June 2020

THE PRESENT AND FUTURE OF AI IN PRE-TRIAL RISK ASSESSMENT INSTRUMENTS

Alexandra Chouldechova
Estella Loomis McCandless Assistant Professor of Statistics and Public Policy,
Heinz College, Carnegie Mellon University, achould@cmu.edu

Kristian Lum
Assistant Research Professor, Department of Computer and Information Science,
University of Pennsylvania, kl1@seas.upenn.edu

SAFETY+JUSTICE CHALLENGE
Supported by the John D. and Catherine T. MacArthur Foundation

This report was prepared following a meeting convened by the John D. and Catherine T. MacArthur Foundation as part of the Safety and Justice Challenge, which seeks to reduce over-incarceration by changing the way America thinks about and uses jails. Core to the Challenge is a competition designed to support efforts to improve local criminal justice systems across the country that are working to safely reduce over-reliance on jails, with a particular focus on addressing disproportionate impact on low-income individuals and communities of color.

Acknowledgments

This critical issue brief was commissioned by the Pretrial Risk Management Project of the John D. and Catherine T. MacArthur Foundation. Members of the Pretrial Risk Management Project reviewed and critiqued drafts of the brief, but the authors are solely responsible for its content.

Project Members

Sarah Brayne
Alexandra Chouldechova
Domingo Corona
Khalil Cumberbatch
Sarah Desmarais
Nneka Jones-Tapia
Logan Koepke
Kristian Lum
Sandra Mayson
John Monahan
David Robinson
Vincent Southerland
Elizabeth Thornton Trosch

More information is available at **www.SafetyandJusticeChallenge.org.**

SAFETY+JUSTICE CHALLENGE

Supported by the John D. and Catherine T. MacArthur Foundation

INTRODUCTION

From music and romantic partner recommendation, to medical diagnosis and disease outbreak detection, to automated essay scoring, "Artificial intelligence" (AI) systems are being used to tackle prediction, classification, and detection tasks that impact nearly every sphere of our lives. Since the fundamental task of pre-trial risk assessment instruments is one of prediction, we anticipate that the success of AI technology in these other domains will inspire an increase in the availability of AI-based risk assessment instruments in the coming years. The purpose of this critical issue brief is primarily to equip practitioners considering adopting an AI-based pre-trial risk assessment tool to consider the relevant questions relevant to determining whether adopting such a system will result in better predictions and ultimately move their jurisdiction towards fairer, more just and decarceral pre-trial decision-making that respects civil and human rights.

AI technologies are not likely to achieve considerably greater predictive accuracy than currently available risk assessment instruments. The primary obstacle is that the behavioral outcomes these tools seek to predict, outcomes such as future arrest or court non-appearance, have an inherently high degree of uncertainty or randomness. Existing evidence suggests that any significant increase in predictive ability would need to come from uncovering hereto unknown and unused highly predictive risk factors. It is reasonable to doubt whether such factors exist. Furthermore, to the extent that incorporating new types of data does produce gains in predictive accuracy, those gains must be weighed against the ethical concerns that reliance on that data might raise. This brief suggests new directions for AI-based risk assessment tools that look beyond predictive accuracy in promoting more just and decarceral pre-trial decision-making.

WHAT IS PRE-TRIAL RISK ASSESSMENT?

Pre-trial risk assessment predicts the likelihood that a defendant will have specific pre-trial outcome if released. In this setting, the pre-trial outcomes of interest are most often failure to appear in court or re-arrest during the pre-trial period. The estimated likelihood of the pre-trial outcomes is translated into concrete policy recommendations regarding the appropriate conditions of release (if any) for the arrested person or categorical descriptions, such as "high risk" or "low risk". Risk assessment tools serve as one input in a much larger decision-making process [8].

