BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Research Publication No. 2018-8
December 2018

Assessing the Assessments: Lessons from Early State Experiences in the Procurement and Implementation of Risk Assessment Tools

Christopher Bavitz
Sam Bookman
Jonathan Eubank
Kira Hessekiel
Vivek Krishnamurthy

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu/ •
cyber@law.harvard.edu

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

ASSESSING THE ASSESSMENTS

Lessons From Early State Experiences In The Procurement And Implementation of Risk Assessment Tools

December 2018

by Christopher Bavitz, Sam Bookman, Jonathan Eubank, Kira Hessekiel, Vivek Krishnamurthy[1]

I. Introduction

Across the United States and around the world, local governments are procuring or developing applications known as "risk assessment tools" or "actuarial risk assessments" (collectively, "RAs") to aid decision-making in criminal courts. These tools are most commonly used in two contexts: pre-trial, informing a court's evaluations of whether a defendant should pay money bail or be subject to other conditions of release pending full adjudication;[2] and in post-trial sentencing decisions.[3] In these contexts, RA tools purport to predict the likelihood that a criminal defendant will reoffend or fail to appear for future proceedings.

Certain RA tools have been validated as significantly predictive,[4] and many criminal justice advocates argue that they open up an alternative pathway to highly-criticized systems of money bail.[5] Such targeted reform could theoretically reduce the United States prison population without compromising public safety. However, there is also evidence to suggest that some tools may amplify racial bias.[6] Still others are difficult to assess, as the results they generate and the methodology that underlies them are effectively uninterpretable to lawyers, judges, and the general public.[7]

1 The authors thank Berkman Klein Center Project Coordinator Adam Nagy for his significant substantive comments and feedback on this paper.

2 *Pretrial Risk Assessment*, Pretrial Justice Institute, http://www.pretrial.org/solutions/risk-assessment/ (last visited June 12, 2018).

3 *State Policies and Legislation*, National Center for State Courts (June 2018), https://www.ncsc.org/microsites/csi/home/In-the-States/State-Activities/State-Policies-and-Legislation.aspx (last visited July 23, 2018).

4 For a survey of validation methods and results, see Sarah L. Desmarais and Jay P. Singh, *Risk Assessment Instruments Validated and Implemented in the United States*, CJG Justice Center (Mar. 27, 2014), https://csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf (last visited July 18, 2018).

5 See e.g. Jonathan Lippman, *Our cash bail system isn't working. We can fix it.*, Washington Post, (Nov. 28, 2017), https://www.washingtonpost.com/opinions/our-cash-bail-system-isnt-working-we-can-fix-it/2017/11/28/3f0dd2ce-cf9f-11e7-a1a3-0d1e45a6de3d_story.html?noredirect=on&utm_term=.836bc22c76c7 (last visited July 19, 2018).

6 Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last visited June 12, 2018); Lauren Eckhouse, *Big data may be reinforcing racial bias in the criminal justice system*, Washington Post (Feb. 10, 2017), https://www.washingtonpost.com/opinions/big-data-may-be-reinforcing-racial-bias-in-the-criminal-justice-system/2017/02/10/d63de518-ee3a-11e6-9973-c5efb7ccfb0d_story.html?utm_term=.bb8405ee6b9d (last visited 23 July, 2018).

7 Megan Stevenson, *Assessing Risk Assessment in Action*, George Mason L. & Econ. Res. Paper No. 17-36, 3 (Aug. 29, 2017), https://ssrn.com/abstract=3016088 (last visited Feb. 14, 2018); *see also* Julia Dressel and Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 Sci. Advances 5580 (2018), available at http://advances.sciencemag.org/content/4/1/eaao5580 (last visited Feb. 14, 2018) ("the widely used commercial risk assessment

One pervasive set of questions about these tools concerns the extent to which they produce biased outcomes or reinforce and augment existing biases in the criminal justice system. For example, University of Michigan Law Professor Sonja Starr has argued that the use of RA tools at sentencing "violates the Equal Protection Clause and is bad policy."[8] Some widely-published statistical analyses seem to support this conclusion, although these analyses are not without their critics.[9]

Legislators typically make the policy decision to adopt RA tools, but the difficult task of developing, procuring, and implementing specific tools for specific purposes often falls to officials in the executive and judicial branches. For these officials, selecting the right pretrial risk assessment tool for their jurisdiction requires either choosing among a range of privately-developed tools, or deciding to commit to a long-term public development process. The technical underpinnings of RA tools are complex and the stakes involved in choosing them could not be higher, given that the criminal justice system involves the most fundamental decisions government can make (deprivation of individual liberty). It is vital that officials charged with procuring such systems dig deep on the technical issues, ask the right questions, and demand concrete answers of developers.

The difficulties of the procurement and implementation tasks in this arena are compounded by the fact that the officials charged with these responsibilities are not usually experts in statistical analysis, data science, and related fields. In some cases, the procurement or development of RA tools has been tasked to expert bodies such as Sentencing Commissions, which do possess broad-ranging expertise across a range of fields. More commonly, however, procurement is tasked to generalists who are responsible for contracting everything from stationery to uniforms. Even if procurement officials have some technical expertise, this may not include knowledge of criminal justice technologies. Informed and responsible procurement of RAs ultimately requires:

a robust understanding of the science behind algorithmic decision-making and the technical, business, legal, and other considerations that drive private developers in this space;

sustained engagement and consultation with relevant communities who are impacted by the use of these tools (particularly minority communities who are often overrepresented in the criminal justice system);

development of robust frameworks for post-procurement training and guidance on implementation; and

deep and sustained commitment to regular evaluation and collaboration with experts and the wider community around assessment of efficacy and bias.

software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise.")

8 Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 Stan. L. Rev. 803, 803, 819 (2014) (arguing that "[t]here is a strong case that most or all of the risk prediction instruments now in use are unconstitutional.")

9 *See, e.g.*, Julia Angwin et al., *supra* note 6; *see also* The 2017 Pulitzer Prize Finalist in Explanatory Reporting, The Pulitzer Prizes, http://www.pulitzer.org/finalists/julia-angwin-jeff-larson-surya-mattu-lauren-kirchner-and-terry-parris-jr-propublica (last visited Feb. 14, 2018); *cf.* Anthony W. Flores et al., *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,'* 80(2) Fed. Prob. J. 3 (2016).

This piece endeavors to provide context for state and local officials considering tasks around development, procurement, implementation, and use of RA tools. It begins in Part II with brief case studies of four states that adopted (or attempted to adopt) RA tools early on and describes their experiences. Part III draws lessons from these case studies and suggests some questions that procurement officials should ask of themselves, their colleagues who call for the acquisition and implementation of tools, and the developers who create them. This paper concludes in Part IV by examining existing frameworks for technological and algorithmic fairness. We offer a framework of four questions that government procurers should be asking at the point of adopting RA tools. Our framework draws from the experiences of the states we study and offers a way to think about:

Accuracy - the RA tool's ability to accurately predict recidivism.

