THE COLUMBIA
SCIENCE &
TECHNOLOGY
LAW REVIEW

VOL. XXI STLR.ORG SPRING 2020

ARTICLE

BAIL OR JAIL? JUDICIAL VERSUS ALGORITHMIC DECISION-MAKING IN THE PRETRIAL SYSTEM

Doaa Abu Elyounes[*]

To date, there are approximately sixty risk assessment tools deployed in the criminal justice system. These tools aim to differentiate between low-, medium-, and high-risk defendants and to increase the likelihood that only those who pose a risk to public safety or who are likely to flee are detained. Proponents of actuarial tools claim that these tools are meant to eliminate human biases and to rationalize the decision-making process by summarizing all relevant information in a more efficient way than can the human brain. Opponents of such tools fear that in the name of science, actuarial tools reinforce human biases, harm defendants' rights, and increase racial disparities in the system. The gap between the two camps has widened in the last few years. Policymakers are torn between the promise of technology to contribute to a more just system and a growing movement that calls for the abolishment of the use of actuarial risk assessment tools in general and the use of machine learning-based tools in particular.

This paper examines the role that technology plays in this debate and examines whether deploying artificial intelligence ("AI") in existing risk assessment tools realizes the fears emphasized by opponents of automation or improves our criminal justice system. It focuses on the pretrial stage and examines in depth the seven most commonly used tools. Five of these tools are based on traditional regression analysis, and two have a machine-learning component. This

[*] **Author** – Harvard Law School; and Berkman Klein Center for Internet and Society at Harvard University (email: dabuelyounes@sjd.law.harvard.edu). I would like to express my gratitude to my supervisor, Professor Yochai Benkler, and to Professor Carol Steiker and Professor Niva Elkin-Koren for their extremely helpful guidance throughout the writing process. Earlier drafts of this work were presented at the AI & Law Conference at Seton Hall Law School, at the Cyber Challenges to Human Rights Conference in the University of Haifa, and at an AI Colloquium at the Berkman Klein Center; I am grateful for the helpful comments provided by the participants at these events.

paper concludes that classifying pretrial risk assessment tools as AI-based tools creates the impression that sophisticated robots are taking over the courts and pushing judges from their jobs, but that impression is far from reality. Despite the hype, there are more similarities than differences between tools based on traditional regression analysis and tools based on machine learning. Robots have a long way to go before they can replace judges, and this paper does not argue for replacement. The long list of policy recommendations discussed in the last chapter highlights the extensive work that needs to be done to ensure that risk assessment tools are both accurate and fair toward all members of society. These recommendations apply regardless of whether machine learning or regression analysis is used. Special attention is paid to assessing how machine learning would impact those recommendations. For example, this paper argues that carefully detailing each of the factors used in the tools and including multiple options to choose from (i.e., not just binary "yes-or-no" questions) will be useful for both regression analysis and machine learning. However, machine learning would likely lead to more personalized and meaningful scoring of criminal defendants because of the ability of machine learning techniques to "zoom in" on the unique details of each individual case.

I. INTRODUCTION

The pretrial phase is intended to be a very short period of time between the determination that there is probable cause to support a criminal charge and the adjudication or dismissal of a case. While there are clear differences between bail standards and procedures across states as well as between state and federal jurisdiction, judges are generally limited in when they can detain individuals before trial. Because of the constitutional presumption of innocence, judges can detain a defendant in the pretrial phase only if there is a high risk that the defendant (1) would fail to attend trial or (2) would commit additional wrongdoing while awaiting trial.[1] However, in practice, around 450,000 presumably innocent people are held in jail awaiting trial on any given day in the United States, the world leader in the number of pretrial detainees.[2] The majority of those defendants will eventually be charged with low-level nonviolent offenses. Poor defendants of color, including Black people, Latinos, and Native Americans, are twice as likely to be detained pretrial, and they are usually assigned higher bail amounts than are White defendants arrested on similar charges.[3] This creates an emotional and physical burden on these defendants and pushes them even further below the poverty line. Defendants are commonly fired from their jobs if they cannot secure and deposit bail quickly.[4] Single parents risk losing their children or shuttling them between relatives until bail is placed. Further, in several cases, defendants have committed suicide in their cells while waiting for family members to secure bail.[5]

To date, the consensus among the criminal justice community is that major reform in pretrial adjudication is needed, that far too many defendants are awaiting trial behind bars, and that racial disparities permeate the criminal justice system.[6] Suggested pretrial reforms often include two major changes: (1) abolishing money bail practices and (2) using actuarial risk assessment tools to

[1] Megan Stevenson & Sandra G. Mayson, *Pretrial Detention and Bail*, *in* 3 REFORMING CRIMINAL JUSTICE: PRETRIAL AND TRIAL PROCESSES 21, 22–23 (Erik Luna ed., 2017).

[2] World Prison Brief, *United States of America*, PRISON STUDIES, http://www.prisonstudies.org/country/united-states-america (last visited Nov. 22, 2017).

[3] Cherise F. Burdeen, *The Dangerous Domino Effect of Not Making Bail*, ATLANTIC (Apr. 12, 2016), https://www.theatlantic.com/politics/archive/2016/04/the-dangerous-domino-effect-of-not-making-bail/477906/.

[4] Lorelei Laird, *Bail's Failings*, 102 A.B.A. J. 54, 55–56 (2016).

[5] Joshua J. Luna, *Bail Reform in Colorado: A Presumption of Release*, 88 U. COLO. L. REV. 1067, 1069 (2017); Laird, *supra* note 5, at 55–56.

[6] Wendy R. Calaway & Jennifer M. Kinsley, *Rethinking Bail Reform*, 52 U. RICH. L. REV. 795, 796–97 (2018).

more accurately assess the risk that a defendant will fail to appear or commit another crime. This paper focuses on the second change.

To differentiate between low-, medium-, and high-risk defendants and to increase the likelihood that only those who pose a risk to public safety or who are likely to flee are detained, an increasing number of jurisdictions use actuarial tools. While statistical risk assessment tools have been used in criminal justice for more than 80 years, the complexity and sophistication of these tools are vastly increasing due to rapid developments in computational capabilities and new methods for data analysis.[7] Proponents of actuarial tools claim that these tools are meant to eliminate human biases and rationalize the decision-making process by summarizing all relevant information more efficiently than can the human brain.[8] Opponents of such tools fear that actuarial tools will reinforce human biases, harm defendants' rights by skewing decision-making in an opaque, inaccurate, and unfair way, and that they will increase racial disparities in the system.[9] The gap between the two camps has widened in the last few years, and there has been a growing movement that calls for the abolishment of the use of actuarial risk assessment tools in general and the use of machine learning-based tools in particular.[10] Articles with headlines such as "AI Is Sending People to Jail—And Getting It Wrong"[11] or "Courts Are Using AI to Sentence Criminals. That Must Stop Now"[12] sensationalize the issue without providing a reasoned description of the nature of these tools and their capabilities. In other words, the articles create the impression that the technology itself is the major factor that is leading to racial disparities and unfair treatment of minorities.

Can deploying AI in existing risk assessment tools improve our criminal justice system without realizing the fears emphasized in the media? To address this question, this paper examines in depth the seven most commonly used risk assessment tools in the United States. These tools are currently deployed widely in six states and in

[7] Chelsea Barabas et al., *Interventions Over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment*, 81 PROCS. MACHINE LEARNING RES. 1, 1 (2018).

[8] ANGELE CHRISTIN ET AL., *Courts and Predictive Algorithms*, PRIMER FOR THE DATA & CIVIL RIGHTS, A NEW ERA OF POLICING AND JUSTICE 1–2 (2015).

[9] DAVID G. ROBINSON & LOGAN KOEPKE, CIVIL RIGHTS AND PRETRIAL RISK ASSESSMENT INSTRUMENTS 1–2 (2019), http:// www. safetyandjusticechallenge.org/wp-content/uploads/2019/12/Robinson-Koepke-Civil-Rights-Critical-Issue-Brief.pdf.

[10] *Id.*

[11] Karen Hao, *AI is Sending People to Jail—And Getting it Wrong*, MIT TECH. REV. (Jan. 21, 2019), https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/.

[12] Jason Tashea, *Courts are Using AI to Sentence Criminals. That Must Stop Now.*, WIRED (Apr. 17, 2017), https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/.

more than 100 U.S. jurisdictions, encompassing communities ranging from large and mid-sized cities to isolated rural areas.[13] Five of these tools are based on traditional regression analysis, and two have a machine learning component. This paper concludes that characterizing all existing tools as opaque AI/machine learning-based tools, as many scholars have done,[14] unjustly creates a negative perception of the technology—a perception that does not reflect reality. First, classifying pretrial risk assessment tools as "AI-based tools" creates the impression that sophisticated robots are taking over the courts and pushing judges from their jobs, but this is far from reality. Second, a lack of understanding of the capabilities and limitations of machine learning tools leads to their being unjustly depicted as opaque and inexplicable. While some machine learning techniques could be characterized this way, this is not the case with present-day tools used in criminal justice. Despite the hype, there are more similarities than differences between tools based on traditional regression analysis and tools based on machine learning. The main difference is that machine learning algorithms are capable of learning from millions of observations; they make predictions and learn simultaneously. In contrast, statistical modeling is generally applied to smaller data sets with fewer attributes.[15] Given the current limited capabilities of machine learning, adding a machine learning component to the existing risk assessment tools, if done properly, is unlikely to harm due process or equal protection. On the contrary, AI has the potential to improve decision making in the pretrial process by personalizing the risk score for each defendant and thereby making it more meaningful. As a result, the judge will be able to better understand the risk that each defendant poses and to impose the appropriate conditions to mitigate this risk. The defendant, too, will have a better understanding of the score and could appeal it on more meaningful grounds, if necessary.

This is not to say that the existing risk assessment tools are problem-free or that machine learning tools are perfect. The long list of policy recommendations discussed in the last section of this paper highlights the extensive work that needs to be done to ensure that any risk assessment tool is both accurate and fair toward all members of society. There is no doubt that automation cannot solve the injustice that has been deeply rooted in the system for decades.

[13] *See infra* Table 5.

[14] P'SHIP ON AI, REPORT ON ALGORITHMIC RISK ASSESSMENT TOOLS IN THE U.S. CRIMINAL JUSTICE SYSTEM 6–7 (2019), https://www.partnershiponai. org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/.

[15] Tavish Srivastava, *Difference Between Machine Learning & Statistical Modeling*, ANALYTICS VIDHYA (Jul. 1, 2015), https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/.

But putting the blame predominantly on technology can prevent policy makers from understanding the real problem. Research shows that the most successful pretrial reforms—the ones that led to significant reduction in preventive detention and reduction in racial disparities – disparities—occurred in jurisdictions that implemented a combination of methods: employing an actuarial risk assessment tool, passing laws limiting money bail to only the riskiest offences, and investing in finding effective and functional alternative systems of detention.[16] This paper aims to understand the role that technology plays in much needed reform and to better guide policymakers in navigating the discourse surrounding risk assessment tools.

This paper will proceed as follows. First, the paper will discuss the unique characteristics of the pretrial system in the United States and the legal framework that governs this regime. Next, the paper will describe the evolution of risk assessment in the criminal justice system, from the past to the future. This chapter will compare regression analysis (the traditional statistical method used in the most tools today)[17] to machine learning. The goal of this comparison is to provide readers with no technical background with some elementary knowledge about the differences between the techniques. In the next chapter, the article will address the seven most commonly used risk assessment tools in the pretrial stage. The paper will then discuss the policy considerations that policymakers need to take into account when deciding whether to adopt a risk assessment tool, as well as which tool to implement. The paper will assess the performance of the seven tools with regards to these policy considerations, focusing on the difference between the traditional tools and the machine learning tools.

This paper does not argue that robots can replace judges. Rather, this paper aims to compare the two types of tools and examine their impacts on the criminal justice system in order to ensure their safe deployment. Because the implementation of AI-based risk assessment tools faces many challenges, piloting them during the pretrial stage would be helpful. This is because the pretrial stage has a short duration and simple and clear aims, involves relatively straightforward legal questions, and has easy-to-measure outcomes.[18]

[16] PRETRIAL JUSTICE INST., THE STATE OF PRETRIAL JUSTICE IN AMERICA 6–8 (Nov. 2017), https://www.prisonpolicy.org/scans/pji/the_state_of_pretrial_in_america_pji_2017.pdf.

[17] Arthur Rizer & Caleb Watney, *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just*, 23 TEX. REV. L. & POL. 181, 181 (2018).

[18] *Id.* at 184.

A. Pretrial in the United States

Each year, almost twelve million defendants are admitted into local jails across the country.[19] At any given time, about 450,000 are in those jails only because they are awaiting trial, and the majority of these defendants will eventually be charged with low-level nonviolent offenses.[20] Soon after a person is arrested, the court must determine whether he or she will be unconditionally released pending trial, released subject to a condition or combination of conditions, or held in jail in preventive detention. Money bail is by far the most common condition of release. Some jurisdictions use "bail schedules," a fixed amount that is attached to each offense in the criminal code, while in other jurisdictions, the bail amount is subject to the discretion of the judge.[21] In some cases, defendants are released only if they pay the full bail amount or a certain percentage of it; in other cases, they must pay the amount only if they do not appear in court or if they commit another crime during the pretrial phase.[22] For defendants who cannot afford the money bail set in their case, a bondsman may post a surety bond on their behalf and charge them a non-refundable premium in exchange for taking on the risk that they will fail to appear.[23] The for-profit bail bond industry is a two-billion-dollar-per-year industry, and it is often perceived as exploiting minorities and poor communities.[24]

In most jurisdictions, the main considerations judges are allowed to take into account in making pretrial decisions are (1) the risk of failure to appear in court and (2) the risk of endangering public safety.[25] Preventive detention, in which the judge determines that there is no condition or combination of conditions that can adequately address those risks, is supposed to be a rare occurrence. However, in recent years, more judges are demanding bail from

[19] RAM SUBRAMANIAN ET AL., VERA INSTITUTE OF JUSTICE, INCARCERATION'S FRONT DOOR: THE MISUSE OF JAILS IN AMERICA 4 (Feb. 2015), http://www.safetyandjusticechallenge.org/wp-content/uploads/2015/01/incarcerations-front-door-report.pdf.

[20] Burdeen, *supra* note 4.