Increasingly, courts are relying on actuarial risk assessment models (commonly referred to as "algorithms") to estimate these likelihoods. Proponents of such tools point to the tools' superiority to humans in predictive accuracy, transparency, and objectivity [8]. Policy simulations have suggested that, if the recommendations of risk assessment tools were adhered to exactly, more people could be released pretrial without increasing rates of pre-trial failure [10, 14]. As such, these tools have become a popular component of broader pre-trial reform efforts. Critics of risk assessment point to racial disparities in some measures of the tools' predictive performance as evidence of unfairness or racial bias. Other critiques point instead to perceived fundamental disconnects between the use of a predictive tool and civil rights guarantees [12]. While some may be pinning their hopes on next generation AI-based risk assessment models to alleviate these concerns, by and large, we anticipate quite the opposite. Rather than resolving today's concerns about the use of risk assessment tools in the pre-trial process, we expect that future uses of AI in risk assessment will raise even more questions. However, we also see opportunities for more advanced AI methods to improve upon existing practice by producing better locally-tailored models, and changing how the risk assessment task is formulated in the first place.

WHAT IS AI?

There is no universally agreed upon definition of artificial intelligence. Some interpret the term to apply only to cases where a computer performs calculations that are not understandable by humans or exhibits intelligence on par with if not superior to humans. However, the term "artificial intelligence" is most often used as a catch-all to describe any computational technologies that are capable of producing reasoned or "intelligent" outputs, even when the computation required to arrive at those outputs is fairly simple and understandable. Under this definition, "intelligence" can take many forms, such as the strategic decision-making, perception, knowledge representation, planning, and problem solving, among others. This encompasses systems like AlphaGo for playing board games, computer vision systems for object detection

and facial recognition, automated language translation systems, and autonomous vehicles. It also encompasses data-driven systems for making predictions about future outcomes. What is common to these systems is not their structure, complexity, or purpose, but rather the property of completing a task without a human precisely specifying all of the steps to do so—there is some aspect of the task the computer has, in some sense, *learned* on its own.

Much of the AI that is being developed and deployed today is driven by machine learning, a term that in practice has become interchangeable with AI. Machine learning is a computational approach by which the computer learns "models" or "algorithms" from the data, often with the aim of classifying unseen cases or predicting future outcomes. Given this objective, it is no surprise that AI and machine learning are increasingly discussed in the context of developing pre-trial risk assessment tools.

What AI is not

AI does not present a fundamentally new approach to criminal risk assessment. As we emphasize throughout this brief, despite the tremendous hype around AI, we have had something akin to "AI-based" risk assessment tools in the criminal justice system for decades. These are what are commonly known as "actuarial" tools. The developers of actuarial tools would not have thought of their methods as machine learning or AI. Yet actuarial risk assessment instruments arrive at predictions about future outcomes precisely by leveraging patterns and structure identified in historical data, and hence are themselves a form of machine learning-driven AI technology. When actuarial tools were introduced they marked a shift away from earlier risk assessment approaches such as structured decision making and unstructured clinical judgment. Machine learning, on the other hand, is best regarded as a family of computational approaches for producing actuarial tools, rather than a completely new risk assessment paradigm.

While the idea of deploying artificial intelligence systems to improve decision-making may sound appealing, it would be wrong to think of these systems as capable of making intelligent, multifaceted decisions. As we discuss in more detail below, these systems or "tools" are constructed with a very narrow and specific scope: they identify relationships between the available data and the specified prediction target, and apply those relationships to make predictions on future cases. To the extent that the prediction target for which the tool is optimized fails to faithfully reflect all relevant decision criteria for the task at hand, the tool will be ill-suited to serve as the sole determinant of the decision taken.

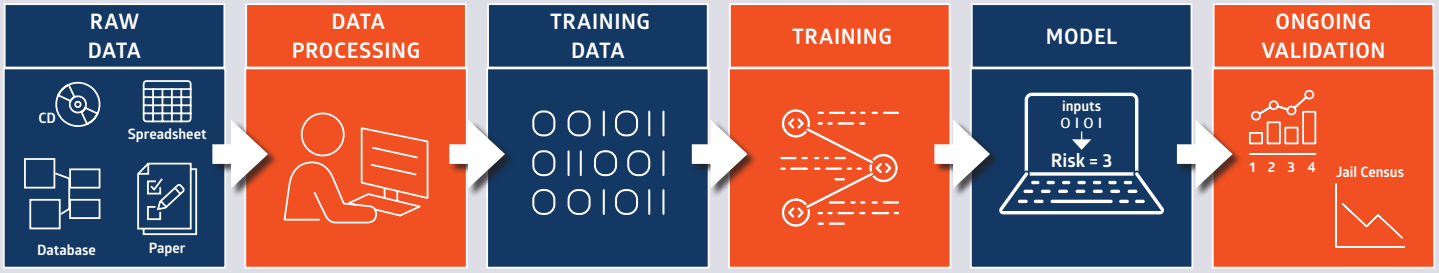
> Who is represented in the training data? Has the model been locally validated?

