Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System

PARTNERSHIP ON AI

This report was written by the staff of the Partnership on AI (PAI) and many of our Partner organizations, with particularly extensive input from the members of PAI's Fairness, Transparency, and Accountability Working Group. Our work on this topic was initially prompted by California's Senate Bill 10 (S.B. 10), which would mandate the purchase and use of statistical and machine learning risk assessment tools for pretrial detention decisions, but our work has subsequently expanded to assess the use of such software across the United States.

Though this document incorporated suggestions or direct authorship from around 30 to 40 of our partner organizations, it should not under any circumstances be read as representing the views of any specific member of the Partnership. Instead, it is an attempt to report the widely held views of the artificial intelligence research community as a whole.

The Partnership on AI is a 501(c)3 nonprofit organization established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society.

The Partnership's activities are determined in collaboration with its coalition of over 80 members, including civil society groups, corporate developers and users of AI, and numerous academic artificial intelligence research labs. PAI aims to create a space for open conversation, the development of best practices, and coordination of technical research to ensure that AI is used for the benefit of humanity and society. Crucially, the Partnership is an independent organization; though supported and shaped by our Partner community, the Partnership is ultimately more than the sum of its parts and makes independent determinations to which its Partners collectively contribute, but never individually dictate. PAI provides administrative and project management support to Working Groups, oversees project selection, and provides financial resources or direct research support to projects as needs dictate.

The Partnership on AI is deeply grateful for the collaboration of so many colleagues in this endeavor and looks forward to further convening and undertaking the multi-stakeholder research needed to build best practices for the use of AI in this critical domain.

Executive Summary

This report documents the serious shortcomings of risk assessment tools in the U.S. criminal justice system, most particularly in the context of pretrial detentions, though many of our observations also apply to their uses for other purposes such as probation and sentencing. Several jurisdictions have already passed legislation mandating the use of these tools, despite numerous deeply concerning problems and limitations. Gathering the views of the artificial intelligence and machine learning research community, PAI has outlined ten largely unfulfilled requirements that jurisdictions should weigh heavily and address before further use of risk assessment tools in the criminal justice system.

Using risk assessment tools to make fair decisions about human liberty would require solving deep ethical, technical, and statistical challenges, including ensuring that the tools are designed and built to mitigate bias at both the model and data layers, and that proper protocols are in place to promote transparency and accountability. The tools currently available and under consideration for widespread use suffer from several of these failures, as outlined within this document.

We identified these shortcomings through consultations with our expert members, as well as reviewing the literature on risk assessment tools and publicly available resources regarding tools currently in use. Our research was limited in some cases by the fact that most tools do not provide sufficiently detailed information about their current usage to evaluate them on all of the requirements in this report. Jurisdictions and companies developing these tools should implement Requirement 8, which calls for greater transparency around the data and algorithms used, to address this issue for future research projects. That said, many of the concerns outlined in this report apply to any attempt to use existing criminal justice data to train statistical models or to create heuristics to make decisions about the liberty of individuals.

Challenges in using these tools effectively fall broadly into three categories, each of which corresponds to a section of our report:

1. Concerns about the validity, accuracy, and bias in the tools themselves;

2. Issues with the interface between the tools and the humans who interact with them; and

3. Questions of governance, transparency, and accountability.

Although the use of these tools is in part motivated by the desire to mitigate existing human fallibility in the criminal justice system, it is a serious misunderstanding to view tools as objective or neutral simply because they are based on data. While formulas and statistical models provide some degree of consistency and replicability, they still share or amplify many weaknesses of human decision-making. Decisions regarding what data to use, how to handle missing data, what objectives to optimize, and what thresholds to set all have significant implications on the accuracy, validity, and bias of these tools, and ultimately on the lives and liberty of the individuals they assess.

In addition to technical concerns, there are human-computer interface issues to consider with the implementation of such tools. Human-computer interface in this case refers to how humans collect and feed information into the tools and how humans interpret and evaluate the information that the tools generate. These tools must be held to high standards of interpretability and explainability to ensure that users (including judges, lawyers, and clerks, among others) can understand how the tools' predictions are reached and make reasonable decisions based on these predictions. To improve interpretability, such predictions should explicitly include information such as error bands to express the uncertainty behind their predictions. In addition, users must attend trainings that teach how and when to use these tools appropriately, and how to understand the uncertainty of their results.

Moreover, to the extent that such systems are adopted to make life-changing decisions, tools and those who operate them must meet high standards of transparency and accountability. The data used to train the tools and the tools themselves must be subject to independent review by third-party researchers, advocates, and other relevant stakeholders. The tools also must receive ongoing evaluation, monitoring, and audits to ensure that they are performing as expected, and aligned with well-founded policy objectives.

In light of these issues, as a general principle, these tools should not be used alone to make decisions to detain or to continue detention. Given the pressing issue of mass incarceration, it might be reasonable to use these tools to facilitate the automatic pretrial release of more individuals, but they should not be used to detain individuals automatically without additional (and timely) individualized hearings. Moreover, any use of these tools should address the bias, human-computer interface, transparency, and accountability concerns outlined in this report.

This report highlights some of the key problems with risk assessment tools for criminal justice applications. Many important questions remain open, however, and unknown issues may yet emerge in this space. Surfacing and answering those concerns will require ongoing research and collaboration between policymakers, the AI research community, and civil society groups. It is PAI's mission to spur and facilitate these conversations and to produce research to bridge these gaps.

Contents

Introduction

Context

Risk assessment instruments are statistical models used to predict the probability of a particular future outcome. Such predictions are accomplished by measuring the relationship between an individual's features (for example, their demographic information, criminal history, or answers to a psychometric questionnaire) and combining numerical representations of those features into a risk score. Scoring systems are generally created using statistical techniques and heuristics applied to data to consider how each feature contributes to prediction of a particular outcome (e.g., failure to appear at court). These scores are often then used to assign individuals to different brackets of risk.[1]

Though they are usually much simpler than the deep neural networks used in many modern artificial intelligence systems, criminal justice risk assessment tools are basic forms of AI.[2] Some use heuristic frameworks to produce their scores, though most use simple machine learning methods to train predictive models from input datasets. As such, they present a paradigmatic example of the potential social and ethical consequences of automated AI decision-making.

The use of risk assessment tools in the criminal justice system is expanding rapidly, and policymakers at both the federal and state level have passed legislation to mandate their use.[3] This has largely occurred as part of a reform effort that is grappling with extremely high incarceration rates in the United States, which are disproportionate to crime rates and to international and historical baselines (see Figures 1-3). Proponents of these tools have advocated for their potential to streamline inefficiencies, reduce costs, and provide rigor and reproducibility for life-critical decisions. Some advocates hope that these changes will mean a reduction in unnecessary detention and provide fairer and less punitive decisions than the cash bail system or systems where human decision-makers like judges have complete discretion.

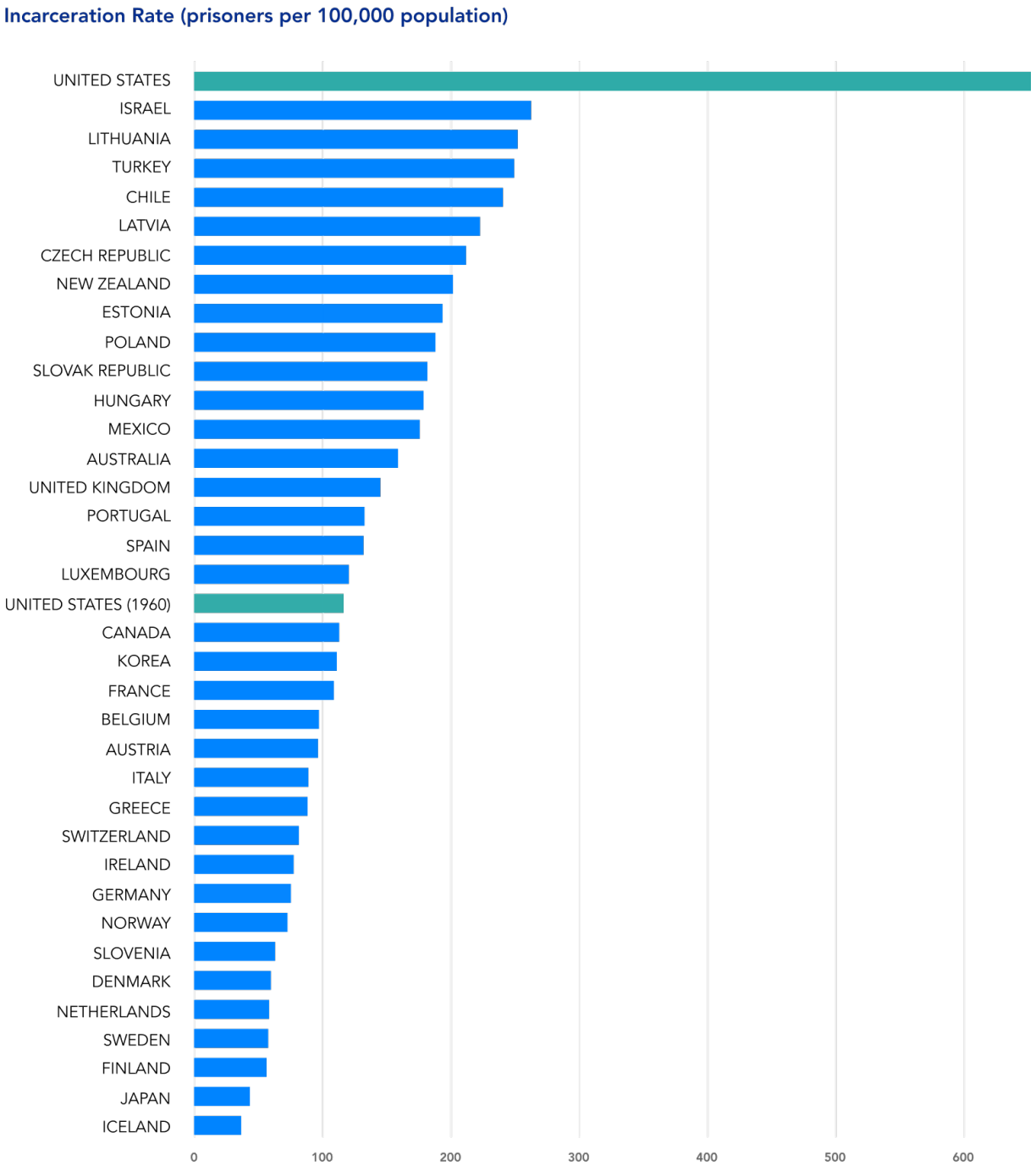
---

[1] For example, many risk assessment tools assign individuals to *decile ranks*, converting their risk score into a rating from 1-10 which reflects whether they're in the bottom 10% of risky individuals (1), the next highest 10% (2), and so on (3-10). Alternatively, risk categorization could be based on thresholds labeled as "low," "medium," or "high" risk.

