TECHNICAL FLAWS OF PRETRIAL RISK ASSESSMENTS RAISE GRAVE CONCERNS

SUMMARY

Actuarial pretrial risk assessments suffer from serious technical flaws that undermine their accuracy, validity, and effectiveness. They do not accurately measure the risks that judges are required by law to consider. When predicting flight and danger, many tools use inexact and overly broad definitions of those risks. When predicting violence, no tool available today can adequately distinguish one person's risk of violence from another. Misleading risk labels hide the uncertainty of these high-stakes predictions and can lead judges to overestimate the risk and prevalence of pretrial violence. To generate predictions, risk assessments rely on deeply flawed data, such as historical records of arrests, charges, convictions, and sentences. This data is neither a reliable nor a neutral measure of underlying criminal activity. Decades of research have shown that, for the same conduct, African-American and Latinx people are more likely to be arrested, prosecuted, convicted and sentenced to harsher punishments than their white counterparts. Risk assessments that incorporate this distorted data will produce distorted results. These problems cannot be resolved with technical fixes. We strongly recommend turning to other reforms.

ACTUARIAL RISK ASSESSMENTS DO NOT ACCURATELY MEASURE PRETRIAL RISKS

When making pretrial release decisions, judges must impose the least restrictive conditions of release necessary to secure the presence of a person at trial and protect the safety of the community. To accomplish this task, judges must identify and mitigate certain pretrial risks, specifically of a person causing serious harm to the community or fleeing the jurisdiction prior to their trial. Today's pretrial risk assessments are ill-equipped to support judges in evaluating and effectively intervening on these specific risks, because the outcomes that these tools measure do not match the risks that judges are required by law to consider. For example, many risk assessments only provide a pretrial failure risk score, which is a combined outcome of missing a court appearance or being rearrested. Many scholars have warned that such a composite score could lead to an overestimation of both flight and danger, and can make it more, not less, difficult to identify effective interventions.[1]

Even when pretrial risk assessments break out risk scores into distinct categories, the data used to define and measure flight and danger are inexact and overly broad. For example, risk assessments frequently define public safety risk as the probability of arrest.[2] When tools conflate the likelihood of arrest for any reason with risk of violence, a large number of people will be labeled a threat to public

[1] *E.g.*, Lauryn P. Gouldin, *Disentangling Flight Risk from Dangerousness*, 2016 BYU L. Rev. 837, 887-88 (2018). The interventions which improve an individuals likelihood of appearing in court (text reminders, transportation services, flexible scheduling) are often quite different from interventions designed to ensure community safety (stay-away orders, curfews, drug testing).

[2] For example, the Colorado Pretrial Assessment Tool (CPAT) defines a risk to "public safety" as any new criminal filing, including for traffic stops and municipal offenses. The Colorado Pretrial Risk Assessment Tool Revised Report 18 (2012).

safety without sufficient justification. Risk assessments that include minor offenses, such as missing a court-debt payment, in their definition of danger, run the risk of increasing pretrial incarceration rates and further exacerbating racial inequalities in pretrial outcomes.[3]

Some risk assessments define public safety risk more narrowly as the risk that a person will be arrested for a violent crime while on pretrial release. But because pretrial violence is exceedingly rare, it is challenging to statistically predict. Risk assessments cannot identify people who are more likely than not to commit a violent crime. The fact is, the vast majority of even the highest risk individuals will not go on to be arrested for a violent crime while awaiting trial. Consider the dataset used to build the Public Safety Assessment (PSA): 92% of the people who were flagged for pretrial violence did not get arrested for a violent crime and 98% of the people who were not flagged did not get arrested for a violent crime.[4] If these tools were calibrated to be as accurate as possible, then they would predict that every person was unlikely to commit a violent crime while on pretrial release. Instead, risk assessments sacrifice accuracy and generate substantially more false positives (people who are flagged for violence but do not go on to commit a violent crime) than true positives (people who are flagged for violence and do go on to be arrested for a violent crime).[5] Consequently, violence risk assessments could easily lead judges to overestimate the risk of pretrial violence and detain more people than is justified.[6]

Finally, current risk assessment instruments are unable to distinguish one person's risk of violence from another's. In statistics, predictions are made within a range of likelihood, rather than as a single point estimate. For example, a predictive algorithm might confidently estimate a person's risk of arrest as somewhere between a range of five and fifteen percent. Studies have demonstrated that predictive models can only make reliable predictions about a person's risk of violence within very large ranges of likelihood, such as twenty to sixty percent.[7] As a result, virtually everyone's range of likelihood overlaps. When everyone is similar, it becomes impossible to differentiate people with low and high risks of violence. At present, there is no statistical remedy to this challenge.