THE PRESENT OF AI AND RISK ASSESSMENT

Although current risk assessment models—such as those that add up integer point values associated with pre-determined risk factors to arrive at a risk score—may seem far too simplistic to be considered AI, in reality, current generation actuarial risk assessment instruments or algorithmic risk assessment tools share much in common with other systems currently marketed as AI. Like many other "AI systems", current risk assessment models are built by learning the patterns in the available data that best predict the developer-defined objective. Any patterns of "bias" encoded in the data are likely to be learned by the model and passed on via the predictions. Thus, scrutiny of any risk assessment model must begin with scrutiny of the *training data*—the data from which the model is built.

The data that informs risk assessment models

The data used to build current generation risk assessment models consists of information about individuals who interacted with the criminal justice system within a specific set of jurisdictions within a specific period of time. The information pertaining to each individual in the dataset can be divided into two categories: information that will be used to make predictions (sometimes called "risk factors" or "features") and the information that the model is built to predict (the "outcomes" or "prediction targets"). Common features include summaries of criminal history, such as the number of previous arrests; demographic information, such as age; and responses to interview questions, such as questions pertaining to residential stability or past drug use. The set of features may be restricted to include only those that are theoretically motivated or it may be entirely unrestricted, incorporating all available information. Which features are included in the data is decided by humans when determining the scope and type of information



Raw data comes in many forms, including paper records, and digitized formats such as spreadsheets and databases.

Preparing the data is a manual process that involves tasks such as standardizing formats, correcting errors, and constructing new fields.

The result of data processing is the training data, which is digitized data that is formatted such that it can be processed by the training algorithm.

The training algorithm is code that instructs the computer how to learn a model from the training data.

The model learned by the training algorithm is the risk assessment tool, which takes inputs like those in the training data and outputs risk scores or categories.

Ongoing validation verifies that the risk assessment tool continues to perform as expected and that its use is leading to better outcomes.

Is the data measured accurately and without bias? What *exactly* is the model predicting?

Does the process for obtaining the inputs for future assessments respect the rights and dignity of the accused?

allowable for use in the prediction task. Typically the outcome(s) to be predicted are measures of pre-trial failure, such as failure to appear or re-arrest.

Even when the data is restricted to this somewhat limited scope, the data may have problematic facets. For example, one serious concern is *non-representativeness*—that is, a mismatch between the population represented in the training data and the population to which the model is applied. Such a mismatch can occur if there are demographic, cultural, institutional, or policy differences between the two populations that impact the relative importance of the risk factors in predicting the outcome of interest. Even a "validated" model will see a degradation in performance if applied to a population that is not well represented by the training data. For example, models built from data collected prior to the implementation of new programs to reduce pre-trial failure are likely—if the programs are successful—to over-state the risk of pre-trial failure [11].

Another concern with training data is *measurement bias*: the systematic over- or under-measurement of the concept of interest. From the point of view of creating a risk assessment tool, this is particularly worrisome if the degree of over- or under-measurement varies based on socially sensitive or legally protected class, such as race or gender. Perhaps the most commonly discussed example of measurement bias in the context of pre-trial risk assessment is the use of re-arrest as a measure of re-offense. Critics argue that due to racially biased policing practices, white people who re-offend are less likely to be arrested than people of color who similarly re-offend, making re-arrest a biased measure of re-offense.

Finally, one non-technical concern has to do with the provenance of the training data and processes for future data collection. For example, pre-trial risk assessment critics have argued that data that is collected by interviews administered without an attorney present jeopardize the rights of defendants to not self-incriminate. While issues of this type have not played so central a role to date, we anticipate that this class of concern will be much more prominent in coming iterations of pre-trial risk assessment.