[21] Rizer & Watney, *supra* note 18, at 183.

[22] LARRY SCHWARTZTOL ET AL., CRIMINAL JUSTICE POLICY PROGRAM AT HARVARD LAW SCHOOL, MOVING BEYOND MONEY: A PRIMER ON BAIL REFORM 5 (2016), http://cjpp.law.harvard.edu/assets/FINAL-Primer-on-Bail-Reform.pdf.

[23] MARC LEVIN, TEX. PUB. POLICY FOUND. CTR. FOR EFFECTIVE JUSTICE, PRETRIAL JUSTICE 101: KEY POINTS FOR POLICYMAKERS 1–2 (2015), https://www.texaspolicy.com/library/doclib/PreTrialJustice-CEJ.pdf.

[24] Stephanie Wykstra, *Bail Reform, Which Could Save Millions of Unconvicted People from Jail, Explained*, VOX (Oct. 17, 2018), https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality.

[25] SCHWARTZTOL ET AL., *supra* note 23, at 4.

defendants awaiting trial, and the amount of bail is increasing.[26] Thus, in reality, many defendants are being held in jail not because of the risk they pose to the community, but because they are poor. More than one-third of pretrial detainees across the country are in jail because they cannot afford to post bail;[27] nationwide, nine of ten felony defendants who were detained pretrial had bail set and would have been released if they had posted it.[28]

In recent years, bail reform movements have advocated for changes to the money bail system to reduce inequities and the rate of pretrial detentions. New Jersey and California have abolished money bail completely, and other states such as Massachusetts, Alaska, Illinois, Connecticut, and Colorado have restricted the types of offenses for which judges can assign money bail or have required judges to consider defendants' financial situation when setting bail fees.[29] Other jurisdictions, such as Kentucky and Washington, D.C., have eliminated the for-profit bail bond industry.[30] Similar initiatives are also happening at the local level. For example, some district attorneys no longer ask for money bail for certain offenses, and grassroots organizations are fighting to eliminate money bail.[31] In addition to reducing the reliance on money bail, local jurisdictions are investing in pretrial services, finding cost-effective ways to increase the rate of court appearances, and making greater use of risk assessment tools.[32] While acknowledging that only a holistic approach that takes into account the different channels of reform will be truly effective, this paper focuses on investing in actuarial risk assessment tools.

1. The Consequences of Awaiting Trial in Jail

Preventive detention has several negative consequences for the defendant and society:

First, detention during pretrial is very expensive. The estimated annual cost to the taxpayer of incarcerating pretrial

[26] Max Ehrenfreund, *How Bail Punishes the Poor for Their Poverty*, WASH. POST (Feb. 13, 2015), https://www.washingtonpost.com/news/wonk/wp/2015/02/13/how-bail-punishes-the-poor-for-their-poverty/.

[27] Kenneth Polite, *Pretrial Justice*, VERA INST. JUST.: THINK JUST. BLOG (Nov. 13, 2015), https://www.vera.org/blog/justice-in-katrinas-wake/pretrial-justice.

[28] Stevenson & Mayson, *supra* note 2, at 22–23.

[29] Wykstra, *supra* note 25; SCHWARTZTOL ET AL., *supra* note 23, at 7.

[30] COLIN DOYLE ET AL., CRIMINAL JUSTICE POLICY PROGRAM AT HARVARD LAW SCH., BAIL REFORM: A GUIDE FOR STATE AND LOCAL POLICYMAKERS 33 (2019), http://cjpp.law.harvard.edu/assets/BailReform_WEB.pdf.

[31] *Id.* at 64, 73.

[32] For a thorough discussion about all needed components of a bail reform, *see id.*

detainees is $13.6 billion.[33] The daily cost in 2017 of detaining a pretrial defendant in a federal facility ranged from a low of $35.41 to a high of $163.35, with an adjusted average daily cost of detention while awaiting trial of $72.67. In stark contrast, the average daily cost of releasing the defendant under the supervision of a federal probation officer was $8.21. In addition, many defendants are held far from the courts in which they appear, which imposes additional transportation costs.[34]

Second, detainees have reduced access to their attorneys, which limits their ability to contribute to the preparation of their cases.[35]

Third, detention exerts high pressure on defendants to agree to a plea bargain, especially for those charged with low-level crimes. Usually, plea bargains give defendants credit for the time spent in jail awaiting conviction, thereby reducing their jail time, and so many defendants take a plea even if they are innocent since it allows them to leave jail faster.[36]

Fourth, the fact that defendants are detained creates negative perceptions in the mind of the court and jury, who may be more likely to convict or sentence the defendants harshly.[37] This consequence is strengthened if the detainee is wears prison jumpsuits to trial or appears cuffed.

Fifth, detention during pretrial is correlated with longer sentences. A study using data from state courts found that defendants who were detained for the entire pretrial period were at least three times more likely to be sentenced to jail or prison, and received significantly longer sentences than defendants who were released at some point pending trial.[38] Another study yielded similar results in the federal system.[39] It is not easy to show a causal

[33] Bernadette Rabuy, *Pretrial Detention Costs $13.6 Billion Each Year*, PRISON POL'Y INITIATIVE (Feb. 7, 2017), https://www.prisonpolicy.org/blog/2017/02/07/pretrial_cost/.

[34] J.C. Oleson et al., *Pretrial Detention Choices and Federal Sentencing*, 78 FED. PROB. 12, 15 (2014).

[35] Samuel R. Wiseman, *Pretrial Detention and the Right to be Monitored*, 123 YALE L.J. 1344, 1356 (2014).

[36] Samuel R. Wiseman, *Fixing Bail*, 84 GEO. WASH. L. REV. 417, 419 (2016).

[37] Timothy P. Cadigan & Christopher T. Lowenkamp, *Implementing Risk Assessment in the Federal Pretrial Services System*, 75 FED. PROB. 30, 32 (2011).

[38] Meghan Sacks & Alissa R. Ackerman, *Bail and Sentencing: Does Pretrial Detention Lead to Harsher Punishment?*, 25 CRIM. JUST. POL'Y REV. 59, 77 (2014).

[39] The study analyzed 1,798 cases drawn from two federal districts (the District of New Jersey and the Eastern District of Pennsylvania) and, after controlling for a number of variables, found that being detained before trial is associated with increased sentence length, Oleson et al., *supra* note 35, at 14; *see also* Brian P. Schaefer & Tom Hughes, *Examining Judicial Pretrial Release Decisions: The Influence of Risk Assessments and Race*, 20 CRIMINOLOGY, CRIM. JUST., L. & SOC'Y 47, 47–48 (2019).

relationship between preventive detention and longer sentences. Defendants who received longer sentences would probably be ranked as high-risk due to dangerousness. This is, at least in part, the reason why they were not released in the pretrial stage. However, given the high correlation between longer sentences and preventive detention, it is very important to ensure that only high-risk defendants will be detained to avoid any potential impact on the length of the sentence because of pretrial detention.[40]

Sixth, detention during pretrial is highly predictive of reoffending in the future, even when detention is relatively short. Defendants who were detained for a longer time were more likely to commit additional crimes after release.[41] Yet again, these findings drawing a causal inference between reoffending and preventive detention should be taken with a grain of salt. If only high-risk defendants are detained pretrial, and those defendants end up committing additional crimes after getting released, this only justifies the initial decision of the court to keep them in preventive detention. However, reoffending could have multiple causes, including the defendant's exposure to the vicious cycle of criminality in jail during preventive detention.

Given the harmful consequences of preventive detention, it is important to ensure that defendants, in the early stages of their trials while they still enjoy the presumption of innocence under the law, are detained only in rare exceptions and not as a rule.[42]

2. The Legal Framework

Given the decentralization that characterizes criminal law enforcement in the United States, the legal framework governing the pretrial phase is very broad and varies from one jurisdiction to another.

a. Constitutional Protections

Various constitutional principles provide protections for pretrial defendants.

The constitutional presumption of innocence dictates that a formal charge against a person is not evidence of guilt and that the government has the burden of proving the person guilty beyond a reasonable doubt.[43] Therefore, any restriction imposed on the

[40] Oleson et al., *supra* note 35, at 14.

[41] LAURA & JOHN ARNOLD FOUND., PRETRIAL CRIMINAL JUSTICE RESEARCH 2 (2013), https://cjcc.doj.wi.gov/sites/default/files/subcommittee/LJAF-Pretrial-CJ-Research-brief_FNL.pdf.

[42] Oleson et. al., *supra* note 35, at 13–14.

[43] Marie VanNostrand & Gena Keebler, *Our Journey Toward Pretrial Justice*, 71 FED. PROB. 20, 21 (2007).

defendant in the pretrial stage is not punishment, but a way to guarantee his or her appearance during the trial where the presumption of innocence will be debated.

The Due Process Clause, anchored in the Fifth and the Fourteen Amendments,[44] provides that the government shall not take a person's life, liberty, or property without due process of law. Due process "comports with the deepest notions of what is fair and right and just."[45] As it relates to restricting a pretrial defendant's liberty, due process requires, at a minimum, that the defendant receive the opportunity for a fair hearing before an impartial judicial officer, that the decision to restrict liberty be supported by evidence, and that the presumption of innocence be honored.[46]

The Eighth Amendment of the Constitution mentions bail explicitly and states, "Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted."[47] Courts have interpreted this requirement as setting an amount that reflects "adequate assurance" that the accused will attend the trial and comply with the sentence.[48]

The Equal Protection Clause, anchored in the Fourteenth Amendment, embodies the principle that persons who are similarly situated ought to be treated alike.[49] The goal of this principle is to ensure equality between individuals and anti-discrimination based on group affiliation like the ability or inability to pay money bail.

In sum, these constitutional principles afford fundamental protections to pretrial defendants.

b. Federal Laws

The Bail Reform Act, enacted in 1966, applies only to defendants in the federal system.[50] It reinforces the principles that the sole purpose of bail is to assure court appearance and that the law favors release pending trial. Additionally, the Act establishes a "presumption of release by the least restrictive conditions" and emphasizes "non-monetary terms of bail."[51] In 1984, an amendment to the Act authorized the federal courts to deny bail to criminal defendants because of the danger they would commit crimes while on bail.[52] For certain severe offenses, the amendment

[44] U.S. CONST. amend. V; U.S. CONST. amend. XIV.

[45] Solesbee v. Balkcom, 339 U.S. 9, 16 (1950).

[46] Marie VanNostrand & Gena Keebler, *Pretrial Risk Assessment in the Federal Court*, 73 FED. PROB. 1, 3 (2009).

[47] U.S. CONST. amend. VIII.

[48] Stack v. Boyle, 342 U.S. 1, 4 (1951).

[49] Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 242 (2015).

[50] Bail Reform Act of 1966, Pub. L. 89-465, 80 Stat. 214.

[51] VanNostrand & Keebler, *supra* note 47.

[52] *Id.* at 4.

states that the default is pretrial detention, and it passes the burden of proof to the defendant to demonstrate otherwise.[53] Further, the Supreme Court upheld the constitutionality of some preventive pretrial detention in its 1987 decision *United States v. Salerno*.[54] The *Salerno* case acknowledged the government's interest in protecting the community; it relies on the limitations set by Congress in the statute and stated that preventive detention should be reserved only for the riskiest of defendants and the most serious offenses.[55]

In the Speedy Trial Act of 1974, Congress established pretrial services agencies in ten judicial districts to reduce the incidence of crime committed by defendants released to the community and to minimize unnecessary pretrial detention.[56] The Act requires the agencies to interview each person charged with any offense other than a minor crime, verify background information, and present a report to the judicial officer considering bail. The agencies are also responsible for supervising defendants released to their custody pending trial and for connecting them with community services.[57] In 1982, pretrial services were extended to every federal judicial district with the enactment of the Pretrial Services Act.[58]

c. State Laws

Most state constitutions include provisions that guarantee a right to bail. A typical right to bail provision states, "All persons shall be bailable by sufficient sureties, unless for capital offenses, where the proof is evident, or the presumption great."[59] In some states, courts interpret the word "shall" broadly; except for defendants charged with serious offenses, defendants are usually granted bail and are detained only if they cannot deposit the money. In other states, courts have more discretion in deciding who will be released and who will be detained, and the right to bail is not automatic. In nine states, the law does not recognize a right to bail, other than that based on the Eighth Amendment.[60] As described earlier, some states have eliminated or limited money bail. Finally, given the increasing

[53] 18 U.S.C. § 3142(e) (2012).

[54] United States v. Salerno, 481 U.S. 739, 755 (1987).

[55] *Id.* at 750-51.

[56] 18 U.S.C. § 3152 (2012).

[57] *Probation and Pretrial Services History*, UNITED STATES COURTS, http://www.uscourts.gov/services-forms/probation-and-pretrial-services/probation-and-pretrial-services-history (last visited Aug. 27, 2019).

[58] Timothy P. Cadigan, *Introduction to Federal Probation's Special Focus on the 30th Anniversary of the Passage of the Pretrial Services Act of 1982*, 76 FED. PROB. 2,2 (2012).

[59] SCHWARTZTOL ET AL., *supra* note 23, at 9.

[60] *Id.*

usage of risk assessment tools, more jurisdictions are anchoring the requirement to adopt such tools in regulation.[61]

II. THE EVOLUTION OF THE PRETRIAL USAGE OF RISK ASSESSMENT TOOLS

Traditionally, judges have relied on their own judgment and experience to assess the risk that each defendant poses. But research shows that when judges rely on their intuition, they do not use information reliably. Instead, judges may assign weight to items that are in fact not predictive, or they may be overly influenced by causal attributions.[62] Although the use of some version of risk assessment tools began in the 1920s,[63] more advanced actuarial pretrial risk assessment tools entered into use in the early 1960s.[64] The Vera Point Scale, considered to be the first actuarial pretrial risk assessment tool, was developed and adopted in New York City in 1961. It classified defendants by the degree of risk they posed; based on this classification, court officers developed a recommendation for release.[65] Since then, the number and sophistication of these algorithms have vastly increased, and today there are about sixty risk assessment tools used across the country.[66] Twenty-four percent of pretrial agencies use tools based on objective factors, mainly criminal history; twelve percent rely on tools that include subjective aspects and are based on interviews with the defendant and data about employment, education, family status, and the like; and sixty-four percent use tools that include a combination of the two.[67]

In many arenas, society is increasingly putting its faith in actuarial instruments, believing that an algorithm can do a much

[61] *See, e.g.*, An Act Relative to Criminal Justice Reform, S.2185, 190th Leg., Reg. Sess. (Mass. 2017).