[2] Whether this is the case depends on how one defines AI; it would be true under many but not all of the definitions surveyed for instance in Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2010, at 2. PAI considers more expansive definitions, that include any automation of analysis and decision making by humans, to be most helpful.

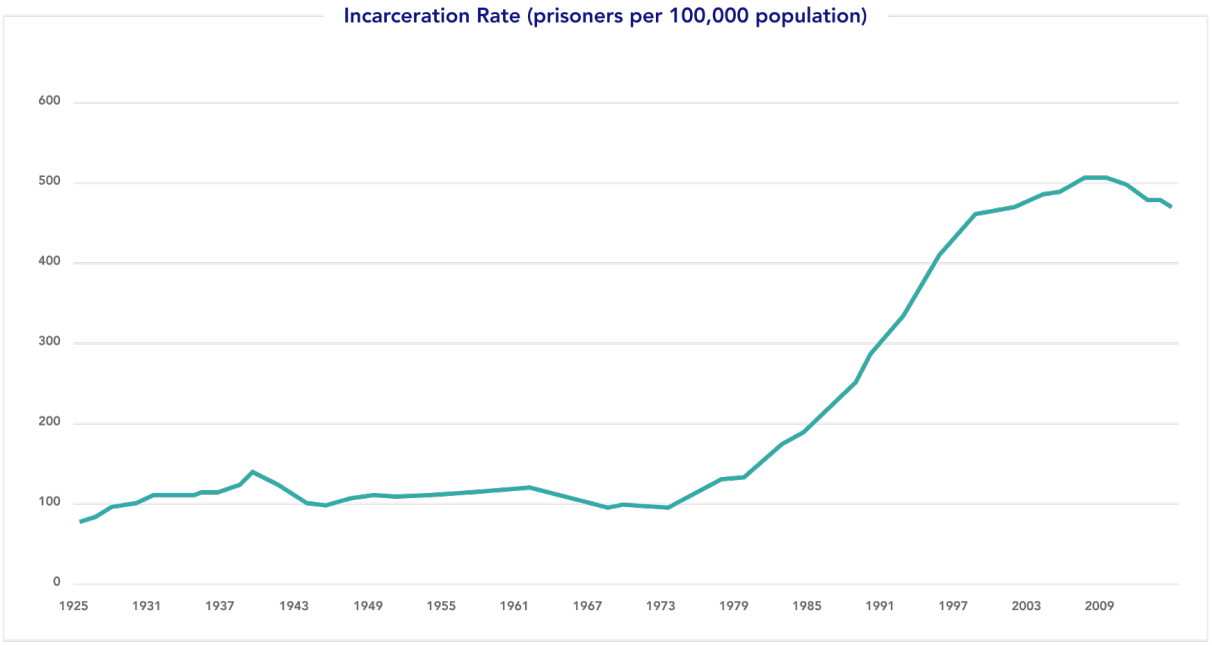
[3] In California, the recently enacted California Bail Reform Act (S.B. 10) mandates the implementation of risk assessment tools while eliminating money bail in the state, though implementation of the law has been put on hold as a result of a 2020 ballot measure; see https://ballotpedia.org/California_Replace_Cash_Bail_with_Risk_Assessments_Referendum_(2020); Robert Salonga, *Law ending cash bail in California halted after referendum qualifies for 2020 ballot*, San Jose Mercury News (Jan. 17, 2019), https://www.mercurynews.com/2019/01/17/law-ending-cash-bail-in-california-halted-after-referendum-qualifies-for-2020-ballot/. In addition, a new federal law, the First Step Act of 2018 (S. 3649), requires the Attorney General to review existing risk assessment tools and develop recommendations for "evidence-based recidivism reduction programs" and to "develop and release" a new risk- and needs- assessment system by July 2019 for use in managing the federal prison population. The bill allows the Attorney General to use currently-existing risk and needs assessment tools, as appropriate, in the development of this system.

Figure 1: Incarceration in the U.S. Relative to OECD and Historical Baselines



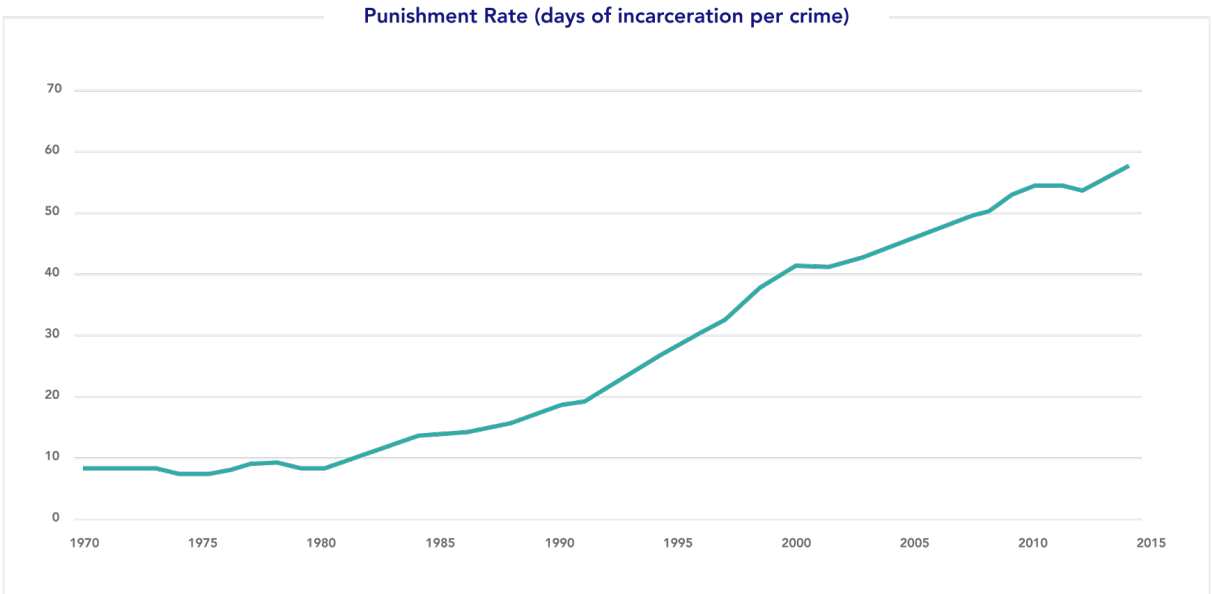
Source: Bureau of Justice Statistics, World Prison Brief—Birkbeck, University of London (2015/2016 data)
Note: U.S. 1960 figure includes those in state or federal institutions only

Figure 2: U.S. State and Federal Incarceration Rates (1925-2014)



Source: Bureau of Justice Statistics & Wikimedia

Figure 3: U.S. State and Federal Incarceration Relative to All Reported Crimes (1970-2014)



Sources: National Archive of Criminal Justice Data, Bureau of Justice Statistics & Wikimedia
Note: Punishment rate is calculated based on the number of people incarcerated per year rather than convictions in a given year

These are critically important public policy goals, but there is reason to believe that these views may be too optimistic. There remain serious and unresolved problems with accuracy, validity, and bias in both the datasets and statistical models that drive these tools. Moreover, these tools are also often built to answer the wrong questions, used in poorly conceived settings, or are not subject to sufficient review, auditing, and scrutiny. These concerns are nearly universal in the AI research community and across our Partnership, though views differ on whether they could realistically be solved by improvements to the tools.

Scope of this report

This Report of the Partnership on AI was written to gather, synthesize, and document the views of the artificial intelligence research community on the use of risk assessment tools in the U.S. criminal justice system. This report focuses on the use of these tools in the pretrial context, but many of the concerns identified with these tools are applicable across other risk assessment contexts (e.g., consideration of parole release and sentencing within the U.S.; design of risk assessment systems generally in other countries). The report attempts to answer: What technical and human-computer interface challenges prevent risk assessment tools from being used to inform fair decisions? And with what transparency, auditing, and procedural protections would it be acceptable to use these tools as possible inputs into criminal justice determinations?

Background on PAI

The Partnership on AI is a 501(c)3 non-profit organization that convenes a coalition of over 80 members, including civil society groups, corporate developers and users of AI, and numerous academic artificial intelligence research labs, to answer important questions about artificial intelligence policy and ethics. This particular report reflects input from conversations that PAI has convened with dozens of its member organizations, as well as numerous experts on fairness and bias in machine learning and the U.S. criminal justice system. Though the report should not be taken as stating an official stance of any particular member, it attempts to report views widely held across our membership and the artificial intelligence research community.

Baselines for Comparison

Some of the controversy about risk assessment tools derives from different baselines against which risk assessment tools are evaluated. Policymakers have many possible baselines they can use in deciding whether to procure and use these tools, including:

A. **Do risk assessment tools achieve absolute fairness?** This is unlikely to be achieved by any system or institution due to serious limitations in data and also unresolved philosophical questions about fairness.

B. **Are risk assessment tools as fair as they can possibly be based on available datasets?** This may be achievable, but only in the context of (a) deciding on a specific measure of fairness and (b) using the best available methods to mitigate societal and statistical biases in the data. In practice, however, given the limitations in available data, this often translates to ignoring biases in the data that are difficult to address.

C. **Are risk assessment tools an improvement over current processes and human decision-makers?** Risk assessment tools can be benchmarked against the performance of the processes, institutions, and human decision-making practices in place before their introduction, or similar systems in other jurisdictions without risk assessment tools. Such evaluations could be based on measurable goals (like better predicting appearance for court dates or recidivism) or lack of susceptibility to human biases. In this sense, risk assessment tools may not achieve a defined notion of fairness, but rather be comparatively better than the status quo.

PAI's work on risk assessment tools in the criminal justice system was initially prompted by the passage of Senate Bill 10 (S.B. 10) in California, which would use risk assessment tools in making pretrial detention decisions. The scope of this project has since expanded, with this report addressing not only the S.B. 10 context but also the concerns more broadly with the use of risk assessment tools around the country.

Objectives of the report

An overwhelming majority of the Partnership's consulted experts agreed that current risk assessment tools are not ready for use in helping to make decisions to detain or continue to detain criminal defendants without the use of an individualized hearing.[4] One objective of this report is to articulate the reasons for this nearly unanimous view of contributors and to help inform a dialogue with policymakers considering the use of these tools. PAI members and the wider AI community do not, however, have consensus on whether statistical risk assessment tools could ever be improved to justly detain or continue to detain someone on the basis of their risk assessment score without an individualized hearing. For some of our members, the concerns remain structural and procedural as well as technical.[5] Regardless of the differing views on these particular issues, this report summarizes the technical, human-computer interface, and governance problems that the community has collectively identified.

Baselines for Comparison (continued)

D. **Are risk assessment tools an improvement over other possible reforms to the criminal justice system?** Other reforms may address the same objectives (e.g., improving public safety, reducing the harm of detention, and reducing the costs and burdens of judicial process) at lower cost, greater ease of implementation, or without trading off civil rights concerns.