Fairness - the extent to which an RA tool treats all defendants fairly, without exhibiting racial bias or discrimination.

Interpretability - the extent to which an RA tool can be interpreted by criminal justice officials and stakeholders, including judges, lawyers, and defendants.

Operability - the extent to which an RA tool can be administered by officers within police, pretrial services, and corrections.

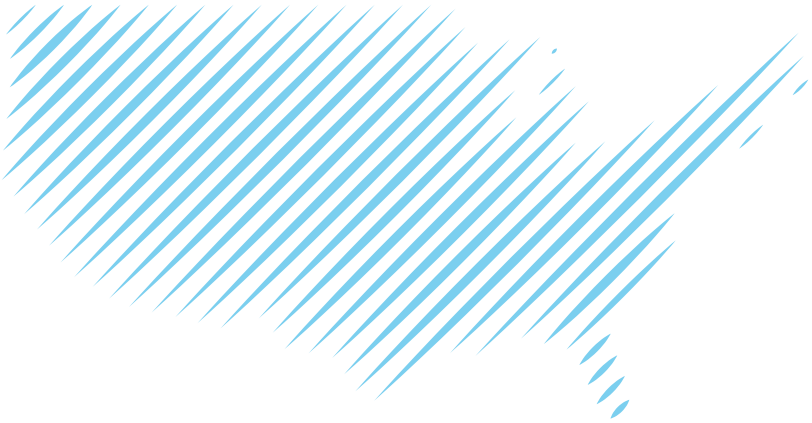
As set out in detail below, transparency about the inner workings of these tools comes through as an overarching theme. Ensuring that the workings of these products are open to scrutiny and review is certainly a key element in algorithmic fairness. It is tempting to fall back on the time-worn solution that "sunlight is [...] the best of disinfectants."[10] That said, a focus on algorithmic transparency is insufficient in two key respects:

First, the need for transparency extends beyond the inner workings of algorithms and the companies and organizations that develop them, to the government agencies that procure and implement tools and the data generated when those tools are used.

Second, transparency is not the only precondition for fair outcomes.[11] Even the most transparent of processes, leading to the procurement of the most transparent of tools, based on publicly-available data, is not guaranteed to produce outcomes that are fair and just. Instead, a comprehensive approach that adequately considers all phases of the development, procurement, and implementation processes is vital to ensure justice and fairness. Transparency may be a necessary condition for securing fair and unbiased outcomes, but it is not a sufficient one.

10 Louis D. Brandeis, Other People's Money—and How the Bankers Use It 92 (1914).

11 *See* Joshua A. Kroll et al, *Accountable Algorithms*, 165 U. Pa. L. Rev. 633 (2017).



II. Lessons from Early State Experiences

Overview

The growing number of jurisdictions that are considering the adoption of RA tools to assist in-court decisionmaking are doing so in the context of a vigorous social conversation on the efficacy and legitimacy of these tools. Jurisdictions now considering adoption of RA tools can benefit from the lessons of several states that were early adopters of these tools. These early-adopter states have had the opportunity to evaluate the use of tools in practice and—in some cases—revisit their initial decisions in the face of unexpected, inconsistent, or improper outcomes.

This set of case studies looks at how four states—Kentucky, Wisconsin, California, and Pennsylvania—implemented (or attempted to implement) RA tools. They offer important context on what mechanisms might help states confront issues such as algorithmic bias and inconsistent implementation. One key lesson is that any legislative mandate to adopt risk assessment tools must go hand-in-glove with careful research, clear policy direction, and an unequivocal commitment to re-evaluation. Another key lesson is that even the most carefully prepared tools can fail to be implemented fairly without careful communication and engagement with the local community.

Kentucky

Facing public safety concerns and rising corrections costs, in 2011 Kentucky's legislature passed an omnibus criminal justice reform bill.[12] The bill required, among other things, pretrial release for defendants who posed a low risk of flight or danger—a mandate that state institutions implemented by adopting an RA tool to assist with pre-trial decisions, which few other states had done at the time.[13]

In a study published in December 2017, Professor Megan Stevenson of George Mason University analyzed the results of the RA tool subsequently adopted in Kentucky. She detected concerning racially discriminatory effects.[14] Within each county, white and black defendants benefitted similarly from the implementation of risk assessments.[15] Yet across the state as a whole, she found a troubling trend: the adoption of risk assessment tools increased pretrial release rates for white defendants more than it did for black defendants.[16] This result stemmed not from bias in the tool itself, but from local differences in how it affected judicial behavior: judges in counties with a high proportion of white defendants were more likely to liberalize bail practices than those in counties with a higher percentage of black defendants.[17]

These situational complexities highlight "[t]he limits of enacting criminal justice reform via statute alone," according to Professor Stevenson.[18] Effective deployment of RA tools requires effective follow-through, from procurement to constant monitoring and validation. Kentucky has benefited from a Pretrial Services Office that worked hard to collect data, track the impact of implementation, and adapt responsively.[19] For instance, at the beginning of 2017, Kentucky mandated no-money pretrial release of all defendants charged with low-level crimes and classified as low- or moderate-risk, thus removing judicial discretion for this category of defendants.[20] This could alleviate some of the cross-county disparities.

In Kentucky, whether RA tools impacted individuals from different racial groups differently depended in large part on whether judges followed the tools' recommendations. New adopters can draw two key lessons from Kentucky's experience:

> **First, inherent bias in technology is not the only challenge states face. Varied implementation by human decision-makers can still cause racial disparities regardless of whatever biases are embedded into a tool.**

> **Second, the effect of adopting RA tools in any given jurisdiction will be hard to predict without express guidance to judges on how to incorporate RA results in their decisions. Any state considering the adoption of RA tools should take care to study their demographics, pretrial procedures, judicial training protocols, and incentives in order to tailor judicial guidance and minimize bias and disparities.[21]**

12 House Bill 463 Implementation Evidence-based Practices and Programs, Kentucky Department of Corrections 4–5 (Jan. 16, 2015), https://corrections.ky.gov/about/Documents/Research%20and%20Statistics/Annual%20Reports/HB%20463%20Report%20on%20Evidence%20Based%20Practices%20and%20Programs-FY%2014.pdf (last visited Feb. 15, 2018).

13 Mike Mullins, Public Safety and Offender Accountability Act (HB 463): Justice Reinvestment Summary, National Conference of State Legislatures 5, http://www.ncsl.org/documents/nalfo/JusticeReinvestmentMikeMullins.pdf (last visited Feb. 15, 2018).

14 Stevenson, *supra* note 7, at 3.

15 *Id.* at 5.

16 *Id.*

17 *Id.*

18 *Id.* at 58.

19 *Id.* at 59.

20 *Id.* at 59.

21 *Id.* at 57–58.