DATA USED TO BUILD PRETRIAL RISK ASSESSMENTS ARE DISTORTED

Risk assessments are frequently posited as a solution to judges' implicit biases. Yet the data used to build pretrial risk assessments are deeply flawed and racially biased. Pretrial risk assessments rely on historical records of arrests, charges, convictions, and sentences to generate predictions about an individuals propensity for pretrial failure. These tools assume that criminal history data are a reliable and neutral measure of underlying criminal activity, but such records cannot be relied upon for this purpose. Arrest records are both under- and over-inclusive of the true crime rate. Arrest records are

[3] For decades, communities of color have been arrested at higher rates than their white counterparts, even for crimes that these racial groups engage in at comparable rates. As a result, people of color are more likely to be labeled as dangerous than their white counterparts when arrest data is used to measure public safety risk. Thus, they will bear a disproportionate amount of the burden that stems from these harmful conflations between arrest and danger.

[4] Public Safety Assessment, PSA Results (2019).

[5] Julia Angwin et al., *Machine Bias*, Propublica (May 23, 2016), https://www.propublica.org/article/machinebias-risk-assessments-in-criminal-sentencing. These inaccuracies are very much mediated by race – African-Americans were twice as likely to be mislabeled as high risk than their white counterparts.

[6] For example, a recent study found that people significantly overestimate the recidivism rate for individuals who are labeled as "moderate-high" or "high" risk on a risk assessment. Daniel A., Krauss, Gabriel I. Cook & Lukas Klapatch, *Risk Assessment Communication Difficulties: An Empirical Examination of the Effects of Categorical Versus Probabilistic Risk Communication in Sexually Violent Predator Decisions*, Behav. Sci. & L. (2018). (Participants greatly overestimated the true recidivism rate for those assessed as moderate-high risk category – the true rate was less than fifty percent of what participants predicted.)

[7] Stephen D. Hart & David J. Cooke, *Another Look at the (Im–)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments*, 31 Behav. Sci. Law 81, 93 (2013).

under-inclusive because they only chart law enforcement activity, and many crimes do not result in arrest.[8] Less than half of all reported violent crimes result in an arrest, and less than a quarter of reported property crimes result in an arrest. Arrest records are also over-inclusive because people are wrongly arrested and arrested for minor violations, including those that cannot result in jail time. Moreover, decades of research have shown that, for the same conduct, African-American and Latinx people are more likely to be arrested, prosecuted, convicted, and sentenced to harsher punishments than their white counterparts. [9] People of color are treated more harshly than similarly situated white people at each stage of the legal system, which results in serious distortions in the data used to develop risk assessment tools:

- **Arrests**: For decades, communities of color have been arrested at higher rates than their white counterparts, even for crimes that these racial groups engage in at comparable rates.[10] For example, African-Americans are 83% more likely to be arrested for marijuana compared to whites at age 22 and 235% more likely to be arrested at age 27, in spite of similar marijuana usage rates across racial groups.[11] Similarly, African-American drivers are three times as likely as whites to be searched during routine traffic stops, even though police officers generally have a lower "hit rate" for contraband when they search drivers of color.[12] This leads to an overrepresentation of people of color in arrest data. Predictive algorithms that rely on this data overestimate pretrial risk for people of color.

- **Charges**: Empirical research has found that African-American defendants face significantly more severe charges than white defendants, even after controlling for a multitude of factors.[13] Persistent patterns of differential charging make prior charges an unreliable variable for building risk assessments.

- **Convictions & Sentences**: Compared to similarly situated white people, African-Americans are more likely to be convicted[14] and more likely to be sentenced to incarceration.[15]

[8] FBI, 2017 Crime in the United States: Clearances, https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/clearances (last visited June 28, 2019).

[9] See generally The Sentencing Project, Report of the Sentencing Project to the United Nations Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia, and Related Intolerance Regarding Racial Disparities in the United States Criminal Justice System (2018); Lynn Langton & Matthew Durose, U.S. Dep't of Justice, Police Behavior During Traffic and Street Stops, 2011 (2013); Stephen Demuth & Darrell Steffensmeier, *The Impact of Gender and Race-Ethnicity in the Pretrial Release Process*, 51 Soc. Probs. 222 (2004); Jessica Eaglin & Danyelle Solomon, Brennan Center for Justice, Reducing Racial and Ethnic Disparities in Jails: Recommendations for Local Practice (2015); Sonja B. Starr & M. Marit Rehavi, *Racial Disparity in Federal Criminal Sentences*, J. Pol. Econ. 1320 (2014); Marc Mauer, Justice for All? Challenging Racial Disparities in the Criminal Justice System (2010).

[10] Megan Stevenson & Sandra G. Mayson, *The Scale of Misdemeanor Justice*, 98 B.U. L. Rev. 731, 769-770 (2018). This comprehensive national review of misdemeanor arrest data has shown systemic and persistent racial disparities for most misdemeanor offences. The study shows that "black arrest rate is at least twice as high as the white arrest rate for disorderly conduct, drug possession, simple assault, theft, vagrancy, and vandalism." Id. at 759. This study shows that "many misdemeanor offenses criminalize activities that are not universally considered wrongful, and are often symptoms of poverty, mental illness, or addiction." Id. at 766.