Baselines A and B are useful for fundamental research on algorithmic fairness and for empirical analysis of the performance of existing systems, but they necessarily produce ambiguous results due to the existence of highly defensible but incompatible definitions of fairness. Nonetheless, they can provide a useful framework for understanding the philosophical, legal, and technical issues with proposed tools.

Baseline C is one of the widely held perspectives by experts operating in the space. It is potentially appropriate for policymakers and jurisdictions purchasing tools under legislative mandates beyond their control, or in situations where political constraints mean that Baseline D is inapplicable. We should, however, stress that in all of the conversations convened by the Partnership on AI, Baseline D has been widely viewed as more fundamentally correct and appropriate as both a policymaking goal and an evaluation standard for risk assessment tools. Therefore, legislatures and judicial authorities should apply Baseline D whenever it is feasible for them to do so.

[4] In addition, many of our civil society partners have taken a clear public stance to this effect, and some go further in suggesting that only individual-level decision-making will be adequate for this application regardless of the robustness and validity of risk assessment instruments. See *The Use of Pretrial 'Risk Assessment' Instruments: A Shared Statement of Civil Rights Concerns*, http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf (shared statement of 115 civil rights and technology policy organizations, arguing that all pretrial detention should follow from evidentiary hearings rather than machine learning determinations, on both procedural and accuracy grounds); see also Comments of Upturn; The Leadership Conference on Civil and Human Rights; The Leadership Conference Education Fund; NYU Law's Center on Race, Inequality, and the Law; The AI Now Institute; Color Of Change; and Media Mobilizing Project on Proposed California Rules of Court 4.10 and 4.40, https://www.upturn.org/static/files/2018-12-14_Final-Coalition-Comment-on-SB10-Proposed-Rules.pdf ("Finding that the defendant shares characteristics with a collectively higher risk group is the most specific observation that risk assessment instruments can make about any person. Such a finding does not answer, or even address, the question of whether detention is the only way to reasonably assure that person's reappearance or the preservation of public safety. That question must be asked specifically about the individual whose liberty is at stake—and it must be answered in the affirmative in order for detention to be constitutionally justifiable.") PAI notes that the requirement for an individualized hearing before detention implicitly includes a need for timeliness. Many jurisdictions across the US have detention limits at 24 or 48 hours without hearings.
Aspects of this stance are shared by some risk assessment tool makers; see, Arnold Ventures' *Statement of Principles on Pretrial Justice and Use of Pretrial Risk Assessment*, https://craftmediabucket.s3.amazonaws.com/uploads/AV-Statement-of-Principles-on-Pretrial-Justice.pdf.

[5] See Ecological Fallacy section and Baseline D for further discussion of this topic.

Minimum Requirements for the Responsible Deployment of Criminal Justice Risk Assessment Tools

Accuracy, Validity, and Bias

What is Accuracy?

Accuracy represents the model's performance compared to an accepted baseline or predefined correct answer based on the dataset available.[6] Most commonly, some of the data used to create the model will be reserved for testing and model tuning. These reserved data provide for fresh assessments that help toolmakers avoid "overfitting"[7] during the process of experimentation.

Measuring accuracy involves assessing whether the model did the best possible job of prediction on the test data. To say that a model predicts inaccurately is to say that it is giving the wrong answer according to the data, either in a particular case or across many cases.

Since accuracy is focused narrowly on how the tool performs on data reserved from the original data set, it does not address issues that might undermine the reasonableness of the dataset itself (discussed in the section on validity). Indeed, because accuracy is calculated with respect to an accepted baseline of correctness, accuracy fails to account for whether the data used to test or validate the model are uncertain or contested. Such issues are generally taken into account under an analysis of validity. Although accuracy is often the focus of toolmakers when evaluating the performance of their models, validity and bias are often the more relevant concerns in the context of using such tools in the criminal justice system.

Fundamental Issues with Using Group-Level Data to Judge Individuals

A fundamental philosophical and legal question is whether it is acceptable to make determinations about individuals' liberty based on data about others in their group. In technical communities, making predictions about individuals from group-level data is known as the ecological fallacy. Although risk assessment tools use data about an individual as inputs, the relationship between these inputs and the predicted outcome is determined by patterns in training data about other people's behavior.

In the context of sentencing, defendants have a constitutional right to have their sentence determined based on what they did themselves instead of what others with similarities to them have done. This concern arose in Wisconsin v. Loomis, where the court prohibited the use of risk scores as the decisive factor in liberty decisions, noting that "offender who is young, unemployed, has an early age-at-first-arrest and a history of supervision failure, will score medium or high on the Violence Risk Scale even though the offender never had a violent offense," illustrating how the predictions of these tools do not necessarily map onto individual cases.

---

[6] Quantitatively, accuracy is usually defined as the fraction of correct answers the model produces among all the answers it gives. So a model that answers correctly in 4 out of 5 cases would have an accuracy of 80%. Interestingly, models which predict rare phenomena (like violent criminality) can be incredibly accurate without being useful for their prediction tasks. For example, if only 1% of individuals will commit a violent crime, a model that predicts that no one will commit a violent crime will have 99% accuracy even though it does not correctly identify any of the cases where someone actually commits a violent crime. For this reason and others, evaluation of machine learning models is a complicated and subtle topic which is the subject of active research. In particular, note that inaccuracy can and should be subdivided into errors of "Type I" (false positive) and "Type II" (false negative) - one of which may be more acceptable than the other, depending on the context.

[7] Overfitting is a statistical problem that is analogous to learning the answer to all the questions on an exam by heart, without having actually understood the true principles that made them correct. When a model is said to have overfitted, this means that it has limited ability to generalize its evaluation to new data, and thus limited application for the complex and varied real-world.

What is Validity?

A narrow focus on accuracy can blind decision-makers to important real-world considerations related to the use of prediction tools. With any statistical model, and especially one used in as critical a context as criminal justice risk assessments, it is important to establish the model's validity, or fidelity to the real world. That is, if risk assessments purport to measure how likely an individual is to fail to appear or to be the subject of a future arrest, then it should be the case that the scores produced in fact reflect the relevant likelihoods. Unlike accuracy, validity takes into consideration the broader context around how the data was collected and what kind of inference is being drawn. A tool might not be valid because the data that was used to develop it does not properly reflect what is happening in the real world (due to measurement error, sampling error, improper proxy variables, failure to calibrate probabilities,[8] or other issues).

Separate from data and statistical challenges, a tool might also not be valid because the tool does not actually answer the correct question. Because validation is always with respect to a particular context of use and a particular task to which a system is being put, validating a tool in one context says little about whether that tool is valid in another context. For example, a risk assessment might predict future arrests quite well when applied to individuals in a pretrial context, but quite poorly when applied to individuals post-conviction, or it might predict future arrest well in one jurisdiction, but not another.[9] Similarly, different models built based on the same data, created with different modeling decisions and assumptions, may have different levels of validity. Thus, different kinds of predictions (e.g., failure to appear, flight, recidivism, violent recidivism) in different contexts require separate validation. Without such validation, even well-established methods can produce flawed predictions. In other words, just because a tool uses data collected from the real world does not automatically make its findings valid.

Fundamental Issues with Using Group-Level Data to Judge Individuals (continued)

The ecological fallacy is especially problematic in the criminal justice system given the societal biases that are reflected in criminal justice data, as described in the sections on Requirements 1 and 2. It is thus likely that decisions made by risk assessment tools are driven in part by what protected class an individual may belong to, raising significant Equal Protection Clause concerns.

While there is a statistical literature on how to deal with technical issues resulting from the ecological fallacy, the fundamental philosophical question of whether it is permissible to detain individuals based on data about others in their group remains. As more courts grapple with whether to use risk assessment tools, this question should be at the forefront of debate and discussed as a first-order principle.

[8] Calibration is a property of models such that among the group they predict a 50% risk for, 50% of cases recidivate. Note that this says nothing about the accuracy of the prediction, because a coin toss would be calibrated in that sense. All risk assessment tools should be calibrated, butthere are more specific desirable properties such as calibration within groups (discussed in Requirement 2 below) that not all tools will or should satisfy completely.

[9] Sarah L. Desmarais, Evan M. Lowder, Pretrial Risk Assessment Tools: A Primer for Judges, Prosecutors, and Defense Attorneys, MacArthur Safety and Justice Challenge (Feb 2019). The issue of cross-comparison applies not only to geography but to time. It may be valuable to use comparisons over time to assist in measuring the validity of tools, though such evaluations must be corrected for the fact that crime in the United States is presently a rapidly changing (and still on the whole rapidly declining) phenomenon.

What is Bias?

In statistical prediction settings, "bias" has several overlapping meanings. The simplest meaning is that a prediction made by a model errs in a systematic direction—for instance, it predicts a value that is too low on average, or too high on average for the general population. In the machine learning fairness literature, however, the term bias is used to refer to situations where the predicted probabilities are systematically either too high or too low for *specific subpopulations*.[10] These subpopulations may be defined by protected class variables (race, gender, age, etc.) or other variables of concern, like socioeconomic class. In this paper, we will primarily use the term "bias" in this narrower sense, which aligns with the everyday use of the term referring to disparate judgments about different groups of people.[11]

Bias in risk assessment tools can come from many sources.[12] Requirement 1 below discusses data bias that is caused by imperfect data quality, missing data, and sampling bias. Requirement 2 discusses model bias that stems from omitted variable bias and proxy variables. Requirement 3 discusses model bias that results from the use of composite scores that conflate multiple distinct predictions. In combination with concerns about accuracy and validity, these challenges present significant concern for the use of risk assessment tools in criminal justice domains.

[10] As a technical matter, a model can be biased for subpopulations while being unbiased on average for the population as a whole.

[11] Note here that the phenomenon of societal bias—the existence of beliefs, expectations, institutions, or even self-propagating patterns of behavior that lead to unjust outcomes for some groups—is not always the same as, or reflected in statistical bias, and vice versa. One can instead think of these as an overlapping Venn diagram with a large intersection. Most of the concerns about risk assessment tools are about biases that are simultaneously statistical and societal, though there are some that are about purely societal bias. For instance, if non-uniform access to transportation (which is a societal bias) causes higher rates of failure to appear for court dates in some communities, the problem is a societal bias, but not a statistical one. The inclusion of demographic parity measurements as part of model bias measurement (see Requirement 2) may be a way to measure this, though really the best solutions involve distinct policy responses (for instance, providing transportation assistance for court dates or finding ways to improve transit to underserved communities).

[12] For instance, Eckhouse *et al.* propose a 3-level taxonomy of biases. Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini, *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, Criminal Justice and Behavior, (Nov 2018).