Wisconsin

Wisconsin began piloting the use of RA tools for sentencing in 2006 under its "Assess, Inform, Measure" (AIM) program.[22] Rather than being driven by legislators, the development of the AIM program was primarily led by judges and other court system stakeholders.[23] The original rationale of the program was to promote alternatives to incarceration by identifying low-risk offenders and triaging them to less severe sentences.[24] The AIM program introduced RA tools to eight counties throughout the state with the goal of improving the information available to sentencing judges. Under the AIM program, defendants were screened using an RA tool prior to sentencing. The results of the assessment were provided to a judge in a presentence investigation report. Although the format of reports was standardized, each county was allowed to select its own tool.[25] Several of the counties selected the LSI-R tool, while others, including La Crosse County, selected the COMPAS tool.[26] COMPAS is a product developed by a private company.[27] Significantly, however, COMPAS was not initially designed to inform a court's sentencing decision and was instead intended for use as a case management tool for corrections agencies.[28]

Thus, RA tools entered into the Wisconsin court system without a legislative mandate. Belatedly, in 2009, the Wisconsin State Legislature passed legislation requiring the use of RA tools but only as a mechanism for triage in the provision of post-conviction community services.[29] The Wisconsin Department of Corrections subsequently adopted COMPAS in 2010.[30] The Department of Corrections has described COMPAS as "the cornerstone of effective supervision" of community-based offenders.[31]

The absence of legislative mandate or statutory safeguards made it likely that the use of the COMPAS tool would be challenged in court. That challenge came in 2016. In *State v. Loomis*, a defendant from La Crosse County challenged the use of the COMPAS tool, which had been used by a state court in deciding to sentence him to six years' imprisonment for involvement in a drive-by shooting.[32] Mr. Loomis argued that because the

22 Suzanne Tallarico et al., *Effective Justice Strategies in Wisconsin: A Report of Findings and Recommendations*, Wisconsin Court System (2012), 14, https://www.wicourts.gov/courts/programs/docs/ejsreport.pdf (last visited July 17, 2018).

23 *Id.* at 30.

24 *Id.* at 30.

25 *Id.* at 15, 32.

26 Pamela M. Casey et al., *Use of Risk and Needs Assessment Information at Sentencing: La Crosse County, Wisconsin*, National Center for State Courts (Jan. 2014), http://www.ncsc.org/~/media/Microsites/Files/CSI/RNA%202015/Final%20PEW%20Report%20updated%2010-5-15.ashx (last visited July 17, 2018).

27 *COMPAS Classification*, Equivant, http://www.equivant.com/solutions/inmate-classification (last visited July 16, 2018).

28 *See further* Malenchik v. State, 928 N.E.2d 564 (Sup. Ct. Ind. 2010).

29 Corrections Code, Wis. Stat. § 301.068(3)(a) (2009).

30 Tallarico et al., *supra* note 22, at 41.

31 *COMPAS*, Wisconsin Department of Corrections, https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx (last visited July 16, 2018).

32 State v. Loomis 881 N.W.2d 739 (Wis. 2016).

algorithmic method required generalizations to be made with recourse to group factors (including gender), it violated his due process rights.[33]

Mr. Loomis's argument was complicated by the fact that COMPAS, as a privately owned instrument, relied upon a proprietary algorithm that was not available to either the defendant or the Court. The Court also acknowledged concerns about racially biased outcomes resulting from the COMPAS tool.[34] Nevertheless, the Wisconsin Supreme Court rejected Loomis's claim, finding that although the COMPAS algorithm grouped Mr. Loomis with past offenders with similar characteristics, COMPAS did not violate the due process right to an individualized sentence because it was not the *sole* basis for sentencing decisions.[35]

Significantly, however, the Wisconsin Supreme Court enumerated a set of safeguards that courts must apply when using the COMPAS tool. COMPAS may only be used to address treatment needs and the risk of recidivism, and not for the purposes of setting a sentence of incarceration.[36] Furthermore, where RA scores are included in presentence investigation reports, they must include a five-part written warning, specifying that:[37]

The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined.

Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders—not a particular high-risk individual.

Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having higher rates of recidivism.

A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed. Risk assessment tools must be constantly monitored and renormed for accuracy due to changing populations and subpopulations.

COMPAS was not developed for use at sentencing but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole.

33 *Id.* at 757.

34 *Id.* at 763.

35 *Id.* at 764-65.

36 *Id.* at 768-69.

37 *Id.* at 769.

The *Loomis* decision has drawn both praise and criticism. One expert described it as "one of the most sophisticated judicial treatments of risk assessment instruments" but nevertheless questioned whether lower court judges would heed the State Supreme Court's ruling.[38] Elsewhere, the decision has been criticized for "failing to specify the vigor of the criticisms of COMPAS, disregarding the lack of information available to judges, and overlooking the external and internal pressures to use such assessments".[39] Setting aside the merits of the *Loomis* decision, the Wisconsin experience raises questions as to why a RA tool was in such widespread use for 10 years before formal safeguards were implemented. Until the *Loomis* decision, COMPAS's use in sentencing was not subject to any regulatory limitations.

38 *See* the comments of Professor Christopher Slobogin in Lauren Kirchner, *Wisconsin Court: Warning Labels are Needed for Scores Rating Defendants' Risk of Future Crime*, ProPublica (July 14, 2016), https://www.propublica.org/article/wisconsin-court-warning-labels-needed-scores-rating-risk-future-crime (last visited July 17, 2018).

39 State v. Loomis: *Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*, 130 Harv. L. Rev. 1530, 1534 (2017).

California

While California has not yet adopted a statewide risk-assessment tool, a 2017 study by a workgroup commissioned by the California Supreme Court recommended abolishing money bail by statute and replacing it with a validated RA tool.[40] Currently, state law mandates "that each of California's 58 superior courts develop a uniform countywide schedule of [money] bail,"[41] limiting the judiciary's ability to institute statewide change. Instead, counties have significant leeway in the adoption of risk assessment tools. Consequently, nearly a dozen different pretrial risk assessment instruments were in use in California as of 2017, including COMPAS and PSA-Court, among others.[42]

Devolved decision-making has the potential to be more responsive to local criminal justice concerns, policy preferences, and procedures, but does not guarantee bias-conscious implementation. For example, Human Rights Watch described results in the Santa Cruz County courts in 2015, where"[j]udges agreed with 84 percent of the [RA tool's] 'detain' recommendations, but just 47 percent of 'release' recommendations."[43] The 2017 judiciary report noted, more broadly, that despite local use of RA tools, "California's current bail system... exacerbates socioeconomic disparities and racial bias."[44] The workgroup's recommendations are broad and holistic, targeting the following areas for reform (though not always in great detail):

Recognition of implicit bias. The workgroup notes that an RA tool must not exhibit "any implicit or explicit bias,"[45] an acknowledgement of the many forms bias can take. But bias can enter an RA-informed judgment at many points, from disparities in policing that affect detection of recidivism, to historical differences in court-ordered incarceration and treatment that disparately affect baseline frequencies of reoffending for some groups in society.[46] The workgroup's acknowledgement of "implicit" bias suggests it might urge courts to consider such factors, but a more robust framework would specify *how* this bias would be measured by providing the alternative baseline against which to compare RA results. Such a task is a societal choice, and should be made by policymakers rather than be left solely to statisticians and forensic psychologists.