Developing a risk assessment model

In talking about "algorithms" it is helpful to distinguish between two different algorithms that are involved in risk assessment. The first is typically called the training algorithm. This is the procedure by which the data available for model construction is processed into what becomes the

Does the model development process require manual adjustment after the training algorithm has been run? If so, how and why were each of the adjustments made?

Are there racial, ethnic, gender, or any other relevant disparities in the model's predictions?

second algorithm—the risk assessment tool itself. Whereas the training algorithm is not AI, in the sense that it is an explicitly programmed set of instructions that tell the computer *how* to learn from the data, the resulting tool, by virtue of having been learned from the data, is.

Most actuarial risk assessment tools in use today are developed through *regression modeling*—a training algorithm that identifies the set of "weights" or "points" to assign to each input factor to best predict the target outcome. In many cases, the developers then adjust the weights that are output by the training algorithm through rounding or to remove counter-intuitive aspects of the model, such as when a weight on a criminal history element reduces the individual's risk score. This is essentially the process that was used to produce tools such as the CPAT[1], PTRA[2], VPRAI[3], PSA[4], the Florida Pretrial Risk Assessment Instrument, and some others developed for local use such as the Santa Clara County tool and the supervised release tool in New York City [2, 13]. This process absent the manual adjustment of the weights is a fairly standard example of machine learning.[5]

"Pure machine learning" tools that do not incorporate any such manual adjustment also exist: for example, the COMPAS[6] Women's Pathways Prison Internal Classification used in the PA Department of Corrections [16, Chapter 3] or the model used in Philadelphia to determine the level of supervision for offenders released on probation or parole [5]. In addition to following a pure machine learning pipeline in the training process, these tools are reportedly based on machine learning methods that produce more complex model structures that are not easily reduced to a set of weights.

The resulting risk assessment model

The result of applying the training algorithm to the training data is the model, which is also often referred to as the tool or, somewhat confusingly, the "algorithm." Regardless of nomenclature, this is the set of instructions that will be used to transform new individuals' input factors into risk predictions.

Even models that have been developed using best practices may be controversial. For example, a risk assessment model's predictions may exhibit racial disparities. What types of disparities are acceptable is an ongoing topic of debate, but it is likely that models that result in different average predictions across relevant groups—whether justifiable based on observed patterns of pre-trial failure or not—will continue to be the subject of vocal criticism.

THE FUTURE OF AI AND RISK ASSESSMENT INSTRUMENTS

As pre-trial risk assessment instruments are increasingly adopted by jurisdictions looking to modernize their pre-trial processes, we anticipate that risk assessment tools will begin to incorporate more types of data and more complicated modeling approaches. We can expect to see more tools being developed through a fully algorithmic training process, without reliance on the type of manual post-processing that is common to most existing actuarial tools. While we cannot be sure what is on the horizon, there are some clues. The first lies in the academic literature, where academics are proposing new techniques for risk prediction that have not yet been widely deployed. The second place we turn is to other criminal justice-related predictive models, such as predictive policing. Finally, we also look to ways that AI is being incorporated into other areas where high-stakes decisions are being made.

New and more data sources do not solve fundamental problems with representivity, bias, or ethical acquisition.

To date, pre-trial risk assessment tools have relied on factors calculable from criminal justice data and interviews. However, as more and more data about individuals is collected, stored, and processed, varied and new sources of information may be added to those that are already in use in pre-trial risk assessment.

Have AI-generated model inputs been locally validated and found to be unbiased?

Are surveillance-based inputs derived from unevenly distributed surveillance systems?

Does the tool expand the definition of unacceptable pre-trial behavior or widen the net of those eligible for pre-trial supervision or detention?

Government administrative data is increasingly collected, systematized, and compiled into unified datasets. For example, some localities merge data collected by a wide variety of government bodies—including law enforcement, social services, mental health, and child and family services—to create unified records that encompass information about the same individual across these different government sources.[7] Commercial data, such as data from repossession and collections agencies, social media, foreclosures, pay parking lots, call data from pizza chains, and rebate information is already being considered for use in some predictive policing systems. [7] In particular, the government's intent to use social media as part of algorithmic profiling in the interest of public safety has already been suggested by recent announcements by the Department of Homeland Security.