Requirement 1: Training datasets must measure the intended variables

Datasets pose profound and unresolved challenges to the validity of statistical risk assessments. In almost all cases, errors and bias in measurement and sampling prevent readily available criminal justice datasets from reflecting what they were intended to measure. Building valid risk assessment tools would require (a) a methodology to reweight and debias training data using second sources of truth, and (b) a way to tell whether that process was valid and successful. To our knowledge, no risk assessment tools are presently built with such methods.[13]

Statistical validation of recidivism prediction in particular suffers from a fundamental problem: the ground truth of whether an individual committed a crime is generally unavailable, and can only be estimated via imperfect proxies such as crime reports or arrests. Since the target for prediction (having actually committed a crime) is unavailable, it is tempting to change the goal of the tool to predicting arrest, rather than crime. If the goal, however, of using these tools is to predict a defendant's risk to public safety—as most risk assessment tools are—the objective must be whether a defendant is likely to commit an offense that justifies pretrial detention, not whether the defendant is likely to be arrested for or convicted of any offense in the future.[14]

One problem with using such imperfect proxies is that different demographic groups are stopped, searched, arrested, charged, and are wrongfully convicted at very different rates in the current US criminal justice system.[15] Further, different types of crimes are reported and recorded at different rates, and the rate of reporting may depend on the demographics of the perpetrator and victim.[16] For example, it is likely that all (or very nearly all) bank robberies are reported to police.[17] On the other hand, marijuana possession arrests are notoriously biased, with black Americans much more likely to be arrested than whites, despite similar use rates.[18]

---

[13] Some of the experts within the Partnership oppose the use of risk assessment tools specifically because of their pessimism that sufficient data exists or could practically be collected to meet purposes (a) and (b).

[14] Moreover, defining recidivism is difficult in the pretrial context. Usually, recidivism variables are defined using a set time period, e.g., whether someone is arrested within 1 year of their initial arrest or whether someone is arrested within 3 years of their release from prison. In the pretrial context, recidivism is defined as whether the individual is arrested during the time after their arrest (or pretrial detention) and before the individual's trial. That period of time, however, can vary significantly from case to case, so it is necessary to ensure that each risk assessment tool predicts an appropriately defined measure of recidivism or public safety risk.

[15] *See, e.g., Report: The War on Marijuana in Black and White*, ACLU (2013), https://www.aclu.org/report/report-war-marijuana-black-and-white; ACLU submission to Inter-American Commission on Human Rights, Hearing on Reports of Racism in the Justice System of the United States, https://www.aclu.org/sites/default/files/assets/141027_iachr_racial_disparities_aclu_submission_0.pdf, (Oct 2017); Samuel Gross, Maurice Possley, Klara Stephens, *Race and Wrongful Convictions in the United States*, National Registry of Exonerations, https://www.law.umich.edu/special/exoneration/Documents/Race_and_Wrongful_Convictions.pdf; but see Jennifer L. Skeem and Christopher Lowenkamp, *Risk, Race & Recidivism: Predictive Bias and Disparate Impact*, Criminology 54 (2016), 690, https://risk-resilience.berkeley.edu/sites/default/files/journal-articles/files/criminology_proofs_archive.pdf (For some categories of crime in some jurisdictions, victimization and self-reporting surveys imply crime rates are comparable to arrest rates across demographic groups; an explicit and transparent reweighting process is procedurally appropriate even in cases where the correction it results in is small).

[16] *See* David Robinson and John Logan Koepke, *Stuck in a Pattern: Early evidence on 'predictive policing' and civil rights*, (Aug. 2016). https://www.upturn.org/reports/2016/stuck-in-a-pattern/ ("Criminologists have long emphasized that crime reports, and other statistics gathered by the police, are not an accurate record of the crime that happens in a community. In short, the numbers are greatly influenced by what crimes citizens choose to report, the places police are sent on patrol, and how police decide to respond to the situations they encounter. The National Crime Victimization Survey (conducted by the Department of Justice) found that from 2006-2010, 52 percent of violent crime victimizations went unreported to police and 60 percent of household property crime victimizations went unreported. Historically, the National Crime Victimization Survey 'has shown that police are not notified of about half of all rapes, robberies and aggravated assaults.'") See also Kristian Lum and William Isaac, *To predict and serve?* (2016): 14-19.

[17] Carl B. Klockars, *Some Really Cheap Ways of Measuring What Really Matters*, in Measuring What Matters: Proceedings From the Policing Research Meetings, 195, 195-201 (1999), https://www.ncjrs.gov/pdffiles1/nij/170610.pdf. [https://perma.cc/BRP3-6Z79] ("If I had to select a single type of crime for which its true level—the level at which it is reported—and the police statistics that record it were virtually identical, it would be bank robbery. Those figures are likely to be identical because banks are geared in all sorts of ways...to aid in the reporting and recording of robberies and the identification of robbers. And, because mostly everyone takes bank robbery seriously, both Federal and local police are highly motivated to record such events.")

[18] ACLU, *The War on Marijuana in Black and White: Billions of Dollars Wasted on Racially Biased Arrests*, (2013), available at https://www.aclu.org/files/assets/aclu-thewaronmarijuana-rel2.pdf.

Thus, "arrest, conviction, and incarceration data are most appropriately viewed as measures of official response to criminal behavior," impacting certain groups disproportionately.[19]

Estimating such biases can be difficult, although in some cases may be possible by using secondary sources of data collected separately from law enforcement or government agencies.[20] For example, arrest or conviction data could be reweighted using the National Crime Victimization Survey, which provides a second method of estimating the demographic characteristics for types of crimes where there is a victim who is able to see the perpetrator, or surveys that collect self-reported data about crime perpetration and arrest such as the National Longitudinal Surveys of Youth. Performing such reweighting would be a subtle statistical task that could easily be performed incorrectly, and so a second essential ingredient would be developing a method accepted by the machine learning and statistical research communities for determining whether data reweighting had produced valid results that accurately reflect the world.

Beyond the difficulty in measuring certain outcomes, data is also needed to properly distinguish between different causes of the same outcome. For instance, just looking at an outcome of failure to appear in court obscures the fact that there are many different possible reasons for such an outcome. Given that there are legitimate reasons for failing to appear for court that would not suggest that the individuals pose a danger to society (e.g., a family emergency or limited transportation options),[21] grouping together all individuals who fail to appear for court would unfairly increase the probability that individuals that tend to have more legitimate reasons for failing to appear in court (e.g., people with dependants or who have limited transportation options) would be unfairly detained. Thus, if the goal of a risk assessment tool is to make predictions about whether or not a defendant will flee justice, data would need to be collected that distinguish between individuals that intentionally versus unintentionally fail to appear for court dates.[22]

Given that validity often depends on local context to ensure a tool's utility, where possible, the data discussed above should be collected on a jurisdiction-by-jurisdiction basis in order to capture significant differences in geography, transportation, and local procedure that affect those outcomes.

---

[19] Delbert S. Elliott, Lies, *Damn Lies, and Arrest Statistics*, (1995), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.9427&rep=rep1&type=pdf, 11.

[20] Lisa Stoltenberg & Stewart J. D'Alessio, *Sex Differences in the Likelihood of Arrest*, J. Crim. Justice 32 (5), 2004, 443-454; Lisa Stoltenberg, David Eitle & Stewart J. D'Alessio, *Race and the Probability of Arrest*, Social Forces 81(4) 2003 1381-1387; Tia Stevens & Merry Morash, *Racial/Ethnic Disparities in Boys' Probability of Arrest and Court Actions in 1980 and 2000: The Disproportionate Impact of "Getting Tough" on Crime*, Youth and Juvenile Justice 13(1), (2014).

[21] Simply reminding people to appear improves appearance rates. Pretrial Justice Center for Courts, *Use of Court Date Reminder Notices to Improve Court Appearance Rates*, (Sept. 2017).

[22] There are a number of obstacles that risk assessment toolmakers have identified towards better predictions on this front. Firstly, there is a lack of consistent data and definitions to help disentangle willful flight from justice from failures to appear for reasons that are either unintentional or not indicative of public safety risk. Policymakers may need to take the lead in defining and collecting data on these reasons, as well as identifying interventions besides incarceration that may be most appropriate for responding to them.

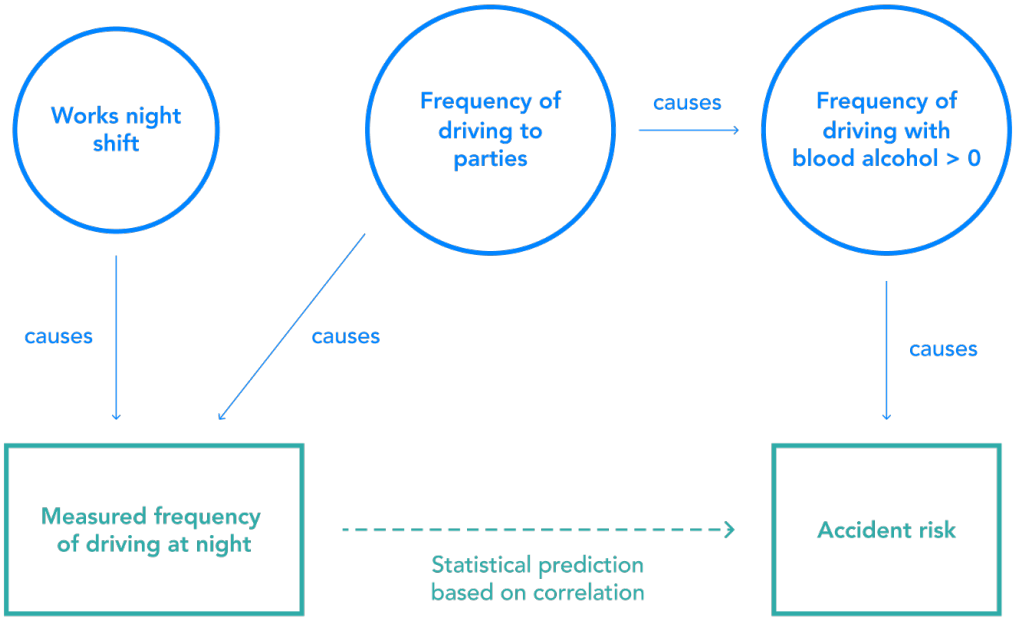
Requirement 2: Bias in statistical models must be measured and mitigated

There are two widely held misconceptions about bias in statistical prediction systems. The first is that models will only reflect bias if the data they were trained with was itself inaccurate or incomplete. A second is that predictions can be made unbiased by avoiding the use of variables indicating race, gender, or other protected classes.[23] Both of these intuitions are incorrect at the technical level.

It is perhaps counterintuitive, but in complex settings like criminal justice, virtually all statistical predictions will be biased even if the data was accurate, and even if variables such as race are excluded, unless specific steps are taken to measure and mitigate bias. The reason is a problem known as omitted variable bias. Omitted variable bias occurs whenever a model is trained from data that does not include all of the relevant causal factors. Missing causes of the outcome variable that also cause the input variable of interest are known as confounding variables. Moreover, the included variables can be proxies for protected variables like race.[24]

Figure 4 illustrates an example of this problem:

Figure 4: Omitted Variable Bias in a Simple Insurance Model



Solid lines indicate causal variables, boxes indicate variables that are measured in the training dataset, ellipses indicate variables that were not measured, and the dotted line indicates the prediction that is made by the final trained model.