Transparency. Promisingly, the workgroup specifies that both "the factors and algorithm used to determine risk levels" must be transparent—presumably requiring public disclosure.[47] This mandate would mark a significant improvement over common practice elsewhere by allowing researchers and the general public to examine the tool.

Acknowledgement of areas requiring specialized tools. The workgroup also writes that "cases involving intimate partner violence or sexual assault" call for "specialized risk assessment."[48] While this recommendation could also benefit from greater specificity, it acknowledges the reality that a general-purpose RA tool may not be the right tool for predicting certain types of risk.

40 Pretrial Detention Reform Workgroup, *Pretrial Detention Reform: Recommendations to the Chief Justice*, California Courts, 1 (2017), http://www.courts.ca.gov/documents/PDRReport-20171023.pdf.

41 *Id.* at 28; Cal. Pen. Code, § 1269b(c).

42 Pretrial Detention Reform Workgroup, *supra* note 40, at 102.

43 *"Not in it for Justice:" How California's Pretrial Detention and Bail System Unfairly Punishes Poor People*, Human Rights Watch (2017), https://www.hrw.org/report/2017/04/11/not-it-justice/how-californias-pretrial-detention-and-bail-system-unfairly.

44 Pretrial Detention Reform Workgroup, *supra* note 40, at 57.

45 *Id.* at 53–54.

46 *See, e.g.*, Stevenson, *supra* note 7, at 20–21.

47 Pretrial Detention Reform Workgroup, *supra* note 40, at 53–54.

48 *Id.*

A recently-introduced criminal justice reform bill in the California Senate spells out in more detail the factors that a pretrial RA tool should exclude, including education, employment, and housing status, and suggests regularly validating such tools and minimizing economic and racial disparities that may be embedded in criminal history.[49] This level of detail, though insufficient to fully address bias concerns, provides important specificity to evaluate particular RA tools. But it is no panacea: there is always a risk that proxy variables will introduce bias into the tools, and judicial override may continue to skew outcomes.

As California's experience with RA tools and the workgroup's research have shown, effective reform will take more than new statutory language. It will take transparency, robust and frequent testing, careful consideration of court processes and demographics, adjustment to contexts such as domestic violence or mental illness, and—perhaps most importantly—engagement with criminal justice stakeholders.

49 S. 10, § 1318.3 Reg. Sess. (Cal. 2017–2018), http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201720180SB10 (last visited Mar. 27, 2018).

Pennsylvania

Pennsylvania offers a recent example of an attempt by a state legislature to introduce RA tools to guide sentencing. In 2010, the Pennsylvania state legislature enacted legislation requiring the state's Sentencing Commission to develop "a sentence risk assessment instrument for the sentencing court to use to help determine the appropriate sentence" for offenders who plead or are found guilty, with the goal of decreasing incarceration rates in the state.[50] Since 2010, the Pennsylvania Commission on Sentencing (PCS) has been developing an in-house Sentencing Risk Assessment Instrument (SRAI) to implement this legislative mandate.[51]

The development of the SRAI has been far from smooth. The PCS intended to bring approval of the SRAI to a vote in June 2018, but strong local opposition has delayed any decision until December 2018.[52] Opposition has been grounded in two primary concerns:[53]

The reliance on static variables and juvenile conviction data. Some of the variables taken into account by the SRAI run against important ideas and intuitions about justice. For example, the tool relies only on "static" variables: those that do not change over time, such as gender, age, and prior convictions. Many of these variables are beyond an individual's control. All offenders under the age of 25 will have their age risk factor weighted the same, regardless of their temperament or personality. Furthermore, because it only evaluates static variables, the SRAI does not take into account rehabilitation efforts by an individual offender. An offender's prior convictions--including those from juvenile court--will always count against an offender, regardless of how much time has passed or what he or she may have done since.

Correlation between race and included variables, leading to disparate impact for African-Americans. The tool does not expressly include race as a variable, and the PCS has taken some steps to ameliorate the racial impact of SRAI (for example, by conducting a racial impact assessment and using conviction rather than arrest data).[54] Indeed, the PCS has claimed that the tool *underpredicts* recidivism for African-American offenders.[55] However, several of the SRAI's chosen variables, particularly prior drug and juvenile offending convictions, are likely to be more prevalent among African-Americans. This reflects a legacy of neglect and underinvestment in African-American communities, as well as a history of racist policing and prosecution.[56] As the Barristers' Association of Philadelphia, an African-American lawyers group, has alleged, "[b]y relying on data resulting from racial profiling and institutional racism, RAT [risk assessment tools] bakes the discrimination in the cake."[57]

50 42 Pa. C.S.A. § 2154.7.

51 Pennsylvania Commission on Sentencing, Proposed Risk Assessment Instrument (2018), http://www.hominid.psu.edu/specialty_programs/pacs/guidelines/proposed-risk-assessment-instrument (last visited Jun. 18, 2018).

52 Pennsylvania House Democratic Caucus, *McClinton motion delays vote on sentencing risk assessment tool*, PAHouse.com, June 14, 2018, 3:48PM, http://www.pahouse.com/InTheNews/NewsRelease/?id=98816 (last visited June 18, 2018).

53 Pennsylvania Commission on Sentencing, Testimony, 2018, http://www.hominid.psu.edu/specialty_programs/pacs/guidelines/proposed-risk-assessment-instrument/testimony (last visited June 18, 2018).

54 Pennsylvania Commission on Sentencing, Risk Assessment Project Phase III: Racial Impact of the Proposed Risk Assessment Scales, May 2018, http://pcs.la.psu.edu/guidelines/proposed-risk-assessment-instrument/additional-information-about-the-proposed-sentence-risk-assessment-instrument/racial-impact-analysis-of-the-proposed-risk-assessment-scales/view, (last visited June 18 2018); Pennsylvania Commission on Sentencing, Risk Assessment Update: Arrest Scales, Feb. 28, 2018, http://pcs.la.psu.edu/publications-and-research/risk-assessment/phase-iii-reports/risk-assessment-update-arrest-as-a-predictive-factor-2018/view (last visited June 18, 2018).

55 Risk Assessment Project Phase III: Racial Impact of the Proposed Risk Assessment Scales, *supra* note 53, at 1.

56 See e.g. Aaron Moselle, *Several African-American cops allege racism, corruption in Philly police unit*, Sept. 6, 2018, https://whyy.org/articles/several-african-american-cops-allege-racism-corruption-in-philly-police-unit/ (last visited June 18, 2018).

57 Charles M. Gibbs, *Testimony Before the Pennslyvania Sentencing Commission*, The Barristers' Association of Philadelphia (June 6, 2018), http://www.hominid.psu.edu/specialty_programs/pacs/guidelines/proposed-risk-assessment-instrument/testimony/charles-m.-gibbs-attorney-the-barristers-associa-

Pennsylvania's experience demonstrates the difficulty of securing community support for the introduction of new RA tools. Although the PCS conducted open hearings from an early stage in the SRAI's development, those hearings were so poorly attended that feedback deadlines had to be extended.[58] The racial impact of the SRAI was left open to confusion after the PCS decided to rely on *arrest* data–which would have amplified racial disparities–before reverting to *conviction* data instead.[59] In public testimony on the SRAI, it was clear that this reversion had not been communicated to all community groups.[60] The lack of widespread community support has postponed the SRAI's implementation, and it remains to be seen whether it will be adopted.

tion-of-philadelphia.-philadelphia-june-6-2018/view (last visited July 2, 2018).