Expanded use of surveillance technology, such as automated toll passes or GPS monitoring, may provide another detailed source of information. In fact, researchers at Purdue University recently received a large grant from the National Institute of Justice to fund work on incorporating GPS information from wearable devices into risk assessment modeling.[8] Facial recognition software, similar to that being rolled out in schools and other public spaces, could provide similar information about the whereabouts of individuals in the past and on an ongoing basis. Biometrics is an as of yet mostly untapped category of data that will likely provide inputs for future risk assessments. For example, several computer vision-based approaches to measuring facial characteristics or movements have been proposed or are currently commercially available for predicting criminality, job performance, mood, or deception. Further, work is already underway to incorporate brain scan data into risk assessment [3, 9].

How do these new data sources change the existing data issues in pre-trial risk assessment? Larger training datasets are not necessarily more representative. Even a model based on nationally representative data may not well represent how factors combine at a local level to best predict pre-trial outcomes. Thus, even AI tools built from massive datasets require local validation on an ongoing basis.

New data sources will still exhibit measurement bias. For example, if surveillance systems are more densely deployed in areas with higher levels of past reported crime, data collected by the surveillance system may be as much of a signal about where an individual lives as it is a signal about their pre-trial behavior. Just as zip code is often considered a "proxy" for race, to the extent that surveillance systems are disproportionately deployed in minority communities, measures drawn from these systems may also reflect high correlations with protected characteristics. The use of such inputs could then drive racially disparate predictions.

Surveillance-based data raises additional concerns. While current risk assessment models mostly define pre-trial failure as failure to appear or pre-trial re-arrest, surveillance technologies could be used to create a more expansive definition. Image processing software similar to what is already being sold commercially could be used to detect minor violations or non-compliance with conditions of release during the pre-trial period that, without the surveillance technology, would go unrecorded. This expansion would result in a larger proportion of people being labeled as "re-offenders" during the pre-trial period, leading to higher predicted rates of pre-trial failure in the future. This could then be used to justify higher rates of pre-trial supervision, including detention.

Does the data contain any information that was obtained via legally or ethically questionable methods?

Does collecting data to administer the assessment in the future require any morally objectionable or overly invasive procedures?

Are any of the inputs derived from proprietary software? If so, can that software be audited and can those inputs be contested?

Is the model understandable?

Are the gains in predictive accuracy sufficient to offset the loss in interpretability?

Legal and ethical concerns are not ameliorated by using expanded data sets; if anything, they are exacerbated. One example comes from the use of social media as an input to a risk assessment model. This has the potential to run afoul of first amendment guarantees should this information raise recommended levels of pre-trial supervision based on protected speech.

Another, more far-fetched example is whether it is ethical to require an individual to undergo an invasive procedure, such as a brain scan, in order to receive a risk assessment score to possibly improve their chances of release.

Finally, using features for a risk assessment model that are themselves the output of other, potentially proprietary, models raises concerns about the transparency and validity of those features. Many of the issues we have identified with risk assessment models in general apply to subsidiary models used to generate inputs. One common concern pertains to the ability to audit the model or contest its predictions. This ability becomes more convoluted when some of the inputs are themselves the outputs of other proprietary systems.

More complex models are difficult for humans to understand and are unlikely to substantively increase predictive accuracy.

Even at this early stage of adoption we have seen examples of criminal risk assessment tools that rely on more complex model structures than traditional regression approaches. These types of modeling approaches produce tools that, unlike more familiar point-based systems, are complex "black boxes".[9] One advantage of such models is that they are able to capture more complex associations between the model predictors and the target outcome. When such associations exist, these models will have greater predictive accuracy than structurally simpler models. But precisely because they capture more complex associations, their logic may be difficult or impossible to meaningfully understand.