[23] This is known in the algorithmic fairness literature as "fairness through unawareness"; see *Moritz Hardt, Eric Price, & Nathan Srebro, Equality of Opportunity in Supervised Learning, Proc. NeurIPS 2016*, https://arxiv.org/pdf/1610.02413.pdf, first publishing the term and citing earlier literature for proofs of its ineffectiveness, particularly Pedreshi, Ruggieri, & Turini, *Discrimination-aware data mining, Knowledge Discovery & Data Mining*, Proc. SIGKDD (2008), http://eprints.adm.unipi.it/2192/1/TR-07-19.pdf.gz. In other fields, blindness is the more common term for the idea of achieving fairness by ignoring protected class variables (e.g., "race-blind admissions" or "gender-blind hiring").

[24] Another way of conceiving omitted variable bias is as follows: data-related biases as discussed in Requirement 1 are problems with the rows in a database or spreadsheet: the rows may contain asymmetrical errors, or not be a representative sample of events as they occur in the world. Omitted variable bias, in contrast, is a problem with not having enough or the right columns in a dataset.

Frequently driving to parties is a confounding variable because it causes both night-time driving and accident risk. A model trained on data about the times of day that drivers drive would exhibit bias against people who work night shifts, because it would conflate the risk of driving to parties with the risk of driving at night.

The diagram also indicates proxy variables at work: frequency of driving at night is a proxy, via driving to parties, for driving while inebriated. It is also a direct proxy for working night shifts. As a result, even though it is not appropriate to charge someone higher insurance premiums simply because they work night shifts, that is the result in this case due to the inclusion of the proxy variable of frequency of driving at night.

Similar networks of proxies apply to criminal risk assessments, from observed input variables such as survey questions asking "How many of your friends/acquaintances have ever been arrested?" or "In your neighborhood, have some of your friends or family been crime victims?"[25] that are proxies for race. As such, it is difficult to separate the use of risk assessment instruments from the use of constitutionally-protected factors such as race to make predictions, and mitigations for this model-level bias are needed.

Methods to Mitigate Bias

There are numerous possible statistical methods that attempt to correct for bias in risk assessment tools. The correct method to employ will depend on what it means for a tool to be "fair" in a particular application, so this is not only a technical question but also a question of law, policy, and ethics. Although there is not a one-size-fits-all solution to addressing bias, below are some of the possible approaches that might be appropriate in the context of US risk assessment predictions:[26]

1. One approach would be to design the model to satisfy a requirement of "equal opportunity," meaning that false positive rates (FPRs) are balanced across some set of protected classes (in the recidivism context, the FPR would be the probability that someone who does not recidivate is incorrectly predicted to recidivate).[27] Unequal false positive rates are especially problematic in the criminal justice system since they imply that the individuals who do not recidivate in one demographic group are wrongfully detained at higher rates than non-recidivating individuals in the other demographic group(s).

[25] These specific examples are from the Equivant/Northpoint COMPAS risk assessment; see sample questionnaire at https://assets.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf

[26] This list is by no means exhaustive. Another approach involves attempting to de-bias datasets by removing all information regarding the protected class variables. See, e.g., James E. Johndrow & Kristian Lum, *An algorithm for removing sensitive information: application to race-independent recidivism prediction*, (Mar. 15, 2017), https://arxiv.org/pdf/1703.04957.pdf. Not only would the protected class variable itself be removed but also variation in other variables that is correlated with the protected class variable. This would yield predictions that are independent of the protected class variables, but could have negative implications for accuracy. This method formalizes the notion of fairness known as "demographic parity," and has the advantage of minimizing disparate impact, such that outcomes should be proportional across demographics. Similar to affirmative action, however, this approach would raise additional fairness questions given different baselines across demographics.

[27] See Moritz Hardt, Eric Price, & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, Proc. NeurIPS 2016, https://arxiv.org/pdf/1610.02413.pdf.

Methods to Mitigate Bias (continued)

One caveat to this approach is that corrections to ensure protected classes have identical or similar false positive rates will result in differences in overall predictive accuracy between these groups.[28] Thus, if an equal opportunity correction is used, then differences in overall accuracy must be evaluated.[29]

2. A second approach would be to prioritize producing models where the predictive parity of scores is the same across different demographic groups. This property is known as "calibration within groups" and has the benefit of making scores more interpretable across groups. Calibration within groups would entail, for instance, that individuals with a score of 60% have a 60% chance of recidivating, regardless of their demographic group.

   The issue with this approach is that ensuring predictive parity comes at the expense of the equal opportunity measure described above.[30] For instance, the COMPAS tool, which is optimized for calibration within groups, has been criticized for its disparate false positive rates. In fact, ProPublica found that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.[31] This indicates that group-calibrated risk assessment tools may impact non-recidivating individuals differently, depending on their race.[32]

3. A third approach involves using causal inference methods to formalize the permissible and impermissible causal relationships between variables and make predictions using only the permissible pathways.[33] An advantage of this approach is that it formally addresses the difference between correlation and causation and clarifies the causal assumptions underlying the model. It also only removes correlation to the protected class that results from problematic connections between the variables, preserving more information from the data. The shortcoming of this approach is that it requires the toolmaker to have a good understanding of the causal relationships between the relevant variables, so additional subject-matter expertise is necessary to create a valid causal model (Figure 4 shows a simple example in a hypothetical insurance case, but recidivism predictions will likely be far more complex). Moreover, the toolmaker needs to identify

---

[28] This is due to different baseline rates of recidivism for different demographic groups in U.S. criminal justice data. See J. Kleinberg, S. Mullainathan, M. Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores.* Proc. ITCS, (2017), https://arxiv.org/abs/1609.05807 and A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Proc. FAT/ML 2016, https://arxiv.org/abs/1610.07524. Another caveat is that such a correction can reduce overall utility, as measured as a function of the number of individuals improperly detained or released. See, e.g., Sam Corbett-Davies et al., *Algorithmic Decision-Making and the Cost of Fairness*, (2017), https://arxiv.org/pdf/1701.08230.pdf.

[29] As long as the training data show higher arrest rates among minorities, statistically accurate scores *must of mathematical necessity* have a higher false positive rate for minorities. For a paper that outlines how equalizing FPRs (a measure of unfair treatment) requires creating some disparity in predictive accuracy across protected categories, see J. Kleinberg, S. Mullainathan, M. Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores.* Proc. ITCS, (2017), https://arxiv.org/abs/1609.05807; for arguments about the limitations of FPRs as a sole and sufficient metric, *see e.g.* Sam Corbett-Davies and Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, working paper, https://arxiv.org/abs/1808.00023.

[30] Geoff Pleiss et al. *On Fairness and Calibration* (describing the challenges of using this approach when baselines are different), https://arxiv.org/pdf/1709.02012.pdf.

[31] The stance that unequal false positive rates represents material unfairness was popularized in a study by Julia Angwin et al. *Machine Bias*, ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, (2016), and confirmed in further detail in e.g, Julia Dressel and Hany Farid, *The accuracy, fairness and limits of predicting recidivism*, Science Advances, 4(1), (2018), http://advances.sciencemag.org/content/advances/4/1/eaao5580.full.pdf. Whether or not FPRs are the right measure of fairness is disputed within the statistics literature.

[32] See, e.g., Alexandra Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, Big Data 5(2), https://www.liebertpub.com/doi/full/10.1089/big.2016.0047, (2017).

[33] See, e.g., Niki Kilbertus et al., *Avoiding Discrimination Through Causal Reasoning*, (2018), https://arxiv.org/pdf/1706.02744.pdf.

Methods to Mitigate Bias (continued)

which causal relationships are problematic and which are not,[34] so validity further depends on the toolmaker exercising proper judgment.

Given that some of these approaches are in tension with each other, it is not possible to simultaneously optimize for all of them. Nonetheless, these approaches can highlight relevant fairness issues to consider in evaluating tools. For example, even though it is generally not possible to simultaneously satisfy calibration within groups and equal opportunity (Methods #1 and #2 above) with criminal justice data, it would be reasonable to avoid using tools that either have extremely disparate predictive parity across demographics (i.e., poor calibration within groups) or extremely disparate false positive rates across demographics (i.e., low equal opportunity).

Given that each of these approaches involves inherent trade-offs,[35] it is also reasonable to use a few different methods and compare the results between them. This would yield a range of predictions that could better inform decision-making.[36] In addition, appropriate paths for consideration include relying on timely, properly resourced, individualized hearings rather than machine learning tools, developing cost-benefit analyses that place explicit value on avoiding disparate impact,[37] or delaying tool deployment until further columns of high quality data can be collected to facilitate more-accurate and less-biased predictions.

---

[34] Formally, the toolmaker must distinguish "resolved" and "unresolved" discrimination. Unresolved discrimination results from a direct causal path between the protected class and predictor that is not blocked by a "resolving variable." A resolving variable is one that is influenced by the protected class variable in a manner that we accept as nondiscriminatory. For example, if women are more likely to apply for graduate school in the humanities and men are more likely to apply for graduate school in STEM fields, and if humanities departments have lower acceptance rates, then women might exhibit lower acceptance rates overall even if conditional on department they have higher acceptance rates. In this case, the department variable can be considered a resolving variable if our main concern is discriminatory admissions practices. *See, e.g.*, Niki Kilbertus et al., *Avoiding Discrimination Through Causal Reasoning*, (2018), https://arxiv.org/pdf/1706.02744.pdf.

[35] In addition to the trade-offs highlighted in this section, it should be noted that these methods require a precise taxonomy of protected classes. Although it is common in the United States to use simple taxonomies defined by the Office of Management and Budget (OMB) and the US Census Bureau, such taxonomies cannot capture the complex reality of race and ethnicity. *See* Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, 62 Fed. Reg. 210 (Oct 1997), https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf. Nonetheless, algorithms for bias correction have been proposed that detect groups of decision subjects with similar circumstances automatically. For an example of such an algorithm, see Tatsunori Hashimoto et al., *Fairness Without Demographics in Repeated Loss Minimization*, Proc. ICML 2018, http://proceedings.mlr.press/v80/hashimoto18a/hashimoto18a.pdf. Algorithms have also been developed to detect groups of people that are spatially or socially segregated. See, e.g., Sebastian Benthall & Bruce D. Haynes, *Racial categories in machine learning*, Proc. FAT* 2019, https://dl.acm.org/authorize.cfm?key=N675470. Further experimentation with these methods is warranted. For one evaluation, see Jon Kleinberg, *An Impossibility Theorem for Clustering*, Advances in Neural Information Processing Systems 15, NeurIPS 2002.