[62] Stephen D. Gottfredson & Laura J. Moriarty, *Clinical Versus Actuarial Judgments in Criminal Justice Decisions: Should One Replace the Other?*, 70 FED. PROB. 15, 15 (2006) .

[63] Barabas et al., *supra* note 8.

[64] Kristin Bechtel et al., *A Meta-Analytic Review of Pretrial Research: Risk Assessment, Bond Type, and Interventions*, 42 AM. J. CRIM. JUST. 443, 445-46 (2017).

[65] For more information on the Manhattan Bail Project, see CYNTHIA A. MAMALIAN, STATE OF THE SCIENCE OF PRETRIAL RISK ASSESSMENT 4–5 (Pretrial Justice Inst., ed., 2011), https://www.bja.gov/publications/pji_pretrialriskassessment.pdf; *see also* PRETRIAL JUSTICE INST., PRETRIAL RISK ASSESSMENT: SCIENCE PROVIDES GUIDANCE ON ASSESSING DEFENDANTS 3 (May 2015), https://www.ncsc.org/~/media/Microsites/Files/PJCC/Pretrial%20risk%20assessment%20Science%20provides%20guidance%20on%20assessing%20defendants.ashx.

[66] Rizer & Watney, *supra* note 18, at 191.

[67] *Id.* at 192.

better job than can a human brain.[68] Most scholars, criminal justice practitioners, and citizens see actuarial methods as efficient, rational, and wealth-maximizing tools to allocate limited law enforcement resources.[69] However, due to the sensitivity of the criminal justice context and the significant impact that actuarial results have on defendants' lives, the use of actuarial instruments in this area is more controversial. There are contradictory findings regarding the performance and validity of risk assessment tools in different jurisdictions and their predictive accuracy across race and gender.[70] For example, a meta-analysis of twenty-two research papers on the implementation of risk assessment tools found that some papers concluded that after the adoption of a tool, fewer minorities were incarcerated, while another set of papers found more ambiguous results.[71] Other studies show that actuarial tools outperform clinical judgment in predicting defendants' recidivism risk.[72] However, in one study, after a risk assessment algorithm was adopted, "[j]udges from predominantly [W]hite areas liberalized their bail setting practices more than judges from more racially mixed urban areas," resulting in bias against Black defendants.[73] Yet there could be many explanations for that finding, including the interplay with human biases. Judges respond to and interact differently with risk assessment tools. In one jurisdiction, judges can use the tool to "liberalize" their practices, while in another jurisdiction, judges can use it to reinforce their internal biases and/or may deviate from the recommendation presented by the tool.[74]

In addition, as will be explained later in the paper, the ways in which each tool operates vary, and small details about the design of the algorithm could impact its performance. Other controversial findings were reported in regard to the predictive validity of the tools

[68] BERNARD HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 2 (2007).

[69] *Id.*

[70] Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 3, 4 (Jay P. Singh et al. eds., 2018); SARAH L. DESMARAIS & EVAN M. LOWDER, PRETRIAL RISK ASSESSMENT TOOLS: A PRIMER FOR JUDGES, PROSECUTORS, AND DEFENSE ATTORNEYS 6–7 (Feb. 2019), http://www.safetyandjusticechallenge.org/resource/pretrial-risk-assessment-tools-a-primer-for-judges-prosecutors-and-defense-attorneys/.

[71] Nicholas Scurich & Daniel A. Krauss, *Public's Views of Risk Assessment Algorithms and Pretrial Decision Making*, PSYCHOL. PUB. POL'Y & L. (forthcoming 2020) (manuscript at 4–5), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3478886.

[72] *See* Sharad Goel et al., *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment* (Dec. 26, 2018) (unpublished article) (manuscript at 2–4), https://ssrn.com/abstract=3306723.

[73] Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303, 309 (2018).

[74] *Id.*

for defendants from different socio-economic backgrounds. One study found that the risk assessment tool released a greater number of wealthy defendants and increased the likelihood of incarceration for poor defendants.[75] This could lead judges to link blameworthiness and socio-economic status and exacerbate existing biases.[76]

These controversies have led civil society organizations and some researchers to object to the use of risk assessment tools in favor of other types of pretrial reform.[77] Proponents of this view claim that risk assessment tools reproduce and legitimize discriminatory practices, particularly for people of color in the criminal justice system. Thus, according to this view, principles of fairness, equity and accuracy cannot be translated to mathematical formulas before fundamentally changing discriminatory policing and courtroom practices, which would change the base rate of criminal activity.[78] In other words, this approach views all efforts to fix bias within the algorithm as "technical fixes" that miss the broader picture and will not lead to long lasting change.[79] Since each criminal justice agency has a very limited set of resources, the fear is that due to the changing nature of technology, the tools will have to be updated and changed frequently even before their validity is examined. In addition, the attention spent on introducing these tools could come at the expense of more important reforms.[80] One solution that supporters of this approach propose is to use risk assessment tools to understand the underlying drivers of crime in order to try to interrupt the cycles of crime.[81] Another suggested solution is to use predictive tools only for violent felonies; in this group, preventive detention should be saved only for those who were ranked medium- or high-risk by the algorithm. Thus, defendants who were charged with misdemeanors or non-violent felonies, and who would be ranked low-risk by the algorithm, will be released.[82]

[75] Jennifer Skeem et al., *Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants*, L. & HUM. BEHAV. (forthcoming 2020) (manuscript at 16-17), https://www.ssrn.com/abstract=3316266.

[76] *Id*. at 2, 18.

[77] ROBINSON & KOEPKE, *supra* note 10, at 3; Barabas et al., *supra* note 8, at 7–9; Chelsea Barabas, *Beyond Bias: Re-Imagining the Terms of "Ethical AI" in Criminal Law* 2–3, 34–40 (Apr. 25, 2019) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3377921.

[78] Barabas, *Beyond Bias*, *supra* note 78, at 21.

[79] *Id.*

[80] ROBINSON & KOEPKE, *supra* note 10, at 4.

[81] Barabas et al., *supra* note 8, at 7–8.

[82] SARAH PICARD ET AL., CTR. FOR COURT INNOVATION, BEYOND THE ALGORITHM: PRETRIAL REFORM, RISK ASSESSMENT AND RACIAL FAIRNESS 12 (2019), https://www.courtinnovation.org/sites/default/files/media/document/2019/Beyond_The_Algorithm.pdf.

In sum, eliminating racial and gender bias in risk assessment tools is a topic that has been receiving a great deal of attention from all actors in the field. Developers of risk assessment tools and jurisdictions that have implemented these tools are pushing hard to conduct validation studies and to implement changes in the algorithm accordingly. There is still a long way to go, but technology can help us to achieve a fairer result.

A. The Difference Between Regression Analysis and Machine Learning

Although both machine learning and statistical regressions have the same objective—to learn from data—each uses a different methodology. The main differences between the two approaches are the volume of data and the extent of human involvement in building a model. Machine learning algorithms are capable of learning from millions of observations; they make predictions and learn simultaneously. In contrast, statistical modeling is generally applied to smaller data sets with fewer attributes.[83] Machine learning usually does not rely on rules, whereas statistical regressions formalize relationships between variables in the form of mathematical equations.[84] In addition, machine learning "refers to the capacity of a system to improve its performance at a task over time. Often this task involves recognizing patterns in datasets, although [machine learning] outputs can include everything from translating languages and diagnosing precancerous moles to grasping objects or helping to drive a car."[85] Enhanced computing capabilities and huge amounts of data have increased the capabilities of all data processing techniques.[86]

Consider the following example. A newly appointed judge needs to decide whether to release John until the end of the trial or to keep him in preventive detention. John is a 28-year-old Black single father of two children who works in a factory. He is accused of raping a 19-year-old woman whom he met at a party. He has two prior convictions, one for sexual assault and the other for possessing drugs for personal use, and has one failure to appear in court. He has a tenth-grade education, a tattoo on his shoulder, a history of alcoholism, and he receives food stamps and other welfare benefits. In the short pretrial hearing session, John denies the accusation and any acquaintance with the woman.

[83] Srivastava, supra note 16.

[84] *Id.*

[85] Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 405 (2017).

[86] Randy Bean, *How Big Data is Empowering AI and Machine Learning at Scale*, MIT SLOAN MGMT. REV. (May 8, 2017), https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/.

The judge is conflicted and cannot decide on an outcome. To reach an informed decision, the judge has three options. First, she can try on her own, based on her prior knowledge and skills, to look for prior similar cases from her jurisdiction and to see how they were treated. She would likely spend hours searching and debating which facts of the case were important for comparison and which she could omit. Should she decide based on marital status, age group, education level, priors, and/or race? The options are endless.[87] Second, the judge can consult a regression-based risk assessment tool. This tool has a pre-specified list of factors, selected by either the company that created the tool or a group of experts from her state/county. These factors reflect certain human hypotheses about what will or will not be predictive and uses these factors to predict risk. Factors can include, for example, age, gender, number of priors, number of children, etc. Third, the judge can consult a machine learning-based tool. This tool might include a longer list of factors, or the way it operates would reveal correlations and patterns that the judge was not expecting or was not able to see on her own. In any case, this tool will not be restricted to a set of defined rules but rather will be given a training set of cases with various outcomes. The algorithm could for example find that having a tattoo is a good predictor, or that the genders of the children of the defendant matter. If these factors improve the prediction, the machine learning algorithm will include them, something that is impossible for a standard regression tool to do. Hence, a machine learning approach to the judge's dilemma takes into account the complex relationship between the predictive variables and the outcome.

B. Types of Machine Learning Algorithms

Machine learning is a broad field that encompasses many methods and approaches. Which algorithm to use depends on many factors, including (1) the amount of the data and its quality, (2) the specific task that the algorithm is to solve—for example, predicting a category, predicting a quantity, or both—and (3) the desired level of explainability of the results.[88] It is beyond the scope of this paper to dive into the technical details of each approach. But to illustrate the capabilities of the various machine learning techniques, this section explains the difference between the three main branches of

[87] *See, e.g.*, Ryan Copus, Machine Learning and the Reliability of Adjudication 4-5 (2017) (unpublished Ph.D. thesis, University of California, Berkeley) (on file with the University of California, Berkeley Electronic Theses and Dissertations), https://escholarship.org/uc/item/7t80m17d.

[88] Paul Blondel, *Which Machine Learning Algorithm to Choose for My Problem?*, SAP CONVERSATIONAL AI (Feb. 2, 2017), https://recast.ai/blog/machine-learning-algorithms/.

machine learning: supervised learning, unsupervised learning, and reinforcement learning.

1. Supervised Learning

This type of algorithm is called a "supervised algorithm" because the learning process is similar to that which takes place in the classroom, where teachers supervise their students and correct them when they make mistakes.[89] In a supervised learning algorithm, the inputs (variables) and the outputs (outcome) are known in advance, and the algorithm applies a learning technique to detect the correlation between them.[90] For example, for the pretrial phase, the input could be the defendant's age, gender, and number of prior convictions, and the outcome could be release or detention. A dataset is used to train the algorithm. For example, if we had information about 1,000 defendants, of whom half appeared at trial and half did not, we would feed the algorithm with information about 900 defendants (450 who appeared and 450 who did not) and let the algorithm figure out why each one appeared or did not. In other words, it would determine what combination of age, gender, and prior convictions increased the likelihood of a failure to appear. Then, we would validate the performance of the algorithm by testing the 100 remaining cases. If the algorithm accurately predicts 89 of the 100 cases, then the algorithm has an 89% rate of accuracy.[91]

One popular method of supervised learning in the context of criminal justice is the random forest algorithm, which is the basis for one of the risk assessment tools discussed below. In the first of two stages, a large number of decision trees are generated. In the second stage, the results are combined from each tree to produce a forecast.[92] Each branch of the tree represents a variable, and according to a rule set up in advance, one can then determine if the data complies with it or not.[93] For instance, consider the following rule: defendants who are unemployed, male, single, and under thirty years of age, with two or more prior convictions and at least one failure to appear, are likely either to fail to appear for trial or to

[89] Jason Brownlee, *Supervised and Unsupervised Machine Learning Algorithms*, MACHINE LEARNING MASTERY (Mar. 16, 2016), https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

[90] *Id.*

[91] *See, e.g.*, Gustavo Machado, *ML Basics: Supervised, Unsupervised, and Reinforcement Learning*, MEDIUM (Oct. 6, 2016), https://medium.com/@machadogj/ml-basics-supervised-unsupervised-and-reinforcement-learning-b18108487c5a.

[92] Richard Berk & Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT'G REP. 222, 223 (2015).

[93] *Id.*

commit a crime during the pretrial phase. In contrast, males who are older than fifty, have a stable job, are in a long-term relationship, and have up to one prior and no failures to appear are unlikely to fail in pretrial. The different portraits can be visualized as "paths down the branches of a tree."[94] The algorithm will split each variable into two categories using a threshold that maximizes any association with the outcome. For example, one tree splits the age variable into under thirty and over thirty, placing those younger than thirty on the right side of the tree and those older than thirty on the left. In a second tree, the split is between those younger than fifty and those older. In a similar way, many trees can be created to represent different variables, and in each one, the split of the variable is at a different point.[95] The forest consists of many classification trees that, when combined, arrive at a forecast. If the majority of the trees label a given defendant as high risk, then he or she will be labeled as a high risk by the algorithm.

2. Unsupervised Learning

Unsupervised algorithms focus on the potential of the data and what we can learn from it. Under this model, we assume that there has to be some kind of relationship or correlation between the data we have, but that these data are too voluminous and complex for us to analyze using statistical techniques. Thus, the role of the algorithm is to model the underlying structure or distribution in the data to discover the relationship between the data and the outcome.[96] To apply this technique, we feed the algorithm as much information as possible about defendants, but do not define in advance the factors that we think correlate best with failure in the pretrial phase. Instead, we let the algorithm work its magic and figure out on its own what makes a given defendant high-, medium-, or low-risk to not appear in court and to commit another crime. In some instances, the model is able to reveal the structure and point out a number of factors on which it is basing that prediction. One of the special characteristics of interpretable unsupervised learning models is that, although they suggest different ways to categorize or order the data, it is up to us to decide if we want to use a certain factor. For instance, the model may suggest that the combination of race, gender, number of priors, and eye color would produce the most accurate result about failure in the pretrial stage. Obviously, eye color would not be a relevant variable, and it need not be included. It is thus up to policymakers to determine which factors to use.[97] This will be possible only so long as the model is interpretable,

[94] *Id.* at 225–26.

