfectly suited to the task at hand. We may then be able to mitigate some of these deficiencies. However, both theory and practice tell us that this will often require making trade-offs between competing desiderata. This in turn will require that we have an explicit understanding of precisely what those trade-offs are.

Along the way, it is critical that we not lose sight of the fact that most decisions do not neatly map on to a simple prediction task. Admissions is about more than GPA prediction, hiring is about more than forecasting hours worked, and sentencing has a lot more to do with blameworthiness and harm done than it does with predicting future arrests.

Disclosure Statement

Alexandra Chouldechova has no financial or non-financial disclosures to share for this article.

References

- American Law Institute (ALI). (2017). Model penal code: Sentencing.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91).

<https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/>

Chouldechova, A., & G'Sell, M. (2017). *Fairer and more accurate, but for whom?* Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML).

Kleinberg, J., & Mullainathan, S. (2019). *Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability*. Cambridge, MA: National Bureau of Economic Research.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Conference on Artificial Intelligence, Ethics, and Society. <https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/>

Rudin, C., Wang, C., & Coker, B. (2020). The *age* of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.6ed64b30>

Wisconsin v. Loomis. 881 N.W.2d 749 (2016)

©2020 Alexandra Chouldechova. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.