[36] The best way to do this deserves further research on human-computer interaction. For instance, if judges are shown multiple predictions labelled "zero disparate impact for those who will not reoffend", "most accurate prediction," "demographic parity," etc, will they understand and respond appropriately? If not, decisions about what bias corrections to use might be better made at the level of policymakers or technical government experts evaluating these tools.

[37] Cost benefit models require explicit tradeoff choices to be made between different objectives including liberty, safety, and fair treatment of different categories of defendants. These choices should be explicit, and must be made transparently and accountably by policymakers. For a macroscopic example of such a calculation see David Roodman, *The Impacts of Incarceration on Crime*, Open Philanthropy Project report, September 2017, p p131, at *https://www.openphilanthropy.org/files/Focus_Areas/Criminal_Justice_Reform/The_impacts_of_incarceration_on_crime_10.pdf.*

Requirement 3: Tools must not conflate multiple distinct predictions

Risk assessment tools must not produce composite scores that combine predictions of different outcomes for which different interventions are appropriate. In other words, the tool should predict the specific risk it is hoping to measure, and produce separate scores for each type of risk (as opposed to a single risk score reflecting the risk of multiple outcomes). For instance, risk assessment tools should not conflate a defendant's risk of failure to appear for a scheduled court date with the risk of rearrest. Many existing pretrial risk assessment tools, however, do exactly this: they produce a single risk score that represents the risk of failure to appear or rearrest occurring.[38] In some cases this may violate local law; many jurisdictions only permit one cause as a basis for pretrial detention. And regardless of the legal situation, a hybrid prediction is inappropriate on statistical grounds.

Different causal mechanisms drive each of the phenomena that are combined in hybrid risk scores.[39] The reasons for someone not appearing in court, getting re-arrested, and/or getting convicted of a future crime are all very distinct, so a high score would not be readily interpretable and would group together people who are likely to have a less dangerous outcome (not appearing in court) with more dangerous outcomes (being convicted of a future crime).[40] In addition, as a matter of statistical validity, past convictions for non-violent crimes that have since been decriminalized (e.g., marijuana possession in many states) arguably should be considered differently from other kinds of convictions if the goal is to predict future crime or public safety risk.

Moreover, different types of intervention (both as a policy and a legal matter) are appropriate for each of these different phenomena.[41] Risk assessment tools should only be deployed in the specific context for which they were intended, including at the specific stage of a criminal proceeding and to the specific population for which they were meant to predict risk. For example, the potential risk of failing to appear to a court date at a pretrial stage should have no bearing in a sentencing hearing.[42] Likewise, predicting risks for certain segments of the population, such as juveniles, is distinct from predicting risks for the general population.

Risk assessment tools must be clear about which of these many distinct predictions they are making, and steps should be taken to safeguard against conflating different predictions and using risk scores in inappropriate contexts.

---

[38] Sandra G. Mayson, *Dangerous Defendants*, 127 Yale L.J. 490, 509-510 (2018).

[39] *Id.*, at 510. ("The two risks are different in kind, are best predicted by different variables, and are most effectively managed in different ways.")

[40] For instance, needing childcare increases the risk of failure to appear (*see* Brian H. Bornsein, Alan J. Thomkins & Elizabeth N. Neely, *Reducing Courts' Failure to Appear Rate: A Procedural Justice Approach*, U.S. DOJ report 234370, *available at* https://www.ncjrs.gov/pdffiles1/nij/grants/234370.pdf) but is less likely to increase the risk of recidivism.

[41] For example, if the goal of a risk assessment tool is to advance the twin public policy goals of reducing incarceration and ensuring defendants appear for their court dates, then the tool should not conflate a defendant's risk of knowingly fleeing justice with their risk of unintentionally failing to appear, since the latter can be mitigated by interventions besides incarceration (e.g. giving the defendant the opportunity to sign up for phone calls or SMS-based reminders about their court date, or ensuring the defendant has transportation to court on the day they are to appear).

[42] Notably, part of the holding in Loomis, mandated a disclosure in any Presentence Investigation Report that COMPAS risk assessment information "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," Wisconsin v. Loomis (881 N.W.2d 749).

Human-Computer Interface Issues

While risk assessment tools provide input and recommendations to decision-making processes, the ultimate decision-making authority still resides in the hands of humans. Judges, court clerks, pretrial services officers, probation officers, and prosecutors all may use risk assessment scores to guide their judgments. Thus, critical human-computer interface issues must also be addressed when considering the use of risk assessment tools.

One of the key challenges of statistical decision-making tools is the phenomenon of automation bias, where information presented by a machine is viewed as inherently trustworthy and above skepticism.[43] This can lead humans to over-rely on the accuracy or correctness of automated systems.[44] The holding in *Wisconsin v. Loomis* [45] indirectly addressed the issue of automation bias by requiring that any Presentence Investigation Report containing a COMPAS risk assessment be accompanied by a written disclaimer that the scores may be inaccurate and have been shown to disparately classify offenders.[46] While disclosure regarding the limitations of risk assessment tools is an important first step, it is still insufficient. Over time, there is the risk that judges become inured to lengthy disclosure language repeated at the beginning of each report. Moreover, the disclosures do not make clear how, if at all, judges should interpret or understand the practical limits of risk assessments.

This section attempts to illustrate how to safeguard against automation bias and other critical human-computer interface issues by ensuring (i) risk assessment tools are easily interpretable by human users, (ii) users of risk assessment tools receive information about the uncertainty behind the tools' predictions, and (iii) adequate resources are dedicated to fund proper training for use of these tools.

---

[43] M.L. Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.2634&rep=rep1&type=pdf.

[44] It is important to note, however, that there is also evidence of the opposite phenomenon, whereby users might simply ignore the risk assessment tools' predictions. In Christin's ethnography of risk assessment users, she notes that professionals often "buffer" their professional judgment from the influence of automated tools. She quotes a former prosecutor as saying of risk assessment, "When I was a prosecutor I didn't put much stock in it, I'd prefer to look at actual behaviors. I just didn't know how these tests were administered, in which circumstances, with what kind of data." From Christin, A., 2017, *Algorithms in practice: Comparing web journalism and criminal justice*, Big Data & Society, *4(2)*.

[45] See Wisconsin v. Loomis (881 N.W.2d 749).

[46] "Specifically, any PSI containing a COMPAS risk assessment must inform the sentencing court about the following cautions regarding a COMPAS risk assessment's accuracy: (1) the proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined; (2) risk assessment compares defendants to a national sample, but no cross- validation study for a Wisconsin population has yet been completed; (3) some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism; and (4) risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations." Wisconsin v. Loomis (881 N.W.2d 749).

Requirement 4: Predictions and how they are made must be easily interpretable

While advocates have focused on the issues mentioned above of bias in risk prediction scores, one often overlooked aspect of fairness is the way risk scores are translated for human users. Developers and jurisdictions deploying risk assessment tools must ensure that tools convey their predictions in a way that is straightforward to human users and illustrate how those predictions are made. This means ensuring that interfaces presented to judges, clerks, lawyers, and defendants are clear, easily understandable, and not misleading.[47]

Interpretability involves providing users with an understanding of the relationship between input features and output predictions. We should caution that this may not mean restricting the model to an "interpretable" but less accurate mathematical form, but instead using techniques that provide separate interpretations for more complex predictions.[48]

Providing interpretations for predictions can help users understand how each variable contributes to the prediction, and how sensitive the model is to certain variables. This is crucial for ensuring that decision-makers are consistent in their understandings of how models work and the predictions they produce, and that the misinterpretation of scores by individual judges does not result in the disparate application of justice. Because interpretability is a property of the tools as used by people, it requires consideration of the use of risk assessments in context and depends on how effectively they can be employed as tools by their human users.

At the same time, developers of models should ensure that the intuitive interpretation is not at odds with intended risk prediction. For instance, judges or other users might intuitively guess that ordered categories are of similar size, represent absolute levels of risk rather than relative assessments, and cover the full spectrum of approximate risk levels.[49] Thus, on a 5-point scale, a natural interpretation would be that a score of one implies a 0% to 20% risk of reoffending (or another outcome of interest), category two a 21% to 40% risk, and so on. However, this is not the case for many risk assessment tools.

One study compared the Pretrial Risk Assessment Tool (PTRA), which converts risk scores into a 5-point risk scale, with the actual likelihood of the outcome (in this case, rearrest, violent rearrest, failure to appear, and/or bail revocation).[50] Only 35% of defendants classified at the highest risk level failed to appear for trial or were rearrested before trial. The probabilities of failure to appear and of rearrest for all risk levels were within the intuitive interval for the lowest risk level.[51]

---

[47] Computer interfaces, even for simple tasks, can be highly confusing to users. For example, one study found that users failed to notice anomalies on a screen designed to show them choices they had previously selected for confirmation over 50% of the time, even after carefully redesigning the confirmation screen to maximize the visibility of anomalies. See Campbell, B. A., & Byrne, M. D. (2009). *Now do voters notice review screen anomalies? A look at voting system usability*, Proceedings of the 2009 Electronic Voting Technology Workshop/ Workshop on Trustworthy Elections (EVT/WOTE '09).

[48] This point depends on the number of input variables used for prediction. With a model that has a large number of features (such as COMPAS), it might be appropriate to use a method like gradient-boosted decision trees or random forests, and then provide the interpretation using an approximation. *See* Zach Lipton, *The Mythos of Model Interpretability*, Proc. ICML 2016, *available at* https://arxiv.org/pdf/1606.03490.pdf, §4.1. For examples of methods for providing explanations of complex models, *see, e.g.*, Gilles Louppe *et al.*, *Understanding the variable importances in forests of randomized trees*, Proc. NIPS 2013, *available at* https://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf; Marco Ribeiro, LIME - Local Interpretable Model-Agnostic Explanations, *at* https://homes.cs.washington.edu/~marcotcr/blog/lime/. For smaller feature sets, as used by some other risk assessment tools (perhaps anything fewer than 10-20 features, depending on collinearity or mutual information), a more interpretable linear model may be appropriate.

[49] Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment*, 46(2) Criminal Justice and Behavior 185–209 (2018), https://doi.org/10.1177/0093854818811379

[50] *See id.*

[51] *See id.*

Similarly, there are also substantial gaps between the intuitive and the correct interpretations of risk categories in Colorado's Pretrial Assessment Tool.[52] In order to mitigate these shortcomings, jurisdictions would need to collect data and conduct further research on user interface choices, information display, and users' psychological responses to information about prediction uncertainty.[53]

Requirement 5: Tools should produce confidence estimates for their predictions

An important component of any statistical prediction is the uncertainty underlying it. In order for users of risk assessment tools to appropriately and correctly interpret their results, it is vital that reports of their predictions include error bars, confidence intervals, or other similar indications of reliability. For example, risk assessment tools often produce a score reflecting a probability of reoffending, or a mapping of those probabilities into levels (like "high," "medium," and "low" risk).[54] This information alone, however, does not give the user an indication of the model's confidence in its prediction. For example, even if a model is calibrated such that an output like "high risk" corresponds to "a 60% probability of reoffending," it is unclear whether the tool is confident that the defendant has a probability of reoffending between 55% and 65%, with a mean of 60%, or if the tool is only confident that the defendant has a probability of reoffending between 30% and 90%, with a mean of 60%. In the former case, the interpretation that the defendant has a 60% probability of reoffending is far more reasonable than in the latter case, where there is overwhelming uncertainty around the prediction.