[95] *Id.*

[96] Brownlee, *supra* note 90.

[97] Machado, *supra* note 92.

meaning not too technically complex. Some unsupervised algorithms operate like a black box: they produce a score from a combination of hundreds of factors without the ability to tell how the score was generated. This particular type of unsupervised algorithm is probably not suitable for criminal justice, due to the high level of obscurity in such a structure and the potential clash with due process since defendants will have hard time appealing the generated score.

To illustrate the difference between supervised and unsupervised machine learning algorithms, let us consider a simpler example of movie reviews. To apply supervised learning, we would collect reviews from various websites, and an individual would have to label those words in the review that are associated with positive reviews and those associated with negative reviews. Based on this labeling, the algorithm would develop a vocabulary list, split into positive and negative phrases. When the training process is done, the performance of the algorithm is evaluated on the validation dataset.[98] If we were to use unsupervised learning, we would only feed the algorithm information about the classification of the review as positive or negative and let the algorithm figure out on its own the words and phrases that make each review positive or negative.[99]

3. Reinforcement Learning

This learning method is not yet commonly used in policymaking, but its predictive capacities are rapidly developing, enabling its potential use in the future. In reinforcement learning, an agent—in this case, an algorithm—learns how to behave based on its interaction with the environment and the positive and negative rewards it receives.[100] The best way to understand how reinforcement learning algorithms operate is to use an example from the world of video games, the field where they are used most commonly.[101] Imagine that the goal of the agent is to collect as many apples as possible while avoiding being eaten by a snake on its way to the top of the screen. In the beginning of the game it may be easier to collect apples, because they are more plentiful and predictable. But the value of the apples near the bottom might be less than those that are closer to the top, whose collection also entails more danger

[98] Pimwadee Chaovalit & Lina Zhou, *Movie Review Mining: A Comparison Between Supervised and Unsupervised Classification Approaches*, 38 PROCS. HAW. INT'L. CONF. ON SYS. SCI. 1, 3 (2005).

[99] *Id.*

[100] Thomas Simonini, *An Introduction to Reinforcement Learning*, FREE CODE CAMP (Mar. 31, 2018), https://www.freecodecamp.org/news/an-introduction-to-reinforcement-learning-4339519de419/.

[101] Kun Shao et al,. *A Survey of Deep Reinforcement Learning in Video Games* (Dec. 26, 2019) (manuscript at 1), https://arxiv.org/abs/1912.10944.

because it requires engaging in intense fighting with the snake. The goal of the agent is to continuously interact with the environment and develop a strategy that will give it as many points as possible and guarantee a quick and safe path to the top to collect the most valuable apples.[102] One way the agent can be taught how to move in the game is by describing to it every possible scenario and explaining what to do in each one. However, because it is not easy to foresee all possible scenarios in advance and further complications might arise in each scenario, another option is to provide a general framework, let the agent to take an action, and based on the reward that they will receive, they will be able to calculate the value of each additional move.[103]

In the real world, it has been suggested that reinforcement learning can be used to control traffic lights based on traffic flow, with the goal being a reduction of time spent in traffic. Reinforcement learning has been tested for websites making news recommendations, where it could be particularly useful because users may get bored quickly and news changes rapidly.[104] This approach, however, may not be suitable for criminal justice because: (1) it is less effective than other methods in conducting a well-defined task; (2) it requires a huge amount of data for its training; and (3) it is more effective in generating a general solution that could match to all types of problems, whereas specific solutions are essential in the criminal justice domain.[105]

C. *Unique Characteristics of Machine Learning Relevant for the Pretrial Phase*

It is important to note that machine learning models are more appropriate for the pretrial phase than for other stages of the criminal justice system. The pretrial stage is unique because pretrial risk measures only two specific behaviors: court appearance and rearrest between initial arrest and the end of the trial. Because the aim of the model is very specific, the result is more accurate.[106] In contrast, determining the appropriateness of sentences requires predicting a complex set of factors, such as long-term recidivism,

[102] Simonini, *supra* note 101.

[103] Justin Gage, *Introduction to Reinforcement Learning*, ALGORITHMIA (May 14, 2018), https://blog.algorithmia.com/introduction-to-reinforcement-learning/.

[104] garychl, *Applications of Reinforcement Learning in Real World*, TOWARDS DATA SCI. (Aug. 1, 2018), https://towardsdatascience.com/applications-of-reinforcement-learning-in-real-world-1a94955bcd12.

[105] Simonini, *supra* note 101.

[106] Spike Bradford, *(Evidence-Based, Actuarial Pretrial) Risk Assessment*, U. PRETRIAL. (Aug. 9, 2017), https://university.pretrial.org/blogs/spike-bradford/2017/08/09/evidence-based-actuarial-pretrial-risk-assessment.

with many more subcategories and factors. Machine learning algorithms will thus be less useful.[107]

1. Correlation Does Not Imply Causation

Machine learning requires no prior assumptions about the underlying relationships between the variables. It focuses on processing the data, discovering patterns, and identifying the factors with the highest correlation with the outcome, so causality is not part of the equation. A correlation quantifies the statistical relationship between two data values but, unlike causation, does not imply that one factor causes the other.[108] With correlations, there is no certainty, only probability. But if a correlation is strong, the likelihood of a causal link is high, and vice versa. "Correlations let us analyze a phenomenon not by shedding light on its inner workings but by identifying a useful proxy for it."[109] Correlations have the advantage of letting the data speak for itself, which is particularly useful in an era where computing capabilities and the amount of data have both increased significantly. They can help unveil connections between factors we have not thought about before, which could encourage new research.[110] In contrast, statistical regression builds on causal relationships between the variables and the outcome.[111] However, relying on causality might encourage researchers to overestimate the effect of certain variables on the outcome. It might lead researchers to ignore relationships between variables or the impact of an external variable that was not considered.[112]

In addition, models that are based on regression analysis depend on relatively few predictors that have strong associations with the outcome. Predictors with weak associations with the outcome are usually considered "noise" and discarded. Yet using machine learning enables the inclusion of those predictors with weak associations, which, in the aggregate, can dramatically improve forecasting accuracy. Each predictor may not matter much on its own, but when combined, they can have a material impact. Let us return to the earlier example of the judge deciding on the disposition of the defendant John in the pretrial phase. By using a machine

[107] *Compare, e.g.*, *id.*

[108] VICTOR MAYER-SCHONBERGER & KENNETH CUKIER, BIG DATA, A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 53 (2013).

[109] *Id.*

[110] Berk & Hyatt, *supra* note 93, at 223.

[111] Aatash Shah, *Machine Learning Vs. Statistics*, EDVANCER EDUVENTURES (Aug. 1, 2016), https://www.edvancer.in/machine-learning-vs-statistics/.

[112] Hal R. Varian, *Causal Inference in Economics and Marketing*, 113 PROCS. NAT'L ACAD. SCI. U.S. 7310, 7311 (2016), https://www.pnas.org/content/pnas/113/27/7310.full.pdf.

learning-based risk assessment tool, the judge can avoid choosing between including factors like the tattoo or the level of education. Even if each single factor has only a small impact on the outcome, it is possible that in the aggregate the result provided will be much more accurate because all the factors will have been included.

Despite the benefits of relying on correlations, giving up on causality poses challenges. First, even if many factors are taken into account, there is a risk that a few factors will be weighted most heavily due to the complexity of the algorithm. However, there are suggested technical solutions for this problem that use optimization methods.[113] Second, when factors that are weakly associated with the outcome are also included, there is a chance that factors that are not suitable to the domain are making their way into the prediction. In the criminal justice and pretrial context, when a defendant's personal liberty is at stake, one should expect a high level of certainty with regards to the factors included in the algorithm: the decision to jail someone should not be based on incidental factors like eye color, shoe size, and the like. As long as the factors that the algorithm claims are correlated with the outcome are transparent and open for debate among experts as well as communities whose safety and liberty will be affected by the algorithm, defendants' rights will still be safeguarded.

2. Avoiding Overfitting

The problem of overfitting refers to a situation in which the algorithm is too specialized on a given dataset and cannot generalize the prediction to other datasets. In the same way that human researchers can conduct their research in a specific way that would reinforce their hypothesis, machines can fall into the same trap. Overfitting can occur in both traditional regression and advanced machine learning techniques. Overfitting happens when the model "learns" the training dataset too well and is not able to distinguish between the actual data and the noise in the dataset; the model then will be unable to generalize and maintain the same level of accuracy with new data because it is too customized to the training data.[114] Consider, for example, that we split one dataset into 90% training and 10% validation. We also know in this case that the number of priors is the variable with the strongest correlation to the outcome. Imagine that we set the threshold to two or fewer priors or more than two priors. We start training the model and we realize that the

[113] P.K. Douglas et al., *Performance Comparison of Machine Learning Algorithms and Number of Independent Components Used in fMRI Decoding of Belief vs. Disbelief*, 56 NEUROIMAGE 544, 544–45 (2011).

[114] Will Koehrsen, *Overfitting vs. Underfitting: A Conceptual Explanation*, TOWARDS DATA SCI. (Jan. 27, 2018), https://towardsdatascience.com/overfitting-vs-underfitting-a-conceptual-explanation-d94ee20ca7f9.

algorithm predicts correctly only 55% of the cases both in the training and validation datasets. We are not satisfied and know that the algorithm can do better, so we run it more times, letting the algorithm get more specific. Now the model distinguishes between violent priors and nonviolent priors and treats them as two separate variables. It also increases the threshold to three or fewer priors or more than three. The predictive ability of the algorithm has improved significantly on the training data and has reached 95%. However, when we run the model on the validation dataset, the predictive power drops to 60%. This is because the algorithm has learned the training data too well; in order to reach a high predictive rate on the training dataset, it started to pick up spurious correlations like the defendant's eye color or the name of the defendant's partner, correlations that will clearly not hold to any other data outside the training data. Although it is acceptable to try different combinations to train the algorithm, passing a certain threshold would mean that we are tweaking the algorithm too much to produce the result we want, an act that can impact its performance on real data. The overarching goal is to maximize the algorithm's predictive accuracy on new data points, not necessarily on the training data.[115]

Being aware of this possibility and validating the algorithm often can help avoid overfitting. A specific technique to reduce the possibility of overfitting and derive a more accurate estimate of model prediction performance is to use *K*-fold cross-validation. This technique allows use of the whole dataset for training, avoiding the trade-off between maximizing the amount of data for training and keeping a reasonably sized set for validation. It is especially useful when dealing with smaller datasets, because we do not want to "waste" data on validation and want the algorithm to use all the available data for training.[116] The first step is to randomly divide the dataset into *K* subsets and then run the algorithm *K* times. Each time, one of the *K* subsets is used for validation and the rest for training. The method is repeated until all *K* subsets have been used for training, and each piece of data appears at least once in the training set and once in the validation set.[117] To illustrate, if we have data about 1,000 defendants and want to predict the likelihood of failure in the pretrial phase, we divide the dataset into ten subsets of 100 defendants each. First, subsets 1–9 are used for training and subset 10 for validation. Next, subsets 1–8 and 10 are used for training and subset 9 for validation. In the next step, subsets 1–7 and 9–10 are used for training and subset 8 for validation. We repeat the process

[115] Tom Dietterich, *Overfitting and Undercomputing in Machine Learning*, 27 ACM COMPUTING SURVS. 326, 326–27 (1995).

[116] Georgios Drakos, *Cross-Validation*, TOWARDS DATA SCI. (Aug. 16, 2018), https://towardsdatascience.com/cross-validation-70289113a072.

[117] *Id.*

until all subsets have been used for both training and validation.[118] Unlike with the human brain, each time we perform the task we can delete the previous knowledge from the memory of the algorithm, and thus avoid overfitting.

3. Explainability

The strongest criticism targeted against using tools based on machine learning techniques in the criminal justice arena is their lack of explainability. Explainability in this context has two meanings. The first one is understandability. That a black box algorithm may take over a task that used to be performed by a judge does not seem acceptable from a legal perspective. Yet discussions of understandability sometimes assume that human decision makers are themselves interpretable because they can explain their actions. But as described earlier, studies show that judges consciously and unconsciously weigh more than just legal factors when making decisions: they are influenced by unconscious biases, and their intuitions can often be inaccurate.[119]

The second meaning of explainability refers to the technicalities of the algorithm and the ability to determine how it reached a certain result.[120] In some cases, after machine learning techniques produce a risk score on a scale of low to high, even the engineers who built the algorithm cannot explain what combination of factors led to this result, let alone the judges.[121] The required level of explainability in the context of the pretrial phase is higher than in other policy domains because judges need to explain and defend their decisions and defendants should know on what grounds to appeal them. However, the following points are worth keeping in mind.

First, there are no clear guidelines on what explainability in the context of criminal justice actually means, what level of explainability should be expected from the algorithm results, and where the line between transparency and a "black box" should be drawn. Do we want to be able to trace each step that the algorithm took until reaching a final result? Or is it sufficient to have a general idea about the workings of the algorithm?[122]

Second, black box algorithms are usually algorithms based on unsupervised learning. But as explained earlier, other methods

[118] Copus, *supra* note 88, at 8–9.

[119] Jennifer Doleac, *Let Computers Be the Judge*, MEDIUM (Apr. 20, 2017), https://medium.com/@jenniferdoleac/let-computers-be-the-judge-b9730f94f8c8.

[120] Berk & Hyatt, *supra* note 93, at 222–28.

[121] Zachary C. Lipton, *The Mythos of Model Interpretability*, 16 ACM QUEUE MAG. 36, 37 (2018).

[122] *Id.* at 38.

like supervised learning can provide explainable results. Whenever possible, it is assumed that supervised algorithms will be chosen, unless the black box algorithms perform significantly better and all the potential trade-offs are taken into account. As explained later in this paper, none of the risk assessment tools currently in use operate like a black box. Even the most advanced pretrial risk assessment tools use machine learning on a small scale and for a very specific task that does not reduce explainability.