One major concern when considering adopting "black box" tools is that that we may be limited in our ability to understand how the different factors are weighed in the risk assessment calculation. Just because a tool relies on a large number of factors in a potentially complex way does not mean that each of those factors receives significant weight. A recent investigation by Stevenson and Slobogin [17] of the COMPAS tool used in sentencing makes this point. Specifically, the authors note that while COMPAS considers over 100 different factors, over half of the risk score is attributable to a single factor, the offender's age.

The fact is, there is generally a large multitude of models that all have approximately the same accuracy in predicting a given measure of pre-trial failure. Some research suggests that applying complex models to existing data can in fact improve accuracy over certain simpler models [6]. It is certainly true that any model that incorporates unstructured data such as free text, images, video, or audio would need to be complex in order to be effective. However, there is mounting evidence that, absent the inclusion of more complex unstructured input data, simple models that rely on a small set of factors such as age and prior system involvement can perform just as well [15, 4]. As we discuss next, in our view, there are potentially bigger gains to be had by using machine learning to predict different types of outcomes.

NEW DIRECTIONS IN RISK ASSESSMENT

One way in which a new generation of AI-based tools can help promote more just and decarceral pre-trial decision-making is by moving toward more dynamic formulations of the risk assessment task. While our view is that more data and more complex models will likely serve mostly to complicate the already politically contentious area of pre-trial risk assessment tools, a more fruitful potential path forward is to reframe the inferential or predictive objective.

In deciding pretrial release, judicial officers are tasked with imposing the "least restrictive conditions of release that will reasonable ensure a defendant's attendance at court proceedings and protect the community, victims, witnesses or any other person" [1]. While there is no agreed upon understanding of what "least restrictive" means, implicit in this task is the notion that different conditions of release entail different risks of pre-trial failure. Existing risk assessment models fail to capture this, instead producing a single number that reflects the likelihood of failure under

some unspecified release conditions. Machine learning combined with methods from causal inference can improve upon this practice by producing counterfactual models that assess risk under different release conditions, e.g. reminders of court appearance dates and times. This is similar to models used for treatment planning in precision medicine. By characterizing risk under specified conditions, these models shift the objective from simply determining risk towards determining what can be done to least intrusively manage the risk.

We also see significant opportunity for machine learning to improve how existing models are adapted for use with local populations. Some methods from machine learning address the problem of adapting (parts of) models trained on a given population to a new one where the patterns of risk may be different. This can be particularly helpful for smaller jurisdictions that do not have sufficient data to reliably produce risk models for their populations from scratch, but have enough to guide the principled adaptation of models trained on larger data sets.

CONCLUSION

Deciding whether a risk assessment instrument is appropriate for a jurisdiction requires carefully balancing many trade-offs. Predictive accuracy comes at a cost. Lately, much of the conversation on this point has focused on the trade-off between predictive accuracy and racial equity. Yet there are other important considerations. Similar trade-offs exist between accuracy and model interpretability. Current generation risk assessment models tend towards more easily understandable model types like point systems, which allow the human decision-maker some insight into how the recommendation is being made. Future AI models may not be so easily understandable. In order to realize the additional predictive accuracy that may come with less understandable models, future risk assessment models will likely need to rely on an expanded set of data sources. The legal and ethical complications of collecting these data also represent a real cost.

We must also bear in mind that there may be severe limits to the predictability of human behavior. Even if we were to rely on all of the data that it is hypothetically possible to collect, we may not be able to significantly improve upon the predictive accuracy of existing tools. To the extent that gains are to be had from the use of AI-based risk assessment tools, we anticipate those gains would come from pursuing new directions in how the risk assessment task is initially framed.

As the tools become more complex, it will become more important to stay anchored to the policy goals of justice and equity that motivate reform efforts. Questions that can be asked at the procurement stage are only part of the story and can only gauge the appropriateness of the model in isolation, not in the real world environment in which it will be used. Follow-up studies that assess how the model impacts pre-trial decision-making, including whether the deployment of the model has been followed by reductions in the pre-trial population and reductions in racial disparities, are vital to understanding whether any model—AI-based or not—is having the intended effects.

ACKNOWLEDGMENTS

We thank Tarak Shah for his work compiling background information for the preparation of this critical issue brief. We thank all the members of the group for thoughtful discussions and feedback.