58 Anna-Maria Barry-Jester, Ben Casselman and Dana Goldstein, *Should Prison Sentences be Based on Crimes that Haven't Been Committed Yet?*, FiveThirtyEight, Aug. 4, 2018, https://fivethirtyeight.com/features/prison-reform-risk-assessment/#risk-assessment-doesnt-eliminate-bias (last visited June 18, 2018).

59 Risk Assessment Project Phase III: Racial Impact of the Proposed Risk Assessment Scales, *supra* note 54, at 1.

60 *See e.g.* Petra K. Gross, *Testimony of Petra K. Gross Before the Pennsylvania Sentencing Commission on the Preliminary Sentencing Risk Assessment Instrument*, Pennsylvania Association of Criminal Defense Lawyers (June 14, 2018), http://www.hominid.psu.edu/specialty_programs/pacs/guidelines/proposed-risk-assessment-instrument/testimony/petra-k.-gross-attorney-pa-association-of-criminal-defense-lawyers.-harrisburg-june-13-2018/view (last visited July 23, 2018).

III. Lessons from Early State Experiences

A set of actionable lessons emerges from the varied experiences of early adopters. Kentucky encountered unexpected variation across judicial districts, including racial disparities. It relied on a robust data infrastructure and institutional willingness to revise processes as a way to reduce the problems that arose. Wisconsin failed to set limits on the use of RA tools at the outset, forcing its Supreme Court to establish parameters in a statutory vacuum. California, facing a patchwork of localities using different algorithms, benefited from a yearlong workgroup that investigated concerns about bias in pretrial risk assessments, but that failed to settle several key questions. In Pennsylvania, the SRAI has failed to get off the ground because of preemptive concerns about fairness. These lessons can be tied to four phases in adopting an RA tool: development, procurement, implementation, and testing.

When developing and procuring an RA tool, research, transparency, and flexibility in approach are essential to addressing bias. In implementation, states should establish processes to ensure judges and other users administer tools correctly and in ways that mitigate bias. Specifically, states should:

Enumerate early the types of information that should and should not be included in an RA tool's inputs. In doing so, they should take special care to identify and exclude potential proxies for race, socioeconomic status, and other inputs that will result in biased scores. In some cases, as the Pennsylvania experience shows, even relatively transparent and sophisticated attempts to eliminate these proxies may fail to satisfy community concerns.

Clearly explain how judges and other users should factor RA scores into their decisions. As Kentucky's experience shows, states must understand how decision-makers weigh RA results in order to anticipate potential biases in outcomes. State governments should set clear standards for how decision-makers factor RA results into their judgments through statutes, judicial rules, or other regulations. The experience of Wisconsin demonstrates that if such guidelines are not clearly established from the outset, they may have to be determined by appellate courts on an ad hoc basis.

Provide training on interpreting results to judges and other users. In addition to examining and setting legal rules to limit judicial discretion, states should work to standardize how decision-makers *interpret* the results of RA tools by investing in repeated training to reduce the likelihood of large discrepancies between individual judges or users.

Finally, when states test and evaluate an RA tool, they should:

Lay out transparent criteria to evaluate efficacy and bias. Not only should the RA tool's inputs and algorithm be available to the public, but so should measures of its predictive accuracy and of bias. Rather than signing non-disclosure agreements with RA tool developers (as Wisconsin has done),[61] states should use their purchasing power to mandate transparency from outside developers. Measures should be clear and specific. Algorithms are trained toward certain ends, and if those ends are uncertain, so too will the tool's outcomes.[62]

Make adjustments to the choice of tool, its design, and use policies after implementation. States must recognize that their first attempts to implement a tool—and perhaps even the tool itself—will require iteration. They should also recognize that some contexts, like domestic violence, may need a specialized approach.[63] Wherever they contract with outside developers, contracts must allow for significant flexibility and constant review.

Involve outside researchers and journalists in testing and evaluation discussions. Academics and journalists have conducted some of the most extensive efforts to date to understand and analyze the impact of RA tools. They provide a valuable—and neutral—resource for states approaching this burgeoning field.

States considering incorporating RA tools can learn from these experiences and improve by considering implementation and evaluation needs early on, when they procure RA tools, negotiate contracts, and draft legislation. But these lessons are equally salient for states that have already adopted RA tools, as they engage in the ongoing processes of testing, public disclosure, training, and review.

61 *See also* the example the experience of New Mexico: Memorandum of Understanding Between Laura and John Arnold Foundation and Bernalillo County Stakeholders, US Bail Reform News 4 (2016), https://www.usbailreform.com/wp-content/uploads/2017/04/Arnold-Foundation-Bernalillo-County-Agreement.pdf.

62 *See* Deven R. Desai and Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 Harv. J. L. & Tech. 1 (2017).

63 For example, several jurisdictions use a standalone tool to assess risk in cases of domestic violence. See for example the Ontario Domestic Assault Risk Assessment (ODARA). *See ODARA 101*, Waypoint Centre, http://odara.waypointcentre.ca/ (last accessed July 2, 2018).

IV. Looking Ahead

Having identified several key lessons learned from early state experiences with RA tools, we now turn to the question of how procurers should frame their decisions when adopting particular instruments. We first identify existing models of impact frameworks, before proposing a four-question framework of our own.

Human Rights Impact Assessment Frameworks

A common way of assessing the impact of significant projects is through an impact assessment, a set of questions designed to guide due diligence. In recent years, such assessments have increasingly considered issues of due process and human rights. In 2011 the United Nations Human Rights Council adopted a set of Guiding Principles for Business and Human Rights ("the Principles").[64] The Principles specify the need to conduct a Human Rights Impact Assessments ("HRIA")—a particularly detailed form of human rights due diligence—whenever there appear to be severe risks to human rights.[65] Although these assessments are designed for private companies rather than government agencies (which are legally held accountable to standards under formal treaty-based law), the Principles and HRIA offer one practical guide for procurers in the field.

Principle 4 of the UN Guiding Principles requires due diligence to be undertaken where "the nature of business operations or operating contexts pose significant risk to human rights." RA tools may affect a number of internationally recognized rights, such as the rights to liberty, freedom from arbitrary detention, a fair and public hearing, the right to be informed of evidence against oneself, and the right to equality and freedom from discrimination.[66] An HRIA could be structured around the impacts of RA tools on these rights.

One strength of the HRIA approach is that it clearly identifies the rights that need to be protected, with succinct recourse to statements of international law. Furthermore, such a framework would have international applicability, and draw on an existing body of templates—such as those produced by the Danish Institute for Human Rights[67] and the law firm Foley Hoag[68]—which have developed sophisticated HRIA models.