There are technical solutions for the explainability problem, such as introducing strong auditing mechanisms that analyze the fairness and level of bias in the output of the algorithm and not in the process itself. In the pretrial context, an auditing mechanism could analyze the results of an algorithm to determine whether it treated Black and White defendants or men and women differently.[123] Auditing as a technique, however, should complement other ways to enhance explainability and should never be used alone, as "justice delayed is justice denied;" discovering through an auditing process that an algorithm unrightfully detained a defendant undermines the individual justice principle and does not provide any solution to the defendant who was unnecessarily "punished" by the algorithm.

Because the goal of the risk assessment tool is to improve the decision making of judges and not to replace them, it is important to educate judges about the limits of an explanation given by an algorithm. Judges must be able to factor the explanation in their decision and provide an explanation that complies with due process requirements, meaning that the defendant can understand it and appeal it if needed.

4. The Proprietary Nature of the Tools

The opacity and lack of explainability of proprietary algorithms in use today is largely due to contract clauses written by the private companies that sell them, rather than the design or operation of the algorithms themselves. Because of their complexity, risk assessment tools require special technical expertise that exists mainly in the private sector.[124] The contracts between the private companies selling these tools and law enforcement agencies usually include nondisclosure agreements that prevent access to the proprietary code. Companies justify these provisions by claiming that existing intellectual property laws in general and trade secret laws in particular do not adequately protect the code from competitors, or from hackers who might tweak their actions to

[123] Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 189–90 (2017).

[124] Natalie Ram, *Innovating Criminal Justice*, 112 NW. U. L. REV. 659, 710 (2018).

circumvent the technology.[125] Yet the manuals and the operational details of the vast majority of pretrial risk assessment tools in use today are public information; only one such tool is proprietary. Even without access to the code, there are ways for the courts, defense lawyers, and third parties to examine the validity and reliability of the tools.[126] In addition, because law enforcement agencies are the only customers of these tools, the market for them is small. Private companies have the ability to shape the functionality and specifications of the technology to their customers, which indirectly creates regulation by design.[127] By working closely with law enforcement agencies, they gain a sense of the factors considered to be the most relevant and the required means of accountability; these policy preferences thus shape the design of the technology.[128] Law enforcement agencies need to acknowledge the power that they have as the only purchasers of such technology when negotiating contracts with the private companies. Agencies could develop a pre-approval process in which experts in technology examine all the available information about the product; debate its utility, going beyond the marketing spin of the private companies; compare it with other available products or assess the ability of the agency to internally develop such tool on their own; and examine any potential clash with important criminal justice principles.[129]

5. Competing Notions of Fairness

It might seem intuitive that pretrial risk assessment tools need to be fair. But when the assessment is done not by an individual but by an algorithm, the term "fairness" must be redefined, and this is not an easy task. Computer science literature refers to more than twenty different notions of fairness.[130] These notions can be divided into three main categories based on their focus: (1) on the individual, (2) on antidiscrimination based on group affiliation, and (3) on the causal relationship between the factors and the outcome.[131] Because of their varying emphases, these notions tilt the balance between

[125] Elizabeth E. Joh, *The Undue Influence of Surveillance Technology Companies on Policing*, 92 N.Y.U. L. REV. ONLINE 101, 105 (2017).

[126] Ram, *supra* note 125, at 669.

[127] Joh, *supra* note 126, at 112.

[128] *Id.* at 111.

[129] Elizabeth E. Joh, *The New Surveillance Discretion: Automated Suspicion, Big Data and Policing*, 10 HARV. L. & POL'Y REV. 15, 41 (2015).

[130] Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, FAIRWARE '18 PROCS.INT'L WORKSHOP ON SOFTWARE FAIRNESS IEEE/ACM 1, 1–7 (2018); Arvind Narayanan, *FAT* 2018 Translation Tutorial: 21 Definitions of Fairness and Their Politics*, YOUTUBE (Apr. 18, 2018), https://www.youtube.com/watch?v=wqamrPkF5kk.

[131] Doaa Abu-Elyounes, *Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness*, U. ILL. J. L., TECH. & POL'Y. (forthcoming 2020).

accuracy and fairness in different ways.[132] For example, some aim to equalize the type of errors that the algorithm makes by arriving at an equal number of false-positive and false-negative returns.[133] Hence, if the algorithm is correct in 85% of the cases, this approach will ensure that among the remaining 15% not all of those who are wrongly sent to jail are Black and not all the defendants who got released and are rearrested are White. This is an interesting approach, but the challenge with equalizing false positives and false negatives is that society values them differently and that their economic costs are different. Setting the threshold and deciding on an error rate our society is willing to tolerate are not easy tasks.

All decisions in the pretrial phase are balancing acts. In deciding whom to incarcerate and whom to release, we balance public safety, the presumption of innocence, and the right to a fair trial.[134] The consequences of false-positive and false-negative results will vary in severity depending on context, and translating them into a numeric error rate is a complicated technical and policy task that must be undertaken after consulting with many relevant stakeholders.

Another notion of fairness calls for creating different algorithms for different groups based on their protected attributes.[135] Imagine that it has been proven that having different algorithms for Black and White defendants will improve their predictive accuracy. Should we allow that as a society? In the context of criminal justice, the common view is that the use of race in any form is unconstitutional and would violate the equal protection clause of the Fourteenth Amendment.[136] However, researchers show that this prohibition against considering race is practically impossible because all the existing risk assessment tools include factors that serve as proxies for race, such as socio-economic factors, education level or type of employment. Even criminal history, which is the factor with the highest correlation of recidivism and failure to appear in pretrial, is highly associated with race.[137] There are suggestions of including race as a factor in criminal justice algorithms, but thus far, the suggestions focus on including race in the training/calculation

[132] Aditya Krishna Menon & Robert C. Williamson, *The Cost of Fairness in Binary Classification*, 81 PROCS. MACHINE LEARNING RES. 1, 1–2 (2018).

[133] Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, 29 ADVANCES NEURAL INFO. PROCESSING SYS. 3315, 3316–17 (2016).

[134] Adam Crawford, *Governing Through Anti-Social Behavior: Regulatory Challenges to Criminal Justice*, 49 BRIT. J. CRIMINOLOGY 810, 819 (2009).

[135] Cynthia Dwork et al., *Decoupled Classifiers for Group-Fair and Efficient Machine Learning*, 81 PROCS. MACHINE LEARNING RES. 119, 119–21 (2018).

[136] Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework* 7 (Oct. 2, 2019) (working paper) (on file with Harvard Law School).

[137] *Id.* at 13.

stage of the algorithm, therefore eliminating the effect of race in the prediction process.[138]

Accommodating different notions of fairness in the risk assessment algorithm significantly affects the result and can only be done after reaching a consensus on complicated legal and moral questions. From a mathematical perspective, each notion of fairness is statistically valid, meaning that it does improve the outcome given a specific definition of fairness.[139] From the legal perspective, most of the algorithmic notions can be accommodated within the existing set of laws. However, the determination of which definition will lead to the best result will change depending on the laws of each jurisdiction, and priorities that policymakers define.[140]

6. Eliminating the Harmful Impact of Discretion

Most jurisdictions leave considerable room for judicial discretion in the pretrial phase. This discretion allows judges to achieve a more just result by enabling them to base it on their experience and professional judgment.[141] Some judges, particularly elected judges, might impose harsher detention decisions because the public may hold them responsible for a crime committed by someone who was released pretrial. The opposite outcome is less likely; in most cases the judge's reputation will not be affected negatively by detaining low-risk defendants.[142] In addition, other governmental branches know more than judges about broader policy matters like the costs of pretrial detention and how to factor them.[143]

Judges may use their discretionary power to make decisions based on their biases, stereotypes, and prejudices.[144] This could happen unconsciously because the human brain itself is a black box. Psychological research shows that people who discriminate are usually not aware of it, acting from rapid automatic responses that the brain generates before the deliberative mind can intervene.[145]

Thus, any comparison between a judge and an algorithm should take into account that judges make their decisions on a range of unconscious and deliberative factors that are unquantifiable and will remain unknown.[146] When we analyze human decisions, the

[138] *Id.* at 50.

[139] Abu-Elyounes, *supra* note 132, at 67.

[140] *Id.* at 66–67.

[141] Charles H. Koch, Jr., *Judicial Review of Administrative Discretion*, 54 GEO. WASH. L. REV. 469, 475 (1986).

[142] Wiseman, *Fixing Bail*, *supra* note 37, at 422–23.

[143] *Id.* at 422.

[144] Gottfredson & Moriarty, *supra* note 63, at 15–16.

[145] Jon Kleinberg et al., *Discrimination in the Age of Algorithms* 10–11 (Nat'l Bureau Econ. Research, Working Paper No. 25548, 2019).

[146] *Id.* at 5.

focus is not typically on explicit bias because of the difficulty of proving biases among judges, the broad discretion given to them, and the flexibility of the legal language designed to be relevant to all scenarios. Deploying an machine learning-based actuarial risk assessment tool can help filter out some of the harmful effects of discretion, but it has to be done cautiously because the engineers who build the algorithm may themselves have unconscious biases, and shifting the harmful effects of discretion to them is very hard to detect.[147] Hence, the issue of shifting discretion has to be taken into consideration of all actors in the field, mainly the policy makers that decide if a certain algorithm will be adopted and under which condition. If not addressed correctly, even the way information is organized and displayed in the algorithm could reflect discretionary choice.[148]

III. THE MOST COMMONLY USED RISK ASSESSMENT TOOLS IN PRETRIAL

A. *Pretrial Risk Assessment ("PTRA"): The Federal Instrument*

The Administrative Office of the U.S. Courts has created its own actuarial risk assessment tool tailored to the characteristics of federal offenses and to the needs of the defendants.[149] The PTRA was developed in 2009 to assist pretrial services officers, to reduce disparities in the system, and to increase the diversion rate of low-risk defendants to alternative programs for detention.[150] It was modeled based on all pretrial cases processed by federal districts (except the District of Columbia) from 2001–2007, a total of 565,178 cases. The tool includes eleven factors divided into two categories: criminal history and others. The manual gives the estimated predictive value of each factor. For example, for the first factor—pending charges—the manual states: "Defendants who had one or more misdemeanor or felony charges pending at the time of arrest were twenty percent more likely to fail pending trial when compared to defendants who did not have a pending charge."[151] The other factors are prior misdemeanor arrests, prior felony arrests, prior failures to appear, employment status, residence status, substance

[147] Peter André Busch, *Conceptualizing Digital Discretion Acceptance in Public Service Provision: A Policy Maker Perspective*, 2018 PACIS PROC. 3.

[148] Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J. L. & TECH. 1, 20 (2017).

[149] Timothy Cadigan et al., *The Re-Validation of the Federal Pretrial Services Risk Assessment (PTRA)*, 76 FED. PROB. 3, 6 (2012).

[150] PRETRIAL SERVICES RISK ASSESSMENT (PTRA): FREQUENTLY ASKED QUESTIONS, http://www.edwinwall.com/PTRA/Federal%20Pretrial%20Risk%20 Assessment%20Instrument%20FAQ%202010.pdf.

[151] VanNostrand & Keebler, *supra* note 52, at 12.

abuse, primary charge category (felony, misdemeanor, or infraction), primary charge type, alcohol use, and foreign ties.[152] The factors and the estimated predictions were used to create the regression analysis algorithm. Each defendant gets a raw score ranging between 0–15, which places him or her in one of five risk categories:

1. PTRA 1: scores 0–4
2. PTRA 2: scores 5–6
3. PTRA 3: scores 7–8
4. PTRA 4: scores 9–10
5. PTRA 5: scores 11–15[153]

Each risk score is associated with a probability that the defendant will fail to appear, be arrested again, or engage in a technical violation. It is based on information gathered from databases and the defendants themselves. Pretrial officers also include their recommendation for release or detention; if it differs from the outcome of the PTRA, they are instructed to consult with their supervisors.[154]

A validation study conducted a year after PTRA's implementation in two federal districts found that it increased the rate of recommendation for release and the rate of actual releases.[155] A new validation study published in 2019 aimed to revalidate the tool and examine whether it is calibrated on gender and race.[156] This study was based on a much larger dataset than that used to develop the tool: 85,369 defendants with closed cases who received PTRA assessment as part of their trials between 2009 and 2015. It found that the PTRA performs well at predicting pretrial violations. For example, of defendants classified in risk category one, only 5% violated their release by either failing to appear or committing a new crime. This number increased gradually as the risk categories increased, so that in risk category five, 36% had a violation.[157] The tool was less accurate in predicting the risk of committing a violent crime when compared with all types of crimes. In addition, the

[152] *Id.*

[153] Christopher Lowenkamp & Jay Whetzel, *The Development of an Actuarial Risk Assessment Instrument for U.S. Pretrial Services*, 73 FED. PROB. 33, 40 (2009).

[154] Cadigan & Lowenkamp, *supra* note 38, at 32.

[155] Cadigan et al., *supra* note 150, at 6.

[156] Thomas H. Cohen & Christopher T. Lowenkamp, *Revalidation of the Federal PTRA: Testing the PTRA for Predictive Biases*, 46 CRIM. JUST. & BEHAV. 234, 234 (2019).

[157] *Id.* at 245.

AUC-ROC value of the PTRA[158] was found to be between 0.67–0.73, meaning that 67%–73% of the time a randomly selected recidivist scores higher on the risk instrument than a randomly drawn non-recidivist.[159]

In terms of racial disparities, the researchers found that the PTRA has good-to-moderate predictive capacities for both Blacks and Whites.[160] Both rearrests for any offense and rearrests for violent offenses increased incrementally among the two groups, and the overall accuracy rate in predicting rearrests among Black and White defendants was between 0.64 and 0.67.[161] The study also found that the PTRA overestimates the likelihood of Hispanic defendants to be arrested for any offense, but for violent rearrests, the predictions were similar.[162] In examining racial disparities, the researchers focused on rearrests, rather than the failure to appear because they considered the former indicator a more objective outcome measure.[163] However, examining potential disparities in failure to appear is equally important. Lack of access to transportation and inability to miss working days in order to go to court are factors that significantly impact failure to appear, and they are not influenced by bias within the criminal justice system, such as re-arrest, which could be impacted by policing practices. In terms of gender, the PTRA was equally accurate in its predictions for men and women.[164]

B. Public Safety Assessment ("PSA")

The Public Safety Assessment ("PSA") is a pretrial risk assessment tool created in 2013 by the nonprofit Laura and John Arnold Foundation.[165] It was developed to provide judges in the early stages of the criminal justice process with neutral, reliable information about the defendant.[166] The PSA was created using a very large dataset of over 750,000 cases drawn from more than 300

[158] The AUC-ROC value is the probability that a randomly drawn recidivist will score higher on the instrument than a randomly drawn non-recidivist. To learn more about calculating AUC-ROC value, *see* Parul Pandey, *Simplifying the ROC and AUC Metrics*, TOWARDS DATA SCI. (Mar. 3, 2019), https://towardsdatascience.com/understanding-the-roc-and-auc-curves-a05b68550b69.