REFERENCES

[1] Pretrial release. URL https://www.americanbar.org/groups/criminal_justice/ publications/criminal_justice_section_archive/crimjust_standards_ pretrialrelease_blk/.

[2] Tools studies URL https://university.pretrial.org/libraryup/topics/assessment/ assessment-studies.

[3] Eyal Aharoni, Gina M Vincent, Carla L Harenski, Vince D Calhoun, Walter Sinnott-Armstrong, Michael S Gazzaniga, and Kent A Kiehl. Neuroprediction of future rearrest. *Proceedings of the National Academy of Sciences*, 110(15):6223–6228, 2013.

[4] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1): 8753–8830, 2017.

[5] Geoffrey Barnes and Jordan M Hyatt. Classifying adult probationers by forecasting future offending. [6] Richard A Berk and Justin Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, 12:513, 2013.

[7] Sarah Brayne. Big data surveillance: The case of policing. *American Sociological Review*, 82(5): 977–1008, 2017.

[8] SL Desmarais and EM Lowder. Pre-trial risk assessment tools: A primer for judges, prosecutors, and defense attorneys. *MacArthur Foundation Safety and Justice Challenge*, 2019.

[9] Kent A Kiehl, Nathaniel E Anderson, Eyal Aharoni, J Michael Maurer, Keith A Harenski, Vikram Rao, Eric D Claus, Carla Harenski, Mike Koenigs, Jean Decety, et al. Age of gray matters: Neuroprediction of recidivism. *Neuroimage: Clinical*, 19:813–823, 2018.

[10] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2017.

[11] John Logan Koepke and David G Robinson. Danger ahead: Risk assessment and the future of bail reform. *Washington Law Review.*, 93:1725, 2018.

[12] Logan Koepke and David Robinson. Civil rights and pretrial risk assessment instruments: A critical issue brief. *MacArthur Foundation Safety and Justice Challenge*, 2019.

[13] Kristian Lum and Tarak Shah. Measures of fairness for new york city's supervised release risk assessment tool. 2019.

[14] Sarah Picard, Matt Watkins, Michael Rempel, and Ashmini Kerodal. Beyond the algorithm. *Center for Court Innovation report*.

[15] Cynthia Rudin and Berk Ustun. Optimized scoring systems: toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

[16] Jay P Singh, Daryl G Kroner, Zachary Hamilton, Sarah L Desmarais, and J Stephen Wormith. *Handbook of Recidivism Risk/Needs Assessment Tools*. John Wiley & Sons, 2018.

[17] Megan T Stevenson and Christopher Slobogin. Algorithmic risk assessments and the double-edged sword of youth. *Behavioral Sciences & the Law*, 36(5):638–656, 2018.

ENDNOTES

1 Colorado Risk Assessment Tool

2 Pre-Trial Risk Assessment

3 Virginia Pre-Trial Risk Assessment Instrument

4 Public Safety Assessment

5 Indeed, the work of "manual adjustment" could itself be automated through a more sophisticated training algorithm. Rather than taking standard regression as a starting point, one can directly create a training algorithm that produces models with rounded weights, ensures all weighs are positive, preferentially includes certain features over others when both produce similar predictive accuracy, and much more.

6 This COMPAS tool was developed for use in correctional settings, not in pre-trial.

7 See, for example, the Los Angeles Enterprise Master Person Index and the Allegheny County Data Warehouse, https://www.alleghenycountyanalytics.us/index.php/dhs-data-warehouse/

8 https://nij.ojp.gov/funding/awards/2019-75-cx-k001

9 In recent years there has been a lot of work on developing improved training algorithms for the express purpose of constructing highly-predictive easily-interpretable models. These approaches can be used, for instance, to produce optimally predictive point system tools, without the need for the type of manual post-processing that has gone into many of the actuarial tools in use today. A vendor may thus have a valid claim of providing a state-of-the-art machine learning based algorithm, even if the tool itself is a simple point system.

Supported by the John D. and Catherine T. MacArthur Foundation

www.SafetyandJusticeChallenge.org