For this reason, risk assessment tools should not be used unless they are able to provide good measures of the certainty of their own predictions, both in general and for specific individuals on which they are used. There are many sources of uncertainty in recidivism predictions, and ideally disclosure of uncertainty in predictions should capture as many of these sources as possible. This includes the following:

- Uncertainty due to sample size and the presence of outliers in datasets. This type of uncertainty can be measured by the use of bootstrapped confidence intervals,[55] which are commonly used by technology companies for assessing the predictive power of models before deployment.

- Uncertainty about the most appropriate mitigation for model bias, as discussed in Requirement 2. One possibility would be to evaluate the outcomes of different fairness corrections as expressing upper and lower bounds on possible "fair" predictions.[56]

---

[52] The lowest risk category for the Colorado Pretrial Assessment Tool (CPAT) included scores 0-17, while the highest risk category included a much broader range of scores: 51-82. In addition, the highest risk category corresponded to a Public Safety Rate of 58% and a Court Appearance Rate of 51%. Pretrial Justice Institute, (2013). Colorado Pretrial Assessment Tool (CPAT): Administration, scoring, and reporting manual, Version 1. Pretrial Justice Institute. Retrieved from http://capscolorado.org/yahoo_site_admin/assets/docs/CPAT_Manual_v1_-_PJI_2013.279135658.pdf

[53] User and usability studies such as those from the human-computer interaction field can be employed to study the question of how much deference judges give to pretrial or pre-sentencing investigations. For example, a study could examine how error bands affect judges' inclination to follow predictions or (when they have other instincts) overrule them.

[54] As noted in Requirement 4, these mappings of probabilities to scores or risk categories are not necessarily intuitive, i.e. they are often not linear or might differ for different groups.

[55] In a simple machine learning prediction model, the tool might simply produce an output like "35% chance of recidivism." A *bootstrapped* tool uses many resampled versions of the training datasets to make different predictions, allowing an output like, "It is 80% likely that this individual's chance of recidivating is in the 20% - 50% range." Of course these error bars are still relative to the training data, including any sampling or omitted variable biases it may reflect.

[56] The specific definition of fairness would depend on the fairness correction used.

- Uncertainty as a result of sampling bias and other fundamental dataset problems, as discussed in Requirement 1. This is a complicated issue to address, but one way to approach this problem would be to find or collect new high quality secondary sources of data to estimate uncertainty due to problems with the training dataset.
- User interfaces to satisfactorily display and convey uncertainty to users are in some respects also an open problem, so the training courses we suggest in Requirement 6 should specifically test and assist users in making judgments under simulations of this uncertainty.

Requirement 6: Users of risk assessment tools must attend trainings on the nature and limitations of the tools

Regardless of how risk assessment outputs are explained or presented, clerks and pretrial assessment services staff must be trained on how to properly code data about individuals into the system. Human error and a lack of standardized best practices for data input could have serious implications for data quality and validity of risk prediction down the line.

At the same time, judges, attorneys, and other relevant stakeholders must also receive rigorous training on how to interpret the risk assessments they receive. For any such tool to be used appropriately, judges, attorneys, and court employees should have regular training to understand the function of the tool itself and how to interpret risk classifications such as quantitative scores or more qualitative "low/medium/high" ratings. These trainings should address the considerable limitations of the assessment, error rates, interpretation of scores, and how to challenge or appeal the risk classification. It should likely include basic training on how to understand confidence intervals.[57] More research is required on how these risk assessment tools inform human decisions, in order to determine what forms of training will support principled and informed application of these tools, and where gaps exist in current practice.[58]

[57] Humans are not naturally good at understanding probabilities or confidence estimates, though some training materials and games exist that can teach these skills; *see e.g.*: https://acritch.com/credence-game/

[58] To inform this future research, DeMichele et al.'s study conducting interviews with judges using the PSA tool can provide useful context for how judges understand and interpret these tools. DeMichele, Matthew and Comfort, Megan and Misra, Shilpi and Barrick, Kelle and Baumgartner, Peter, *The Intuitive-Override Model: Nudging Judges Toward Pretrial Risk Assessment Instruments*, (April 25, 2018). Available at SSRN: https://ssrn.com/abstract=3168500 or http://dx.doi.org/10.2139/ssrn.3168500;

Governance, Transparency, and Accountability

As risk assessment tools supplement judicial processes and represent the implementation of local policy decisions, jurisdictions must take responsibility for their governance. Importantly, they must remain transparent to citizens and accountable to the policymaking process. Such governance requires (i) stakeholder and broad public engagement in the design and oversight of such systems;[59] (ii) transparency around the data and methods used for creating these tools;[60] (iii) disclosure of relevant information to defendants to allow them to contest decisions informed by these tools; and (iv) pre-deployment[61] and ongoing evaluation of the tool's validity, fitness for purpose, and role within the broader justice system.

Requirement 7: Policymakers must ensure that public policy goals are appropriately reflected in these tools

The use of risk assessment tools has the potential to obscure—and remove from the public eye—fundamental policy decisions concerning criminal justice. These include choices about the point at which societal risk outweighs the considerable harm of detention to a defendant and their family, and how certain a risk must be before the criminal justice system is required to act on it (i.e., how accurate, valid, and unbiased a prediction needs to be before it should be relied upon to deprive an individual of liberty). Use of these tools also includes choices about the nature and definition of protected categories and how they are used. In addition, important decisions must be made about how such tools interact with non-incarcerative measures aimed at rehabilitation, such as diversion measures or provision of social services. These are challenging policy questions that cannot and should not be answered by toolmakers alone, and will instead require active engagement from policymakers, judicial system leaders, and the general public.

One key example of how seemingly technical decisions are actually policy decisions is the choice of thresholds for detention. California's S.B. 10 legislation, for example, would create a panel to establish thresholds that define probabilistic risk as "low," "medium," or "high" of failing to appear for court, or committing another crime that poses a risk to public safety.[62] Meanwhile, the First Step Act requires the Attorney General to develop a risk assessment system to classify inmates as having a minimum, low, medium, or high risk of committing another crime in the future.[63] The selection of these thresholds will ultimately determine how many people are detained versus released.

Risk thresholds like those mandated by S.B. 10 and the First Step Act are policy choices that must be chosen with respect to the broader criminal justice process, specific criminal justice policy objectives, and appropriate data to inform those objectives. Policymakers at both the state and federal level must decide which trade-offs to make to ensure just outcomes and lower the social costs of detention.

---

[59] See the University of Washington's Tech Policy Lab's Diverse Voices methodology for a structured approach to inclusive requirements gathering. Magassa, Lassana, Meg Young, and Batya Friedman, *Diverse Voices*, (2017), http://techpolicylab.org/diversevoicesguide/.

[60] Such disclosures support public trust by revealing the existence and scope of a system, and by enabling challenges to the system's role in government. See Pasquale, Frank. *The black box society: The secret algorithms that control money and information*. Harvard University Press, (2015). Certain legal requirements on government use of computers demand such disclosures. At the federal level, the Privacy Act of 1974 requires agencies to publish notices of the existence of any "system of records" and provides individuals access to their records. Similar data protection rules exist in many states and in Europe under the General Data Protection Regulation (GDPR).

[61] Reisman, Dillon, Jason Schultz, Kate Crawford, Meredith Whittaker, *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, AI Now Institute, (2018).

[62] See Cal. Crim. Code §§ 1320.24 (e) (7), 1320.25 (a), effective Oct 2020.

[63] First Step Act, H.R.5682 — 115th Congress (2017-2018).

For example, if a major goal is to reduce mass incarceration in the criminal justice system, thresholds should be set such that the number of individuals classified in higher risk categories is reduced. In addition to gathering input from relevant stakeholders, threshold-setting bodies (whether legislatures, panels, or other agencies) should practice evidence-based policymaking informed by relevant and timely crime rates data, and plan to revisit and revise their decisions on an ongoing basis.

Requirement 8: Tool designs, architectures, and training data must be open to research, review, and criticism

Risk assessment tools embody important public policy decisions made by governments, and must be as open and transparent as any law, regulation, or rule of court. Thus, governments must not deploy any proprietary risk assessments that rely on claims of trade secrets to prevent transparency.[64]

In particular, the training datasets, architectures, algorithms, and models of all tools under consideration for deployment must be made broadly available to all interested research communities—such as those from statistics, computer science, social science, public policy, law, and criminology, so that they are able to evaluate them before and after deployment.[65]

We note that much of the technical research literature on fairness that has appeared in the past two years resulted from ProPublica's pioneering work in publishing a single dataset related to the Northpointe COMPAS risk assessment tool, which was obtained via public records requests in Broward County, Florida.[66] Publishing such datasets enables the independent research and public discourse required to evaluate their effectiveness.

However, it is important to note that when such datasets are shared, appropriate de-identification techniques should be used to ensure that non-public personal information cannot be derived from the datasets.[67] Given increasingly sophisticated information triangulation and re-identification techniques,[68] additional measures might be necessary, such as contractual conditions that the recipients use the data only for specific purposes, and that once those purposes are accomplished, they delete their copy of the dataset.[69]

---

[64] For further discussion on the social justice concerns related to using trade secret law to prevent the disclosure of the data and algorithms behind risk assessment tools, see Taylor R. Moore, *Trade Secrets and Algorithms as Barriers to Social Justice, Center for Democracy and Technology* (August 2017), https://cdt.org/files/2017/08/2017-07-31-Trade-Secret-Algorithms-as-Barriers-to-Social-Justice.pdf.

[65] Several countries already publish the details of their risk assessment models. See, e.g., Tollenaar, Nikolaj, et al. *StatRec-Performance, validation and preservability of a static risk prediction instrument*, Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 129.1 (2016): 25-44 (in relation to the Netherlands); *A Compendium of Research and Analysis on the Offender Assessment System* (OaSys) (Robin Moore ed., Ministry of Justice Analytical Series, 2015) (in relation to the United Kingdom). Recent legislation also attempts to mandate transparency safeguards, see Idaho Legislature, House Bill No.118 (2019).

[66] *See, e.g.*, Jeff Larson *et al. How We Analyzed the COMPAS Recidivism Algorithm*, ProPublica (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. For a sample of the research that became possible as a result of ProPublica's data, see https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=propublica+fairness+broward (showing 154 academic citations to ProPublica's dataset as of April 2019, many of which build on or reanalyze it). Data provided by Kentucky's Administrative Office of the Courts has also enabled scholar's to examine the impact of the implementation of the PSA tool in that state. Stevenson, Megan, Assessing Risk Assessment in Action (June 14, 2018). Minn. L. Rev, 103, Forthcoming; *available at* https://ssrn.com/abstract=3016088.