[159] Cohen & Lowenkamp, *supra* note 157, at 245.

[160] *Id.* at 245–46.

[161] *Id.* at 245.

[162] *Id.* at 249.

[163] *Id.* at 253.

[164] *Id.* at 260.

[165] The foundation's core objective is to "maximize opportunity and minimize injustice"; in particular, in the context of criminal justice, the foundation aims to advance community safety, fairness and racial justice. *About*, ARNOLD VENTURES, https://www.arnoldventures.org/about/ (last visited Apr. 12, 2020).

[166] Laura & John Arnold Found., *PSA Background*, PUB. SAFETY ASSESSMENT (Aug. 20, 2019), https://www.psapretrial.org/about/background (last visited Aug. 20, 2019).

U.S. jurisdictions. It produces two risk scores: one for failure to appear and one for committing a new crime. In addition, the PSA flags defendants at high risk to commit a violent crime.[167] After testing hundreds of factors that could potentially be included in the algorithm, the PSA's developers decided to rely only on those that can be obtained without an interview.

The nine factors that the PSA considers are (1) the person's age at the time of arrest, (2) whether the current offense is for a violent crime, (3) whether the person had a pending charge at the time of the current offense, (4) whether the person has a prior misdemeanor conviction, (5) whether the person has a prior felony conviction, (6) whether the person has prior convictions for violent crimes, (7) whether the person has failed to appear at a pretrial hearing in the last two years, (8) whether the person failed to appear at a pretrial hearing more than two years ago, and (9) whether the person has previously been sentenced to incarceration.[168] The factors, their weights, and the technique's scoring procedures are available to the public on the Laura and John Arnold Foundation's website.

After each one of the factors is weighted, the PSA produces two scores on a scale of 1–6: one for failure to appear and one for a new crime arrest. Defendants also receive a "yes" or "no" flag for whether they are at risk of committing a violent crime.[169] The PSA is designed to be a national risk assessment tool, and to date, it has been adopted by more than thirty-eight jurisdictions, including the states of Arizona, Kentucky, Utah, and New Jersey, and cities like Phoenix, Chicago, and Houston.[170] The PSA is offered for free to jurisdictions that wish to implement it, and the Laura and John Arnold Foundation funds technical support to improve implementation of the tool.

Researchers have conducted validation studies involving more than 650,000 cases in several jurisdictions, and many more are being planned.[171] A 2018 study that examined the validity of the PSA on a dataset from Kentucky found that the overall predictive

[167] *Id.*

[168] LAURA & JOHN ARNOLD FOUND., PUBLIC SAFETY ASSESSMENT: RISK FACTORS AND FORMULA 2 (2013), https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/PSA-Risk-Factors-and-Formula.pdf [hereinafter RISK FACTORS AND FORMULA].

[169] *Id.* at 3-4.

[170] Mathew DeMichele et al., *The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky* 5 (Apr. 25, 2018), https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/3-Predictive-Utility-Study.pdf; Laura & John Arnold Found., *What is the PSA?*, PUB. SAFETY ASSESSMENT, https://www.psapretrial.org/about/intro (last visited Aug. 23, 2019).

[171] LAURA & JOHN ARNOLD FOUND., *Research*, PUB. SAFETY ASSESSMENT, https://www.psapretrial.org/about/research (last visited Aug. 21, 2019).

utility of the PSA is between 0.64 to 0.66, a value that the authors characterize as a "good" level of overall predictive utility relative to other risk assessment tools.[172] In terms of predictive accuracy by race, the PSA was found to be a fair predictor of new crime arrest but there are disparities when it comes to predicting failure to appear and new violent crime arrest. The PSA assigns Black defendants lower risk scores than White defendants who fail to appear.[173] In terms of predictive accuracy across gender, the study did not find an indication of predictive bias for failure to appear or a new crime arrest, but did find some differences when predicting new violent crime; however, the dataset examined was particularly small so the findings need to be considered cautiously.[174]

Another study of judges, prosecutors and public defenders who are using the PSA found that in general, judges are more satisfied with the tool than are prosecutors and public defenders. This is probably because of their ability to diverge from the recommendation of the algorithm and to use it as an assistive tool.[175] Thus, it has been reported that approximately 80% of judges always or often rely of the PSA's recommendation, while the range of reliance on the algorithm among prosecutors and defenders is approximately 40%.[176] Overall, research shows that jurisdictions that implemented the PSA are experiencing decreases in the size of their jail populations without corresponding increases in crime rates.[177]

The Arnold Foundation is currently funding several research institutes, such as the Access to Justice Lab at Harvard University, the MDRC organization, and Research Triangle International, to further examine the validity of the PSA and its impact on actual decision making. One interesting current study, conducted by the Access to Justice Lab, is evaluating the PSA's effectiveness in Dane County, Wisconsin. Each new pretrial case in Dane County is assigned to one of two groups: (1) the treatment group in which the PSA score will be made available to the judge, prosecutor, and defense attorney and (2) the control group in which the PSA score will be unknown. When the study is completed, the differences in

[172] DeMichele et al., *The Public Safety Assessment*, *supra* note 171, at 48. A predictive utility between 0.64 and 0.66 means that "when drawing two random cases from the dataset, one of which had the pretrial outcome and the other did not, between 64 and 66 percent of the time the case with the pretrial outcome would have a higher score than the successful case." *Id.*

[173] *Id.* at 50–51.

[174] *Id.* at 52.

[175] Matthew DeMichele et al., *What do Criminal Justice Professionals Think About Risk Assessment at Pretrial?* 1, 16 (Apr. 30, 2018), https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/4-Criminal-Justice-Professionals.pdf.

[176] *Id.* at 17.

[177] Laura & John Arnold Found., *Research*, *supra* note 172.

decisions produced when judges were exposed to the PSA score or not will be analyzed.[178]

C. Virginia Pretrial Risk Assessment Instrument ("VPRAI")

The Virginia Department of Criminal Justice Services developed the Virginia Pretrial Risk Assessment Instrument ("VPRAI") in 2005 and completed its implementation in all Virginia's pretrial services agencies that same year. The VPRAI determines defendants' risk of failure to appear and risk of rearrest for other crimes and is provided to judges as part of the investigatory report.[179] Since the VPRAI was first implemented, Virginia has professionally maintained and revalidated the tool every few years. The first revalidation study was conducted after two years of statewide use, and its purpose was to examine whether factors that can change over time, such as crime patterns, law enforcement practices, and demographic factors, affected the accuracy of the VPRAI. The examination confirmed the tool's general accuracy and led to minor revisions that were implemented in early 2009.[180] In 2014, a second thorough revalidation study was launched; in addition to examining again the impact of changing factors, it analyzed the race and gender neutrality of the tool. The study confirmed that the VPRAI is statistically significant in predicting failure to appear and new crime arrests. In terms of racial differences, the study found slight disparities between White and Black defendants, such that Black defendants are more likely to be flagged as high risk. Although the rates of failure to appear were relatively equal between men and women, men had a higher rate of new crime arrests (5.8% compared to 4.5% for women).[181]

Several changes were implemented in the VPRAI after the 2014 study. It found that the factor "lived at the same residence for less than one year" was not a statistically significant predictor for Black defendants and women, and thus it was replaced with a new factor that had higher predictive value: "the defendant was on active community supervision at the time of their arrest."[182] In addition,

[178] Access to Justice Lab, *Pretrial Release*, A2J LAB, https://a2jlab.org/pretrial-release/.

[179] VA. DEP'T OF CRIM. JUST. SERVS., VIRGINIA PRETRIAL RISK ASSESSMENT INSTRUMENT (VPRAI) INSTRUCTION MANUAL 1–2 (2003), http://www.pacenterofexcellence.pitt.edu/documents/VPRAI_Manual.pdf [hereinafter VPRAI INSTRUCTION MANUAL] (last visited Apr. 18, 2020).

[180] Marie VanNostrand, *Pretrial Risk Assessment—Perpetuating or Disrupting Racial Bias?*, PRETRIAL JUST. INST. (Dec. 20, 2016), https://university.pretrial.org/viewdocument/pretrial-risk-assessment-perpetuat.

[181] MONA J.E. DANNER ET AL., RACE AND GENDER NEUTRAL PRETRIAL RISK ASSESSMENT, RELEASE RECOMMENDATIONS, AND SUPERVISION: VPRAI AND PRAXIS REVISED 9 (2016).

[182] VanNostrand, *supra* note 181.

the employment factor was also modified. The length of employment was removed, and subcategories such as "primary caregiver, full-time student, or retired" were added. Lastly, the factor "the current charge is a felony" was found to be a good predictor, so it was refined and now includes the subcategories "felony drug, theft or fraud."[183]

The current version of the VPRAI includes the following eight factors: (1) active community criminal justice supervision; (2) charge is a felony drug, theft, or fraud charge; (3) pending charge; (4) criminal history; (5) two or more failures to appear; (6) two or more violent convictions; (7) unemployed at time of arrest, primary caregiver, full-time student, or retired; and (8) history of drug abuse.[184] The VPRAI Instruction Manual was last updated in April 2018. It includes for each factor a clear and comprehensive explanation on how the pretrial officer should determine the answer to each question. For example, for risk factor 3, pending charge, the manual provides the following explanation:

> The defendant has a pending charge(s) when there is an open criminal case that carries the possibility of a period of incarceration, and the pending charge has an offense date that is before the offense date of the current charge. (A charge with a disposition of "deferred" is NOT counted as a pending charge.) EXCEPTION: If the current arrest is solely for a failure to appear, the underlying charge related to the failure to appear does not constitute a pending charge. In addition, if a defendant is arrested, remains incarcerated pending trial, and is served with new warrants, this does not constitute a pending charge. Select "Yes" if the defendant had one or more charges for jailable offenses pending in a criminal or traffic (not civil) court at the time of arrest. Select "No" if the defendant had no pending charge(s) at the time of arrest.[185]

After the factors are weighted, defendants are assigned a score of 1–6, from low to high.[186]

[183] DANNER ET AL., *supra* note 182, at 17.

[184] VIRGINIA DEP'T OF CRIMINAL JUSTICE SERVS., VIRGINIA PRETRIAL RISK ASSESSMENT INSTRUMENT–(VPRAI): INSTRUCTION MANUAL–VERSION 4.3, at 7-10 (Apr. 2, 2018), https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/publications/corrections/virginia-pretrial-risk-assessment-instrument-vprai_0.pdf.

[185] *Id.* at 9.

[186] KENNETH ROSE, PRETRIAL COORDINATOR, VIRGINIA DEP'T OF CRIMINAL JUSTICE SERVS., PRETRIAL SERVICES AGENCIES: RISK-INFORMED

Accompanying the VPRAI is Praxis, a decision grid that helps translate the VPRAI score into the type of release and level of supervision. The VPRAI measures the risk, and Praxis helps manage that risk.[187] The combined process consists of four steps:

1. The VPRAI score should be calculated.

2. After examining all charges, the most serious charge category should be identified.

3. Based on the first two steps, one of the following recommendations should be chosen: release, release with monitoring, release with pretrial supervision levels 1–3, or detain.

4. If one of the charges is for failure to appear, the severity of recommendation should be increased one level.[188]

A revalidation study that analyzed the use of Praxis found that judges released defendants 1.9 times more often than judges who did not use it.[189]

As a result of the good documentation of all the stages of calculating the VPRAI and developing final recommendations, as well as the extensive validation studies analyzing the VPRAI since it was first implemented, the VPRAI has been adopted by counties in more than twelve states and used as a model for other jurisdictions interested in implementing a pretrial risk assessment tool.[190]

D. Colorado Pretrial Assessment Tool ("CPAT")

The Colorado Pretrial Risk Assessment Tool ("CPAT") was developed in 2013 as part of the Colorado Pretrial Reform Act, which required pretrial agencies to "make all reasonable efforts to implement an empirically developed pretrial risk assessment tool

PRETRIAL DECISION MAKING IN THE COMMONWEALTH OF VIRGINIA (Nov. 10, 2016), http://vscc.virginia.gov/Virginia%20Pretrial%20Services%20Presentation-1.pdf.

[187] DANNER ET AL., *supra* note 182, at 1.

[188] *Id.* at 31–32.

[189] KENNETH ROSE, VIRGINIA PRETRIAL RISK ASSESSMENT INSTRUMENT (VPRAI) & PRAXIS OVERVIEW 16 (Jun. 11, 2018), https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/announcements/vpraipraxisoverview6112018.pdf.

[190] *See* VanNostrand, *Pretrial Risk Assessment—Perpetuating or Disrupting Racial Bias?*, *supra* note 181; *see also* STANFORD L. SCH. POL'Y LAB, RISK ASSESSMENT FACTSHEET: VIRGINIA PRETRIAL RISK ASSESSMENT INSTRUMENT 4 (VPRAI) (Jun. 19, 2019), https://www-cdn.law.stanford.edu/wp-content/uploads/2019/06/VPRAI-Factsheet-FINAL-6-20.pdf.

and a structured decision-making design based on the person's charge and the risk assessment score."[191] The goal of the CPAT is to improve pretrial services that are delivered locally.[192] To develop the tool, data were collected from ten counties that represented 81% of Colorado's population and their local services, and factors in tools used by other jurisdictions such as Virginia and New York City were considered.[193] Eventually, 12 items were selected for inclusion in the tool: (1) having a home or cell phone, (2) owning or renting one's residence, (3) contributing to residential payments, (4) past or current problems with alcohol, (5) past or current mental health treatment, (6) age at first arrest, (7) past jail sentence, (8) past prison sentence, (9) having active warrants, (10) having other pending cases, (11) currently on supervision, and (12) history of revoked bond or supervision.[194] Each factor is assigned a number associated with its influence on pretrial misconduct. For example, if having a past or current problem with alcohol increases the risk of pretrial misconduct by 4%, a defendant that does not have a problem with alcohol will get zero points for this factor and a defendant who has problems will get four points. The sum of the total points of all factors ranges from 0–82.[195] The total number of points is associated with a risk score on a scale of 1-4; the lower one's total point value on a scale of 0-82, the higher the final risk score. Typically, only those with scores 3 and 4 will be given cash bonds.[196]

The pretrial officer is required to conduct an interview with the defendant to obtain the information for items 1–8 and to consult available criminal records for items 9–12. In practice, pretrial officers also check the criminal records to verify the defendants' answers regarding the first eight items.[197] In addition, a good amount of discretion is given to the officers when resolving inconsistencies between information in the records and the defendants' answers, and this could lead to bias and subvert the

[191] COLO. REV. STAT. tit. 16, § 4-106(4)(c) (2013).