[11] "[R]acial disparity in drug arrests between black and whites cannot be explained by race differences in the extent of drug offending, nor the nature of drug offending." Ojmarrh Mitchell & Michael S. Caudy, *Examining Racial Disparities in Drug Arrests*, Just. Q., Jan. 2013, at 22.

[12] Ending Racial Profiling in America: Hearing Before the Subcomm. on the Constitution, Civil Rights and Human Rights of the Comm. on the Judiciary, 112th Cong. 8 (2012) (statement of David A. Harris).

[13] Sonja B. Starr M. Marit Rehavi, *Racial Disparity in Federal Criminal Charging and its Sentencing Consequences*, (U. Mich. L. Econ. Working Paper Series, Working Paper No. 12-002, 2012).

[14] Shamena Anwar, Patrick Bayer & Randi Hjalmarsson, *The Impact of Jury Race in Criminal Trials*, 127 Q. J. Econ. 1017, 1019 (2012).

[15] David S. Abrams, Marianne Bertrand & Sendhil Mullainathan, *Do Judges Vary in Their Treatment of Race*, 41 J. L. Stud. 347, 350 (2012).

Risk assessments that incorporate this distorted data will produce distorted results.[16] There are no technical fixes for these distortions.

CONCLUSION

Pretrial risk assessments do not guarantee or even increase the likelihood of better pretrial outcomes. Risk assessment tools can simply shift or obscure problems with current pretrial practices. Some jurisdictions that have adopted risk assessment tools have seen positive trends in pretrial outcomes, but other jurisdictions have experienced the opposite. Within jurisdictions that have achieved positive outcomes, it is uncertain whether the risk assessment tools were responsible for that success or whether that success is due to other reforms or changes that happened at the same time. Given these mixed outcomes, it is impossible to predict the impact of pretrial risk assessments in any jurisdiction.

Beyond the technical flaws outlined in this statement, a broader and growing body of research questions the validity, ethics, and efficacy of actuarial pretrial risk assessments. For example, most risk assessments are proprietary technology, and defendants assessed by these tools are not allowed the opportunity to inspect and critique the algorithms or their underlying data. Poor implementation and lack of judicial training and buy-in can undermine reforms. Validity and fairness questions arise when tools are trained on data from one jurisdiction but deployed in a jurisdiction with different demographics, judicial culture, and policing practices.

This statement specifically addresses fundamental, technical problems with actuarial risk assessment instruments. These technical problems cannot be resolved. We strongly recommend turning to other reforms.

Chelsea Barabas
Research Scientist
MIT Media Lab

Ruha Benjamin, PhD
Associate Professor
Princeton University

John Bowers
Research Associate
Harvard Law School

Meredith Broussard, PhD
Associate Professor
New York University

Joy Buolamwini
Founder
Algorithmic Justice League

Sasha Constanza-Chock, PhD
Associate Professor
MIT

[16] There have been attempts to solve this problem on the back end by mitigating outcome disparities in risk assessment predictions, but they overlook and do not address the fundamental distortions outlined above.

Kate Crawford, PhD
Distinguished Professor, Co-Founder, Co-Director
AI Now Institute, NYU

Karthik Dinakar, PhD
Research Scientist
MIT Media Lab

Colin Doyle, Staff Attorney
Criminal Justice Policy Program
Harvard Law School

Timnit Gebru, PhD
Co-founder
Black in AI

Bernard E. Harcourt
Professor of Law & Political Science
Columbia University

Stefan Helreich, PhD
Professor & Elting E. Morison Chair
Anthropology, MIT

Brook Hopkins, Executive Director
Criminal Justice Policy Program
Harvard Law School

Joichi Ito, PhD
Director, MIT Media Lab
Visiting Professor of Practice, Harvard Law School

Martha Minow
300th Anniversary University Professor
Harvard University

Cathy O'Neil, PhD
Author
Weapons of Math Destruction

Rodrigo Ochigame
Doctoral Candidate, HASTS
MIT

Heather Paxson, PhD
Professor of Anthropology
MIT

Venerable Tenzin Priyadarshi
Director, Ethics Initiative
MIT Media Lab

Rashida Richardson
Director of Policy Research
AI Now Institute, NYU

Bruce Schneier
Fellow and Lecturer
Harvard Kennedy School

Jason Schultz
Professor
NYU School of Law

Jeffrey Selbin
Clinical Professor of Law
UC Berkeley School of Law

Vincent M. Southerland, Executive Director
Center on Race, Inequality, & the Law
NYU School of Law

Jordi Weinstock
Lecturer on Law
Harvard Law School

Jonathan Zittrain
George Bemis Professor of International Law
Harvard Law School

Ethan Zuckerman
Director, Center for Civic Media
MIT Media Lab