[67] For an example of how a data analysis competition dealt with privacy concerns when releasing a dataset with highly sensitive information about individuals, see Ian Lundberg et al., *Privacy, ethics, and data access: A case study of the Fragile Families Challenge* (Sept. 1, 2018), https://arxiv.org/pdf/1809.00103.pdf.

[68] See Arvind Narayanan et al., *A Precautionary Approach to Big Data Privacy* (Mar. 19, 2015), http://randomwalker.info/publications/precautionary.pdf.

[69] *See id.* at p. 20 and 21 (describing how some sensitive datasets are only shared after the recipient completes a data use course, provides information about the recipient, and physically signs a data use agreement).

Requirement 9: Tools must support data retention and reproducibility to enable meaningful contestation and challenges

In order for defendants to contest decisions made by risk assessment tools, they must have access to information about how the tools' predictions are made.[70] As discussed above, there are many potential technical concerns related to the use of these tools, in particular with regard to bias. Given the adversarial nature of the U.S. criminal justice system, which depends on defendants and their attorneys to advance any arguments in their favor, denying defendants the ability to access information about how these decisions are made hampers their ability to contest these decisions.

Individual-level information used in the assessments should be recorded in an audit trail that is made available to defendants, counsel, and judges. Such audit trails must be maintained in an immutable form for future review, so auditors can achieve a reproducible calculation for that individual's level of risk.[71] Defendants should also have an opportunity to contest any inaccuracies in the input information or inferences in the resulting risk classification and to present additional mitigating information.[72] This is especially important given the potential for risk assessment tools to be manipulated. For example, risk assessments often rely on questionnaires administered to arrestees, which presents the opportunity for abuse by administrators, as illustrated by instances of "criteria tinkering."[73] Adversarial analysis is likely the best way to protect against such manipulations. Through these processes, defendants can demonstrate how applicable and robust risk assessments are or are not with respect to their particular circumstances.

---

[70] For a discussion of the due process concerns that arise when information is withheld in the context of automated decision-making, see Danielle Keats Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249 (2007), https://ssrn.com/abstract=1012360. See also, Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer,* 43 Hastings L. J. 1321 (1992).

[71] Additionally, the ability to reconstitute decisions evidences procedural regularity in critical decision processes and allows individuals to trust the integrity of automated systems even when they remain partially non-disclosed. *See* Joshua A. Kroll et al., *Accountable algorithms*, 165 U. Pa. L. Rev. 633 (2016).

[72] The ability to contest scores is not only important for defendant's rights to adversarially challenge adverse information, but also for the ability of judges and other professionals to engage with the validity of the risk assessment outputs and develop trust in the technology. *See* Daniel Kluttz et al., *Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems* (January 2019), https://dx.doi.org/10.2139/ssrn.3311894

[73] "Criteria tinkering" occurs when court clerks manipulate input values to obtain the score they think is correct for a particular defendant. *See* Hannah-Moffat, Kelly, Paula Maurutto, and Sarah Turnbull, *Negotiated risk: Actuarial illusions and discretion in probation*, 24.3 Canada J. of L. & Society/La Revue Canadienne Droit et Société 391 (2009). *See also* Angele Christin, *Comparing Web Journalism and Criminal Justice*, 4.2 Big Data & Society 1.

Requirement 10: Jurisdictions must take responsibility for the post-deployment evaluation, monitoring, and auditing of these tools

Jurisdictions must periodically publish an independent review, algorithmic impact assessment, or audit of all risk assessment tools they use to verify that the requirements listed in this report have been met.[74] Subsequent audits will need to examine the outcomes and operation of the system on a regular basis. Such review processes must also be localized because the conditions of crime, law enforcement response, and culture among judges and clerks are all local phenomena.[75] These processes should ideally operate with staff support and buy-in within judicial institutions, while also drawing on external expertise.

To ensure transparency and accountability, an independent outside body (such as a review board) must be responsible for overseeing the audit. This body should be comprised of legal, technical, and statistical experts, currently and formerly incarcerated individuals, public defenders, public prosecutors, judges, and civil rights organizations, among others. These audits and their methodology must be open to public review and comment. To mitigate privacy risks, published versions of these audits should be redacted and sufficiently blinded to prevent de-anonymization.[76]

A current challenge to implementing these audits is a lack of data needed to assess the consequences of those tools already deployed. When some partners of PAI tried to assess the consequences of California's pretrial risk assessment legislation, they found inadequate data on the pretrial detention population in California and could not identify data or studies to understand how the definition of low, medium, high risk and their thresholds could impact how many people are held or released pre-trial. Similarly, evaluating or correcting tools and training data for error and bias requires better data on discrimination at various points in the criminal justice system. In order to understand the impact of current risk assessment tools, whether in pretrial, sentencing, or probation, court systems should collect data on pretrial decisions and outcomes. In addition, data on individual judges' decisions before and after an intervention should be collected and analyzed.

To meet these responsibilities, whenever legislatures or judicial bodies decide to mandate or purchase risk assessment tools, those authorities should simultaneously ensure the collection of post-deployment data, provide the resources to do so adequately, and support open analysis and review of the tools in deployment. That requires both (i) allocation or appropriation of sufficient funding for those needs and (ii) institutional commitment to recruiting (or contracting with) statistical/technical and criminological expertise to ensure that data collection and review are conducted appropriately.

---

[74] For further guidance on how such audits and evaluations might be structured, *see*, AI Now Institute, *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, https://ainowinstitute.org/aiareport2018.pdf; Christian Sandvig et al., *Auditing algorithms: Research methods for detecting discrimination on internet platform* (2014).

[75] *See* John Logan Koepke and David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 Wash. L. Rev. 1725 (2018).

[76] For a discussion Latanya Sweeney & Ji Su Yoo, *De-anonymizing South Korean Resident Registration Numbers Shared in Prescription Data*, Technology Science, (Sept. 29, 2015), https://techscience.org/a/2015092901. Techniques exist that can guarantee that re-identification is impossible. See the literature on methods for *provable privacy, notably differential privacy*. A good introduction is in Kobbi Nissim, Thomas Steinke, Alexandra Wood, Mark Bun, Marco Gaboardi, David R. O'Brien, and Salil Vadhan, *Differential Privacy: A Primer for a Non-technical Audience*, http://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_0.pdf.

Conclusion

Efforts to move the U.S. criminal justice system to evidence-based policymaking and public-safety-oriented decision-making are laudable and extremely important. As a matter of historical and international comparison, the U.S. incarcerates an abnormally high number of people (in absolute numbers, per capita, and per crime rate: see Figures 1-3). Thus, significant reforms to address that problem are justified and urgent based on the available data. This context has driven the adoption of risk assessment tools, and it is crucial to note that nothing in this report should be read as calling for a slowing of criminal justice reform and efforts to mitigate mass incarceration.

Rather, our aim is to help policymakers make informed decisions about the risk assessment tools currently in deployment and required under legislative mandates, and the potential policy responses they could pursue. One approach is for jurisdictions to cease using the tools in decisions to detain individuals until they can be shown to have overcome the numerous validity, bias, transparency, procedural, and governance problems that currently beset them. This path need not slow the overall process of criminal justice reform. In fact, several advocacy groups have proposed alternative reforms that do not introduce the same concerns as risk assessment tools.[77] Accordingly, the choice is not simply between current systems like cash bail and newer algorithmic systems.

Another option is to embark on the project of trying to improve risk assessment tools. That would necessitate procurement of sufficiently extensive and representative data, development and evaluation of reweighting methods, and ensuring that risk assessment tools are subject to open, independent research and scrutiny. The ten requirements outlined in this report represent a minimum standard for developers and policymakers attempting to align their risk assessment tools—and how they are used in practice—with well-founded policy objectives.

While the widespread use of risk assessments continues, administrative agencies and legislatures driving deployment have a responsibility to set standards for the tools they are propagating. In addition to the ten requirements we have outlined in this report, jurisdictions will also need to gather and incorporate significant expertise from the fields of machine learning, statistics, human-computer interaction, criminology, and law in order to perform this task. At this stage, we should emphasize that we do not believe that any existing tools would meet properly set standards on all of these points, and in the case of Requirement 1, meeting an appropriately set standard would require major new data collection efforts.

PAI believes standard setting in this space is essential work for policymakers because of the enormous momentum that state and federal legislation have placed behind risk assessment procurement and deployment. But many of our members remain concerned that standards could be set with the aim of being easy to meet, rather than actually confronting the profound statistical and procedural problems inherent in using such tools to inform detention decisions. It would be tempting to set standards that gloss over complex accuracy, validity, and bias problems, and to continue deployment of tools without considering alternative reforms.

For AI researchers, the task of foreseeing and mitigating unintended consequences and malicious uses has become one of the central problems of our field. Doing so requires a very cautious approach to the design and engineering of systems, as well as careful consideration of the ways that they will potentially fail and the harms that may occur as a result. Criminal justice is a domain where it is imperative to exercise maximal caution and humility in the deployment of statistical tools. We are concerned that proponents of these tools have failed

[77] Brandon Buskey and Andrea Woods, *Making Sense of Pretrial Risk Assessments*, National Association of Criminal Defense Lawyers, (June 2018), https://www.nacdl.org/PretrialRiskAssessment. Human Rights Watch proposes a clear alternative: "The best way to reduce pretrial incarceration is to respect the presumption of innocence and stop jailing people who have not been convicted of a crime absent concrete evidence that they pose a serious and specific threat to others if they are released. Human Rights Watch recommends having strict rules requiring police to issue citations with orders to appear in court to people accused of misdemeanor and low-level, non-violent felonies, instead of arresting and jailing them. For people accused of more serious crimes, Human Rights Watch recommends that the release, detain, or bail decision be made following an adversarial hearing, with right to counsel, rules of evidence, an opportunity for both sides to present mitigating and aggravating evidence, a requirement that the prosecutor show sufficient evidence that the accused actually committed the crime, and high standards for showing specific, known danger if the accused is released, as opposed to relying on a statistical likelihood." Human Rights Watch, *Q & A: Profile Based Risk Assessment for US Pretrial Incarceration, Release Decisions*, (June 1, 2018), https://www.hrw.org/news/2018/06/01/q-profile-based-risk-assessment-us-pretrial-incarceration-release-decisions.

to adequately address the minimum requirements for responsible use prior to widespread deployment.

Going forward, we hope that this report sparks a deeper discussion about these concerns with the use of risk assessment tools and spurs collaboration between policymakers, researchers, and civil society groups to accomplish much needed standard-setting and reforms in this space. The Partnership on AI would, where it is constructive, be available to provide advice and connections to the AI research community to facilitate such efforts.

PARTNERSHIP ON AI