However, there may be some limitations associated with HRIAs. Fundamentally, existing HRIA

64 Olivier De Schutter, *Guiding principles on human rights impact assessments of trade and investment agreements: Report of the Special Rapporteur on the right to food, Olivier De Schutter*, United Nations General Assembly (December 19, 2011), https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session19/A-HRC-19-59-Add5_en.pdf (last visited June 12, 2018).

65 The Danish Institute for Human Rights, *Human rights impact assessment guidance and toolbox*, https://www.humanrights.dk/business/tools/human-rights-impact-assessment-guidance-and-toolbox (last visited June 12, 2018).

66 International Covenant on Civil and Political Rights (adopted December 16, 1966, entered into force March 23, 1976) https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx (last visited June 12, 2018).

67 *See* The Danish Institute for Human Rights, *supra* note 65.

68 *Practices: Corporate Social Responsibility*, Foley Hoag LLP, http://www.foleyhoag.com/practices/business/corporate-social-responsibility (last visited June 12, 2018).

frameworks have been designed to identify static variables. The most common application is in the extractive industries context: a HRIA helps a company assess what effect a mine will have on the local population, taking into account where people live, how they generate income, and where they get their food and water from. Such impacts can be determined and assessed with reasonable certainty based on past experiences, which will usually be highly relevant to the situation at hand. One of the difficulties in designing impact assessment frameworks for algorithms is that the technology and its likely impacts are constantly shifting. Identifying the potential discriminatory impact of an algorithm depends on how racial patterns of prediction play out over time, which can be very difficult to build into a one-time HRIA designed to inform a particular procurement process.

A second problem may be that the due diligence process reveals that two tools under consideration raise no baseline human rights concerns. This is because human rights are often interpreted as minimum standards, rather than criteria for comparative evaluation. When this occurs, an HRIA may be ineffective in assisting a decisionmaker in deciding how to choose between two different tools. Some other mechanism would then be required to supplement an HRIA.

Data Protection Impact Assessments

Impact assessments are gradually becoming a core expectation within the technology sector. The European Union has recently made it compulsory for certain controllers of personal information to conduct a "data protection impact assessment" (DPIA) wherever significant rights and freedoms are stake.[69] A DPIA comprises four components:

> a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller;
>
> an assessment of the necessity and proportionality of the processing operations in relation to the purposes;
>
> an assessment of the risks to the rights and freedoms of data subjects [...]; and
>
> the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation, taking into account the rights and legitimate interests of data subjects and other persons concerned.

In the United States, there is evidence that state and city authorities are similarly taking privacy and data due diligence seriously. The City of Oakland, for example, has created a Privacy Commission to advise the city on equipment and systems with data or privacy implications, conducting open hearings and publishing public reports.[70]

These processes and frameworks offer a useful way to structure assessments of new technologies. The European DPIA is particularly useful in encouraging private and government actors to pause and consider the data implications of specific systems: something that is not intuitive to everyone. Importantly, it represents a firm regulatory commitment to mainstream this approach throughout all of government (as well as the private sector). But as the earlier case studies illustrate, algorithms raise their own unique concerns. They require a framework which can guide a procurer through the particular pitfalls that algorithms create in the criminal justice context.

69 2018 O.J. (L 119) General Data Protection Regulation Art. 35.

70 City of Oakland, *City Administration*, http://www2.oaklandnet.com/government/o/CityAdministration/d/PrivacyAdvisoryCommission/index.htm (last visited June 18, 2018).

The AI Now Algorithmic Impact Assessment

Recently, the AI Now Institute, a project affiliated with New York University, has suggested a four-part algorithmic impact assessment ("AIA").[71] The proposed AIA is not yet a tool that can used by government procurers, and is instead a broad framework for the integration of critical thinking about automated decisionmaking throughout the whole of government. The four goals included in AI Now's AIA are: (1) to provide the public with information about the algorithmic systems; (2) to give external researchers meaningful access to review and audit systems; (3) to increase the capacity within public agencies to assess fairness, due process and disparate impact; and (4) to strengthen due process by offering the public the opportunity to engage with the AIA process before, during and after the assessment.

This AIA framework is a useful start in building an assessment that can assist government procurement officers. WIth its heavy reliance on the principle of algorithmic transparency, the AIA framework might be able to overcome some of the shortcomings of traditional HRIAs. Our focus in this article is slightly different, however. Rather than suggesting a procedural framework for how tools are assessed, it suggests four substantive areas for evaluation. Procedural and substantive rigor are both vital in developing an effective assessment model.

71 Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker, *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, AI Now (April 2018). https://ainowinstitute.org/aiareport2018.pdf (last visited June 12, 2018).

A Modest Proposal: Four Questions

Against that backdrop, four questions emerge that procurers should ask when selecting an appropriate RA tool. Our 4-question framework sharpens the necessary analysis at the "self-assessment" phase of AI Now's proposed framework. In so doing, our proposal provides a useful outline for risk assessment tool procurement. Our framework is based around four crucial concepts: accuracy, fairness, interpretability and operability. We have framed these questions to assist both seasoned experts in data science, and everyday government procurers. We emphasize, however, that they are a starting point for further analysis, and not an exhaustive model of assessment.

Question One
Is the Tool Accurate?

Perhaps one of the most obvious questions that procurers should be asking is whether the tool has a high rate of predictive accuracy. The recent ProPublica expose of the COMPAS tool in Florida suggested that, among those already classified as being at a higher risk of recidivism, it was only marginally more accurate than a coin toss.[72] That is simply not good enough: a tool should, at the very least, have a comparable or higher rate of predictive accuracy than any non-algorithmic tool (such as structured expert judgments).[73] Government procurers should invite and review independent validation studies of the tool's accuracy, and not simply rely on claims made by its developers. Crucially, they should carefully consider whether validation studies performed in other jurisdictions will indicate predictive accuracy in their own: the fact that a tool is accurate elsewhere does not necessarily mean it will be predictive here.[74]

Furthermore, it is vital that the tool's predictive accuracy be made available to all stakeholders within the justice system. No risk algorithm will be a perfect predictor. Even some of the most widely used tools have accuracy scores of 0.65-0.75,[75] under the widely used "area under curve" metric.[76] There is always a risk that once a formal score is produced in the context of a particular bail or sentencing hearing that it is taken as definitive. It has been suggested that data-driven scores could lead to "automation bias" or "quantification bias", whereby decision-makers favor factors that have been computer-generated[77] or

72 Julia Angwin et al., *supra* note 6.

73 For an empirical assessment of this question with specific reference to the COMPAS tool, see Julia Dressel and Hany Farid, *The accuracy, fairness and limits of predicting recidivism*, 4 Sci. Adv. 1 (2018).

74 For example, the California Department of Corrections and Rehabilitation has concluded that although the COMPAS Tool may be predictive in other jurisdictions, it lacks predictive accuracy in California. See Jennifer L. Skeem and Jennifer Eno Loudon, R*eport Prepared for the California Department of Corrections and Rehabilitation (CDCR)*, Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) at 5 (2007).