[192] COLO. ASS'N OF PRETRIAL SERVS., THE COLORADO PRETRIAL ASSESSMENT TOOL (CPAT): ADMINISTRATION, SCORING AND REPORTING MANUAL VERSION 2 1 (Jun. 2015), https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=47e978bb-3945-9591-7a4f-77755959c5f5.

[193] PRETRIAL JUST. INST., THE COLORADO PRETRIAL ASSESSMENT TOOL (CPAT) REVISED REPORT 10 (Oct. 19, 2012), https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=64908e23-bf3e-9379-1a1f-f2d5b9e1702f&forceDialog=0.

[194] COLO. ASS'N OF PRETRIAL SERVS., *supra* note 193, at 3.

[195] Memorandum from Hillary Smith, Senior Researcher at Colo. Legislative Council Staff, to Interested Persons, 13 (Apr. 16, 2012) (on file with the Colorado State Publications Library Digital Repository), http://hermes.cde.state.co.us/drupal/islandora/object/co%3A20977/datastream/OBJ/view.

[196] Luna, *supra* note 6, at 1094.

[197] COLO. ASS'N OF PRETRIAL SERVS., *supra* note 193, at 4.

neutrality of the tool.[198] Additional interviews that would be helpful for determining the score and recommendation can be conducted with the defendant's family members or the victim.

The final reporting to the judge is done in the following format: "[Defendant's name] has a CPAT risk score consistent with other Colorado defendants whose average public safety rate is [##]% and whose average court appearance rate is [##]%."[199] In addition to the score, the pretrial officer includes the recommendation for suitable conditions for release or detention.

Data collected from the city of Denver shows an increase in release without money bail.[200] However, it is difficult to tell if this increase is due to the implementation of the CPAT or to other pretrial reform measures, such as the abolition of felony bond schedules.

Critics of the CPAT have claimed recently that the tool has not been validated properly for all jurisdictions where it was implemented and that it has not eliminated potential racial bias.[201] For example, one factor that can be problematic is home ownership, which is strongly associated with class, socio-economic levels, and race. Data from Denver County show that, in 2015, 63.7% of Whites in Denver owned their homes compared to only 29.1% of Blacks.[202] A defendant who does not own a home has four points added to the final score. The CPAT also considers past or current mental health treatment, which could be considered discrimination according to the Americans with Disabilities Act.[203]

A revalidation study of the CPAT began in January 2018 and is expected to be completed in mid-2020. The revalidation study seems comprehensive, consisting of a survey of officers, focus groups, and observations, as well as a pilot study that compares the performance of the CPAT to alternative tools by randomly assigning cases to both.[204]

[198] *Id.* at 4–5.

[199] *Id.* at 9.

[200] AUBREE COTE, CMTY. CORRS. DIV. CITY AND COUNTY OF DENVER, PRETRIAL SERVS. (Jan. 9, 2018), https://cdpsdocs.state.co.us/ccjj/Committees/PRTF/Handout/2018-01-09_CCJJ-PRTF_DenverPretrialServices.pdf.

[201] *See, e.g.* American Bail Coalition, *Colorado's Pretrial Risk Assessment Tool Violates the Americans with Disabilities Act, Fourteenth Amendment* (Mar. 23, 2019), http://ambailcoalition.org/colorados-pretrial-risk-assessment-tool-violates-the-americans-with-disabilities-act-fourteenth-amendment/.

[202] *Id.*

[203] *Id.*

[204] VICTORIA TERRANOVA & KYLE WARD, COMM'N ON CRIMINAL & JUVENILE JUSTICE, COLORADO PRETRIAL ASSESSMENT TOOL REVISION (Dec. 5, 2017), https://cdpsdocs.state.co.us/ccjj/Committees/PRTF/Handout/2017-12-05_CCJJ-PRTF_CPAT-Revision.pdf

E. Ohio Pretrial Assessment Tool ("PAT")

The Ohio Pretrial Assessment Tool ("PAT") is part of the Ohio Risk Assessment System ("ORAS"), a collection of ten tools that can be used throughout the criminal justice process, starting from pretrial, community supervision, in prison, and in preparation for release. The development of the ORAS began in 2006 as a collaboration between the Ohio Department of Rehabilitation and Correction and the University of Cincinnati Center for Criminal Justice Research.[205] The system was developed to better classify the risk level of defendants, to match defendants with the most useful support mechanisms, to identify criminogenic needs, and to better allocate resources.[206] An additional goal was to promote consistent measurement of risk across Ohio, given that before its development there was large variation between counties.[207] The ORAS was developed based on data from 1,834 cases adjudicated in 29 locations. To map out the factors to be included in the system, semi-structured 26-question interviews were conducted with defendants; in addition, there was a two-page self-reporting instrument that included 96 questions related to criminal history, criminal thinking, employment, education, aggression, and financial stress.[208] In 2011, House Bill 86 was enacted into law, requiring the Ohio Department of Rehabilitation and Correction to adopt a risk assessment tool for statewide implementation. The bill also required criminal justice agencies to develop policies and guidelines regarding data collection, staff training, oversight and data sharing.[209] In addition to Ohio, the ORAS has been implemented in several other states including Indiana, Texas and Massachusetts, and in over thirty local jurisdictions.[210]

The PAT was developed based on 452 cases from seven Ohio counties. The interviews identified more than 100 potential factors that could be included in the tool, and ultimately, seven were selected: (1) age at first arrest, (2) number of failure to appear warrants in the past 24 months, (3) three or more prior jail

[205] Edward J. Latessa et al., *The Creation and Validation of the Ohio Risk Assessment System (ORAS)*, Univ. of Cincinnati Final Report 1, 5 (2009) [hereinafter *ORAS Final Report*].

[206] Edward J. Latessa et al., *The Creation and Validation of the Ohio Risk Assessment System (ORAS)*, 74 FED. PROB. 1, 1 (Jul. 2009) [hereinafter *ORAS Federal Probation*].

[207] *Id.* at 2.

[208] Latessa et al., *ORAS Final Report*, *supra* note 206, at 9.

[209] Edward J. Latessa et al., *The Ohio Risk Assessment System*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 147, 159–160 (Jay P. Singh et al. eds., 2018) [hereinafter *ORAS Handbook of Recidivism*].

[210] *Id.* at 148–49.

incarcerations, (4) employed at the time of arrest, (5) residential stability, (6) illegal drug use during the past six months, and (7) a severe drug use problem.[211] The PAT collects data from the file of the defendant, from a face-to-face interview, and from a self-report questionnaire. It scores defendants on a scale of 0–9: scores 0–2 are considered low risk; 3–5, moderate risk; and 6–9, high risk. There is a web-based system that allows officers to enter the data; when the assessment is completed, the system also informs officers of the main factors that drive the risk level.[212]

In addition to a validation study conducted when the tool was developed in 2009, a two-part study conducted in 2018 reevaluated its validity and reliability. The first part examined inter-rater reliability; that is, the degree to which professionals converged or diverged on the appropriate score for certain defendants. For this purpose, the researchers presented professionals who worked with the tool with four hypothetical cases that they had to score.[213] Using the PAT to score those cases, the agreement rate among professionals about the seven factors was on average 87%. The participants diverged only in regard to two factors—employed at the time of the arrest and severe drug use problem—for which the level of agreement was less than 80%.[214] One possible explanation for this divergence is the discretion given to the pretrial officers to incorporate their impressions from the interviews into the assessment. They are instructed to score 0 for no drug problem and 1 if they do have a drug problem. Regarding employment, officers score 0 for those in full-time employment, 1 for those in part-time employment, and 2 for unemployed defendants.[215] The two factors on which the officers did not agree are very different. Employment is a static factor, whereas severity of drug usage is a dynamic factor that is more prone to interpretation.

Perhaps that study teaches us that any type of factor is open for interpretation and that the ideal extent of discretion and interpretation allowed for pretrial officers is not clear. Relatively high agreement was reported on two cases: in one case 89% of participants scored the defendant as at low risk to reoffend, and in the other case 82% set a level of moderate risk. In contrast, only 70% of the participants scored the third case as low risk, and 78% of participants scored the fourth case as moderate risk.[216] The researchers did not attribute the differences to the professionals'

[211] Latessa et al., *ORAS Final Report*, *supra* note 206, at 49–50, Appendix A.

[212] Latessa et al., *ORAS Handbook of Recidivism*, *supra* note 210, at 151–52.

[213] Edward J. Latessa et al., THE OHIO RISK ASSESSMENT SYSTEM (ORAS): A REVALIDATION & INTER-RATER RELIABILITY STUDY, FINAL REPORT 1, 12–13 (2018).

[214] *Id.* at 29.

[215] Latessa et al., *ORAS Final Report*, *supra* note 206, at 45.

[216] *Id.* at 30.

gender, educational level, or amount of training in the use of the tool.[217]

The second part of the study examined the tool's validity and any potential differences in scores between male and female defendants and between White and non-White defendants. The study found that the majority of defendants (58%) were scored as moderate risk, only 24% as low risk, 19% as high risk, and that defendants of different races and genders were proportionally distributed. However, scores on individual factors varied by race and gender. For one factor—severe drug use problem during the last six months—there was a significant difference between White and non-White defendants, with whites scoring higher.[218] The tool predicted relatively accurately new arrests for White defendants but, for non-White defendants, detected no significant correlation between the levels of risk and the rate of rearrests. Similar findings were reported for new convictions: for White defendants, the reconviction rate increased as the level of risk increased, but the rate of reconviction for low-risk non-White defendants was actually higher than that for moderate-risk non-White defendants.[219] For both men and women, there was a direct correlation between the increased risk level and an increase in new arrests; however, the tool predicted only weakly to moderately new convictions for men and moderately predicted new convictions for women.[220]

F. Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS")

Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS") is an empirical risk and needs assessment tool integrated into the Northpointe Suite, a web-based assessment and case management system for criminal justice practitioners.[221] It was developed by the private company Northpointe, now owned by Equivant. COMPAS is the only risk assessment tool in-use that is based on machine learning.[222] The tool uses, among other techniques, Random Forest and Support Vector

[217] *Id.* at 31–32.

[218] *Id.* at 36–37.

[219] *Id.* at 43–45.

[220] *Id.* at 41–42.

[221] *Practitioner's Guide to COMPAS Core*, EQUIVANT 2 (Apr. 4, 2019), http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf.

[222] Tim Brennan & William Dieterich, *Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 49, 70–72 (Jay P. Singh et al. eds., 2018).

Machines.[223] Equivant is also working on more advanced techniques for improving predictions and methods for handling a large amount of data.[224] COMPAS comprises 43 risk scale models used in different stages of the criminal trials. This section focuses on four models. The General Recidivism Risk Scale and the Violent Recidivism Risk Scale are the primary risk models of COMPAS, and they are widely used. The Pretrial Release Risk (PRRS) is a risk model designed especially for pretrial, and the Recidivism Risk Screen (RRS) is a scale designed to predict new crime arrests in the next two years.[225] These four models were chosen because, despite the existence of a specialized model for the pretrial phase, most jurisdictions use a combination of models in that phase. The motivation for developing so many different risk models was to closely match the circumstances of the case to the appropriate tool so as to yield the most accurate and relevant score.[226] But in practice, as will be described below, a multiplicity of risk models creates confusion about which tool to use in each circumstance, leads to less uniformity among jurisdictions, and, when more than one model is used, may cause factors such as criminal history to be counted twice, which could inaccurately inflate the defendant's score.

Data for COMPAS are provided by a long questionnaire consisting of 137 questions that are either answered by the defendant in an interview or collected from criminal records. It asks defendants whether one of their parents was ever imprisoned, how many of their friends are taking drugs illegally, and how often they get in fights in school. It also asks whether they agree or disagree with statements such as "[a] hungry person has a right to steal" and "[i]f people make me angry or lose my temper, I can be dangerous."[227] According to Northpointe, the majority of the 137 questions are used to determine defendants' needs, and COMPAS risk models include a much shorter list of factors.[228]

[223] Random Forest and Support Vector Machines are machine learning techniques used for classifying data points into different groups. *See supra* Section 2.2.1.

[224] Brennan & Dieterich, *Correctional Offender Management Profiles*, *supra* note 223, at 70–72.

[225] EQUIVANT, *Practitioner's Guide to COMPAS Core*, *supra* note 222, at 2–3.

[226] *Id.* at 1–2.

[227] Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[228] *Official Response to Science Advances,* EQUIVANT (Jan. 18, 2018), https://www.equivant.com/official-response-to-science-advances/.

1. Pretrial Release Risk Score ("PRRS")

The Pretrial Release Risk Score ("PRRS") was developed in 2009 based on a sample of 2,831 felony defendants arrested in Kent County, Michigan.[229] It includes eight factors: (1) felony top charge, (2) pending case, (3) prior failure to appear, (4) prior arrest on bail, (5) prior jail sentence, (6) drug abuse history, (7) employment status, and (8) length of residence.[230] The tool scores defendants on a scale of 1–10, in which scores 1–4 indicate low risk; 5–7, medium risk; and 8–10, high risk. To date, it has been implemented in two counties in California and has been validated only by internal Northpointe studies.[231]

2. Recidivism Risk Screen ("RRS")

The Recidivism Risk Screen ("RRS") is designed to predict the defendant's risk of being arrested for any misdemeanor or felony offense within the next two years. It includes five factors: (1) age, (2) age at first arrest, (3) number of prior arrests, (4) employment status, and (5) prior parole revocations. According to Northpointe, this scale is meant to supplement the general assessment and the pretrial assessment, and its purpose is to flag those defendants who might need a complete and longer assessment. Since its focus is prediction within the next two years, it could be particularly useful in the pretrial phase.[232] Information on where the tool was implemented and whether any revalidation studies have been conducted is not available at this time.