75 Tim Brennan, William Dieterich, and William Oliver, *Risk Assessment,* Criminology and Criminal Justice (Sept. 2017) http://criminology.oxfordre.com/view/10.1093/acrefore/9780190264079.001.0001/acrefore-9780190264079-e-100 (last visited June 12, 2018).

76 *See* Marie E. Rice & Grant T. Harris, *Comparing Effect Size in Follow-Up Studies: ROC Area, Cohen's d, and r*, 29 L. & Hum. Behavior 615, 618 (2005) ("AUC equals the probability that a score (on an ordinal or continuous measure such as a risk-assessment instrument) drawn at random from one sample or population (e.g., recidivists' scores) is higher than that drawn at random from a second sample or population (e.g., nonrecidivists' scores.")

77 *See* Danielle Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249 (2008).

can be given a numerical value.[78] Decision-makers must be able to contextualize a risk score. By ensuring that everyone in the system knows how accurate the tool is, they are better to make decisions that give the tool appropriate, rather than exclusive, weighting.

Question Two
Does the RA Tool Account for Bias and Discrimination?

At least in theory, risk assessment tools should be able to reduce the racial discrimination that pervades the American criminal justice system. If that discrimination is at least in part explainable by human factors, such as overt or implicit racial bias, then transferring part of decision-making to an automated system should reduce overall bias.

The reality is more complex. Different studies have reached different conclusions. ProPublica, an investigative journalism organization, claimed that holding all other variables constant, the COMPAS tool erroneously labeled black defendants as likely to reoffend at twice the rate as it did white defendants (false positive) and also mislabeled white defendants who did go on to reoffend as low risk at a greater rate than black defendants (false negative). [79] The results of that investigation have been contested by some scholars, leading to widespread public controversy.[80]

Beyond methodological disagreements, the heart of the COMPAS controversy is an argument about the definition of fairness. The ProPublica study acknowledged that the COMPAS algorithm correctly predicted reoffending among whites and blacks at the same rate. If the definition of fairness is "correctly predicting reoffending in equal measure across different groups", then COMPAS would pass the test. However, the key finding of the ProPublica was that it was twice as likely to misclassify black offenders as likely to reoffend, as it was for white offenders (42% versus 22%).[81] In other words, if the definition of fairness is "has an error rate consistent across all groups", then COMPAS appears to fail the test.

A starting point for procurers is to set a definition of fairness. This task is nuanced. RA tools don't explicitly use race as a risk factor. But as noted in the case studies above, they often weigh factors in which racial minorities (and particularly African-Americans) are over-represented, such as unemployment and prior convictions. The mere use of such variables will not necessarily lead to racial bias, but as the early experience demonstrates, in many instances they will. Each tool must be carefully validated, and meaningful engagement with affected communities is vital.

Furthermore, RA tools are trained on datasets in which minorities are already over-represented, meaning that there is a risk, through what some

78 *See e.g.* Tricia Wang, *The Human Insights Missing from Big Data*, TED, Sept. 2016, https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data (last visited July 2, 2018).

79 *See* Angwin, *supra* note 6.

80 *See* Flores et al., *supra* note 9.

81 Sam Corbett-Davies et al., *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.*, Washington Post (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.832e06cd60fc (last visited July 23, 2018).

scholars have called "zombie predictions",[82] that the tools will simply replicate or amplify existing patterns of racial discrimination. Indeed, Human Rights Watch has called for the abandonment of algorithmic RA tools on the basis that the elimination of racial bias may not be possible.[83]

Procurers doing due diligence on a new risk assessment tool should carefully scrutinize tools for any evidence of racial bias. The importance of risk factors which closely correlate to characteristics of racial minority defendants should be examined, and developers should be required to explain why they are necessary.[84]

Ex post, validation studies of the tool should assess racial bias. Policymakers must be ready to confront difficult ethical questions. Ultimately, racial bias can only be eliminated if we have a clear idea of what it is.[85] This is a moral and policy question that cannot be left to tool developers alone. For instance, what if the overall racial bias of the tool is the same as existing processes, but the overall number of defendants denied bail has fallen? Should this be considered an overall improvement which would justify the retention of the tool? This trolley-problem-esque question has no easy answer, but should placed firmly on the table for public debate. Ultimately, policymakers and procurers must be clear what their definitions and parameters of fairness are. In some instances, such as the competing definitions of fairness in the ProPublica tool, it will be mathematically impossible to satisfy *all* competing definitions.[86]

All these issues of fairness demand that the algorithm be transparent. So too must the datasets, subject to competing demands of individual privacy.[87] Variables cannot be interrogated and validated for bias unless the algorithm is publicly known. Unfortunately, many privately-developed algorithms are closely-guarded trade secrets, as was the case in Wisconsin. Fairness and non-discrimination cannot be guaranteed without transparency.

82 John Logan Koepke, David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, Wash L. Rev. (March 21, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3041622 (last visited June 12, 2018).

83 John Raphling, *Human Rights Watch advises against using profile-based risk assessment in bail reform*, Human Rights Watch (July 17, 2017), https://www.hrw.org/news/2017/07/17/human-rights-watch-advises-against-using-profile-based-risk-assessment-bail-reform (last visited June 12, 2018).

84 *See* George Joseph, *Justice by Algorithm*, CityLab (December 8, 2016) https://www.citylab.com/equity/2016/12/justice-by-algorithm/505514/ (last visited June 12, 2018).

85 *See* Kroll, *supra* note 11.

86 See Corbett-Davies et al., supra note 81; see also Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, Proceedings of Innovations in Theoretical Computer Science (2017).

87 For an attempt to reconcile these competing values, *see* Michael Veale and Reuben Binns, *Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data*, Big Data & Society 1 (July-December 2017).

Question Three
Can the Tool's Outputs be Interpreted by Everyone in the Criminal Justice System?

The adversarial criminal justice system relies on judges, defendants and their lawyers being able to understand the nature of the evidence before them. This is reflected in numerous constitutional and human rights instruments, from the United States Constitution[88] to the International Covenant on Civil and Political Rights.[89] This right has recently been recognized in Europe with specific reference to algorithms, as the GDPR creates a "right to explanation".[90]

If defendants are unaware of their risk score, or unaware of what factors have contributed to that score, they will be unable to rebut the powerful and reified numerical score that is presented to a judge. There are three reasons why it is important for defendants to know their risk score and the factors which produced it. The first is that information inputted into the tool could be incorrect, as a result of human error, a misspeak by the defendant during an interview, or an incorrect police record. Secondly, the defendant should have an opportunity to challenge the overall reliability of the algorithm.[91] Although this may be beyond the means of most criminal defendants, it is an important source of accountability to ensure that the assessment is reliable. Finally, the defendant should have an opportunity to explain why, although the RA tool has identified several risk factors and as a result received a high score, the algorithm has got it wrong in their particular case. For example, the algorithm may have scored them at a greater risk of not appearing because of historic substance abuse, whereas the defendant may be able to show that they have overcome those issues. This last reason explains some of the opposition to the inclusion of static variables in the Pennsylvania SRAI.

It is also important that judges and bail commissioners can interpret the tool. A numerical score does not tell a judge much in and of itself. The score must be capable of being translated into some kind of non-expert predictive terminology.

Judges should know whether or not the tool is recording dynamic factors which could be reduced through effective counselling or social services. Many tools take account of these factors in order to identify defendants who might be able to be helped through social services as they go through the criminal justice system.[92] But if judges are unaware that these factors make up part of the risk assessment score—regardless of whether that defendant has taken steps to improve substance abuse—then those factors will cloud the accuracy of the tool as a predictor of future risk.

88 U.S. Const. art. V.

89 *See* International Covenant on Civil and Political Rights, *supra* note 64.

90 *See* Bryce Goodman and Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"*, 38 AI Mag. 50 (2017); Andrew D. Selbst and Julia Powles, *Meaningful information and the right to explanation*, 7 Int. Data Privacy L. 233 (2017). For an argument that the GDPR does not create such a right, *see* Sandra Wachter, Brent Mittelstadt and Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the European General Data Protection Regulation*, 7 Int. Data Privacy L. 76 (2017).

91 *See* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stanford Law Review (forthcoming 2018) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920883 (last visited June 12, 2018).

92 *See* Joseph, *supra* note 84.

Another piece of information that judges need is precisely *what* risk the tool is identifying - a RA tool must be fit for purpose. A risk score is more than just a number. It is identifying something specific, such as the risk of *violent* or *general* offending, or the risk of offending within a particular time period. At the bail stage, it may be relevant to the decisionmaker's assessment whether the defendant is likely to reoffend only in the short-term, while they are awaiting trial. If the RA tool is assessing risk over a five-year period, it is less useful. At the sentencing stage, it may be important for the judge to know whether the score is assessing particular types of offending, such as violent offending or offenses against children. It is vital that judges and other stakeholders understand how the tool's score is produced, and the implications for its accuracy in each situation.

Engaging judges and other decision-makers is a crucial task. In many instances, this will require a basic educational program in statistics. Many judges have been crying out for such education.[93] Part of the very rationale for RA tools is to overcome the implicit bias of these adjudicators. Implicit bias is notoriously difficult to overcome, and may not be possible unless judges have enough confidence in the tool so that it can overcome their preconceived notions about defendants.

On a wider level, interpretability is vital to ensuring public confidence in the justice system. It is not good enough for a defendant, querying why he or she has been denied bail or set a very high bond, to be told "because a risk score says so". Stakeholders, including community organizations, must be included in the procurement or development process. This is not easy feat: as a study by the Marshall Project and FiveThirtyEight observed during the development of the SRAI, public hearings by the Pennsylvania Sentencing Commission were so poorly attended that public comment periods had to be extended.[94] But without effective community engagement, the tools adopted will lack public confidence. Here, analogies to campaigns to improve credit scores may be fruitful.[95] The defendant should be able to look behind the numerical score and know the exact reasons for the court's assessment of the severity of the flight risk that they pose.

Question Four
Can the Tool be Reliably and Easily Administered?

Finally, even a perfectly-designed tool may be useless or dangerous if it is not properly administered. Here, simplicity is an important factor. Some risk tools require answers to hundreds of questions, amplifying the risk of human error on the part of the interviewer or misspoken responses from the interviewee. Procurers should be satisfied that their institution has the expertise to administer the tool, and that the developer will provide ongoing training and support.

93 *See e.g.* the results of a survey of judges in Wisconsin prior to the Loomis decision, reproduced in Tallarico et al., *supra* note 22, at 51.

94 Jester et al., *supra* note 58.

95 *What's in my FICO Scores*, myFICO, https://www.myfico.com/credit-education/whats-in-your-credit-score/ (last visited June 12, 2018); *see also* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1 (2014), where the authors highlight opacity and complexity as barriers to individuals being able to challenge their credit scores.

Even simple things such as the framing of interview questions can impact the risk score produced. For example, asking a defendant "are you using drugs at the moment?" as opposed to "do you use drugs?" may produce different responses.[96] Procurers should be aware of cultural barriers within relevant communities that might make it difficult for some defendants to comprehend the questions asked. If necessary, tools and questionnaires should be carefully translated into other languages.

Furthermore, a tool should be capable of producing predictions based on risk factors that can be gathered from a relatively small range of sources. Tools that require information to come from multiple government agencies, as well as defendants and victims, could prove too costly or impractical to administer, and increase the potential for human error. One danger here is that predictions are made based on incomplete information with certain risk factors missing. Procurers must be confident that they have the resources not only to procure the tool in the first place (such as licensing fees), but also to administer it on an ongoing basis.

Principles of interpretability and operability may come into conflict with the equally important principles of accuracy and fairness. For example, it may be the case that a more sophisticated algorithm, taking into account a large number of dynamic variables from a range of sources, has a higher level of accuracy than a simple tool using a small range of static variables. Ultimately, however, a tool is only as good as the manner in which it is used. The Kentucky experience discussed above is illustrative: inconsistent applications of the same tool may yield arbitrary or discriminatory results. A complex or confusing tool will only exacerbate these concerns. It is up to procurers to fine-tune to difficult balance between the four principles we have identified.

96 *See* Joseph, *supra* note 83.

V. Conclusion

Much has been written about the use of algorithms in the criminal justice system. We do not purport to canvass the entire debate, nor do we endorse or reject their use on philosophical grounds. We offer this piece as algorithms become an increasingly central part of the criminal justice landscape, conscious that serious questions need to be asked about how developers and procurers are going about their work. While the framework we offer is far from comprehensive, it provides a starting point for a more detailed impact assessment process to avoid some of the pitfalls encountered by early adopters.

None of the four proposed questions can be assessed by procurers without transparency. Procurers need to have access to the algorithm and enough information to conduct and interpret validation studies. They need to be able to effectively pass that information on to judges, defendants and other stakeholders in the criminal justice system. Risk assessment tools require algorithmic transparency.[97]

Transparency is not a panacea, however. Although it allows for an iterative process and watchdogs to call out poor development and procurement, these decisions need to be right from the start. The four questions offered in this paper are one possible starting point for a more formalized impact assessment for risk assessment tools. We hope that they will pave the way for a culture among government procurers which takes the efficacy, risks and dangers of these tools seriously. The integrity of the criminal justice system depends on it.

97 *See* Electronic Privacy Information Center, *Algorithms in the Criminal Justice System*, https://epic.org/algorithmic-transparency/crim-justice/ (last visited June 12, 2018).

Permalink: https://cyber.harvard.edu/publication/2018/assessing-assessments

Suggested Citation: Bavitz, Christopher and Bookman, Sam and Eubank, Jonathan and Hessekiel, Kira and Krishnamurthy, Vivek; Berkman Klein Center Research Publication No. 2018-8. December 2018. Available at: https://cyber.harvard.edu/publication/2018/assessing-assessments

Layout: Daniel Dennis Jones