3. The General Recidivism Tool and the Violent Recidivism Tool

These tools are often used simultaneously. The general recidivism tool predicts the risk to be rearrested for both misdemeanor and felony offenses, and it is considered to be more thorough than the RSS because it includes more factors. These factors include prior arrests and prior sentences to jail, prison, and probation; vocational/educational problems; drug history; age at assessment; and age at first arrest. The Violent Recidivism Risk tool

[229] Stanford Law Sch. Policy Lab, RISK ASSESSMENT FACTSHEET: CORRECTIONAL OFFENDER MANAGEMENT PROFILING FOR ALTERNATIVE SANCTIONS (COMPAS) PRETRIAL RELEASE RISK SCALE-II (PRRS-II) 1 (Jun. 6, 2019), https://www-cdn.law.stanford.edu/wp-content/uploads/2019/06/COMPAS-PRRS-II-Factsheet-Final-6.20.pdf.

[230] EQUIVANT, *Practitioner's Guide to COMPAS Core*, *supra* note 222, at 31.

[231] STANFORD LAW SCH. POLICY LAB, *Risk Assessment Factsheet*, *supra* note 230, at 3.

[232] EQUIVANT, *Practitioner's Guide to COMPAS Core*, *supra* note 222, at 32.

includes a similar list of factors, with the addition of a history of violence and a history of noncompliance.[233] Each tool provides a separate score on a scale of 1–10, in which scores 1–4 indicate low risk; 5–7, medium risk; and 8–10, high risk.[234] They have been implemented in counties in Florida, Wisconsin, and California and are also used in the pretrial phase.

4. Challenges to COMPAS's Validity

COMPAS is a product of a for-profit company, and the inner workings of its algorithm and score calculations are not public information. This lack of transparency has raised questions about its validity and whether it is prone to bias. In 2016, the news outlet ProPublica conducted an investigation into 7,000 cases of people arrested in Broward County, Florida.[235] The court system of this county used both the general recidivism risk tool and the violent recidivism risk tool. ProPublica looked at how many defendants were actually charged with new offenses within two years of their release versus how many were predicted to do so, and concluded that COMPAS is biased because the false-positive rate was much higher among Black defendants. The algorithm falsely labeled Black defendants as future criminals nearly twice as often as it did White defendants: 42% of Black defendants who were released from jail and did not commit any future crimes were wrongly labeled as high risk compared to 22% of White defendants.[236]

Northpointe challenged the findings and published the results of its own investigation showing how COMPAS is equally fair to Black and White defendants, claiming that, at each score level, equal percentages of Blacks and Whites were rearrested. For example, among defendants who received a score of 7, 60% of White defendants and 61% of black defendants were rearrested. In addition, Northpointe claimed that COMPAS is fair because, as directed by law, it does not take race into account explicitly. Finally, Northpointe pointed out that, given the different base rate of Black and White defendants, the disparity that ProPublica referred to will always exist regardless of COMPAS.[237]

[233] *Id.* at 31–32.

[234] STANFORD LAW SCH. POLICY LAB, *Risk Assessment Factsheet*, *supra* note 230, at 3.

[235] Angwin et al., *supra* note 228.

[236] Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually not that Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.ef319d030999.

[237] William Dieterich et al., *Northpointe Inc. Res. Dep't, COMPAS Risks Scales: Demonstrating Accuracy Equity and Predictive* 2 (2016),

The ProPublica story attracted the attention of the media and researchers, and it engendered controversy.[238] Some academics supported ProPublica's finding and criticized the use of algorithms in the criminal justice system,[239] whereas others attributed the disparity that ProPublica found to external factors, such as the different base rate among Black and White defendants and statistical errors made by ProPublica.[240]

This controversy centers on differing ways of defining fairness. For ProPublica, fairness means that the algorithm should make the same type of error equally for Black and White defendants. For Northpointe, in contrast, the algorithm was fair because it was calibrated—meaning that for each race category, the same percentage of Black and White defendants recidivated.[241] As explained earlier, the two notions of fairness are diametrically opposed and cannot be satisfied simultaneously.[242]

But fairness is only one challenge that a sophisticated risk assessment tool like COMPAS raises. To deal with the explainability problem and to challenge the assumption that the black box algorithm of COMPAS provides a better result than a transparent algorithm, a group of researchers tried to open that black box by providing an interpretable model. Their model used the same dataset that ProPublica used, and sought to achieve the same level of accuracy as the black box COMPAS algorithm while still providing a set of rules that explained why their model made its decision.[243] They used specialized tools from the fields of discrete optimization and artificial intelligence. Specifically, they introduced a branch-and-bound algorithm, called Certifiably Optimal Rule Lists, that provides (1) the optimal solution, (2) a certificate of optimality, and (3) optionally, a collection of near-optimal solutions and the distance between each solution and the optimal one. They were able to produce certifiably optimal, interpretable rule lists that achieved the same accuracy as black box tools. Using only the

http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[238] Tashea, *supra* note 13; Anthony W. Flores et al., *False Positives, False Negatives and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."* 80 FED. PROB. 38, 38 (2016).

[239] Hao, *supra* note 12.

[240] Corbett-Davies et al., *supra* note 237.

[241] *Id.*

[242] Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, *in* SIGMETRICS '18 ABSTRACTS OF THE 2018 ACM INTERNATIONAL CONFERENCE ON MEASUREMENT AND MODELING OF COMPUTER SYSTEMS 40, 40 (2018); Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 153–63 (2017).

[243] Elaine Angelino et al., *Learning Certifiably Optimal Rule Lists for Categorical Data*, 18 J. MACHINE LEARNING RES. 1, 2 (2018).

current charge, gender, age, and number of priors, the researchers were able to attain the same accuracy level as COMPAS.[244]

In addition to media and academic interest in COMPAS, courts have also examined its validity. The issue was raised in *State v. Loomis*, a case that made its way to the Supreme Court of Wisconsin.[245] The court concluded that using a risk assessment tool in the sentencing phase did not violate the defendant's right to due process, because the output of the algorithm was not the determinative factor in deciding the length of the sentence. The output was one factor among many others, and the judge has the discretion to diverge from it if needed.[246] There was an attempt to challenge this decision in the U.S. Supreme Court, but certiorari was not granted.[247]

G. The Kleinberg et al. Tool: A Machine Learning Tool Suggested by Researchers

Although the tool discussed in this section is not being used yet in practice in any jurisdiction, given its credibility within the academic community and its reliance on machine learning, it is discussed here.

John Kleinberg and his colleagues at Cornell University are studying the use of the gradient-boosted decision tree technique in pretrial risk assessment, because this technique enables a higher degree of interactivity among the variables and yields a score that is more tailored to each defendant.[248] They built an algorithm based on a large dataset of cases heard in New York City from 2008–2013. In New York City, judges make release recommendations based on a six-item checklist developed by a local nonprofit agency. The researchers compared the performance of the algorithm they developed against the performance of the judges using the checklist.[249] The data included these factors: age, current offense, criminal record (including prior failures to appear), and the outcome of the case, which included release, failure to appear, or rearrest while awaiting trial.[250] The algorithm only had three input variables—current offense, priors, and age—and the outcome variable was the likelihood that the defendant would fail to appear. They built a decision tree for each case divided through a sequence

[244] *Id.* at 1–2.

[245] 881 N.W. 2d 749 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[246] Recent Case, State v. Loomis*, 881 N.W.2d 749 (Wis. 2016)*, 130 HARV. L. REV. 1530, 1532–33 (2017).

[247] *Loomis*, 881 N.W. 2d 749.

[248] Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 237 (2017).

[249] *Id.* at 245–51, 260.

[250] *Id.* at 247.

of binary splits. Starting from the bottom of the tree, the first question was if the defendant had ever been arrested before. For each subsequent step, a similar split would be made based on the information gathered in the previous splits.[251]

Kleinberg's tool used both regression analysis and machine learning. Regression analysis identified the factors to be included in the algorithm, and the machine learning aspect enabled the algorithm to be trained. Hence, the researchers let the algorithm come to its own conclusions about whether the defendant will flee or commit another crime, rather than outputting a percentage chance of each event occurring.[252] The algorithm focused on predicting flight risk and not recidivism, because that is the only factor that judges in New York are allowed to consider; however, the researchers obtained qualitatively similar findings from a national dataset.[253]

The results of the study are quite promising. They show that using machine learning, crime can be reduced up to 24.7% with no change in the rate of detention, or the detention rate can be reduced up to 41.9% with no increase in crime rates.[254] Moreover, all categories of crime, including violent crime, showed reductions, and these gains can be achieved while simultaneously reducing racial disparities.[255]

In addition, the researchers concluded that any additional information that judges are exposed to, other than the necessary factors for prediction, act as noise and distract them from reaching a fair decision. They attributed some of the distraction to what they call a "selective labels problem," meaning that judges rely on many factors that are hard to measure, such as mood, or specific features of the case such as the defendant's appearance.[256]

As a result of this study's findings, the Criminal Justice Agency in New York City is considering redesigning its system.[257] Although these results are quite promising, it is important to remember that Kleinberg and colleagues' work is the first attempt to apply such a technique in pretrial. Yet the fact that its findings are reshaping risk assessment in this jurisdiction indicates that future research that will apply similar techniques in other contexts can play a significant role in designing policy in this field.

[251] *Id.* at 252–53.

[252] Jens Ludwig, *Man vs Machine Learning: Criminal Justice in the 21st Century | Jens Ludwig | TEDxPennsylvaniaAvenue*, YOUTUBE (May 5, 2017), https://www.youtube.com/watch?v=iAsOq-tAe_s.

[253] Kleinberg, *supra* note 249, at 239-41.

[254] *Id.* at 241.

[255] *Id.*

[256] *Id.* at 242–43.

[257] *Redesign of CJA's Risk Assessment System Discussed By Panel*, N.Y.C. CRIMINAL JUSTICE AGENCY (Sept. 22, 2017), https://web.archive.org/web/20181122140417/http://www.nycja.org/resources/details.php?id=1388.

IV. COMPARISON OF THE TOOLS

This section describes the policy considerations that jurisdictions should consider when implementing any risk assessment tools. These policy issues derive from a growing literature produced by civil society organizations and academics that attempts to help policy makers pick the right risk assessment tool for them and address some of the risks that those tools entail.[258] This section examines the performance of the seven risk assessment tools discussed earlier, keeping the policy considerations in mind. The goal is to assess differences between the traditional and the more advanced tools in their compliance with those considerations and in their impact on the criminal justice system.

A. Factors Used in Each Tool

In Table 1 below, all the factors from the tools were clustered into categories, for example the category criminal history includes many factors that are paraphrased differently in the tools.

Table 1. Factors Used in Seven Pretrial Risk Assessment Tools

Tool	Criminal History	Current Offense	Socio-Economics	Age	Substance Abuse	Others
Pretrial Risk Assessment (PTRA): the federal tool	✓	✓	✓		✓	foreign ties
Public Safety Assessment (PSA): Arnold Foundation	✓	✓		✓		
Virginia Pretrial Risk Assessment Instrument (VPRAI)	✓	✓	✓		✓	active community supervision

[258] *See* Christopher Bavitz, Berkman Klein Ctr., *An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System*, MEDIUM (Nov. 9, 2017), https://medium.com/berkman-klein-center/the-following-letter-signed-by-harvard-and-mit-based-faculty-staff-and-researchers-chelsea-7a0cf3e925e9; *see also* P'SHIP ON AI, *supra* note 15 at 3; SARAH PICARD-FRITSCHE ET AL., DEMYSTIFYING RISK ASSESSMENT: KEY PRINCIPLES AND CONTROVERSIES 1 (Ctr. For Court Innovation eds., Mar. 2017), https://www.courtinnovation.org/sites/default/files/documents/Monograph_March2017_Demystifying%20Risk%20Assessment_1.pdf.

Colorado Pretrial Risk Assessment Tool (CPAT)	✓	✓	✓	✓	✓	mental health issues
Ohio Pretrial Assessment Tool (PAT)	✓		✓	✓	✓	
COMPAS	✓	✓			✓	length of time in current community
Kleinberg et al. tool	✓	✓		✓		

Each of the seven tools includes between three and twelve factors.[259] The Kleinberg tool has the fewest factors, although "priors" is a broad category, which includes many subfactors. The main focus across tools is on criminal history and its variance; other commonly used factors include age, community ties, residential stability, employment, and substance abuse. During the lengthy period of development of each tool, hundreds of factors were considered for inclusion. All the developers relied on long-standing criminogenic theories and regression analysis to identify which factors to include in the final tool. None of the developers, even the Kleinberg group, used machine learning to identify the factors for inclusion. It is likely, however, that future risk assessment tools will be based on factors identified by machine learning. It would be interesting to examine whether there is any difference between the list of factors that the machine learning algorithm identified and the list of factors identified by regression analysis. It would also be interesting to dig deep to determine why certain factors may not have been chosen using machine learning. This information could be very useful to policy makers in designing criminal justice reforms.[260]

Another task that can be performed by machine learning is determining how best to split each one of the factors. Kleinberg and colleagues allowed the algorithm to decide which age groups to use to divide the dataset, which has proven successful. Machine learning allows researchers to try different combinations and different partitions, a task that is prohibitively complicated when using regression analysis.

Looking at the factors considered by each tool highlights the importance of giving pretrial officers multiple options to choose

[259] *See supra* Table 1.

[260] JOINT TECH. COMM., USING TECHNOLOGY TO IMPROVE PRETRIAL RELEASE DECISION-MAKING 15 (2016), https://www.ncsc.org/~/media/Files/PDF/About%20Us/Committees/JTC/JTC%20Resource%20Bulletins/IT%20in%20Pretrial%203-25-2016%20FINAL.ashx.

from: having more categories into which to group defendants can change the final score. For example, in regard to employment status, the tool used in the federal system, PTRA, distinguishes only between employed and unemployed defendants; in contrast, the VPRAI in Virginia considers being a student, a primary caregiver, and retired as other forms of employment, and so defendants in those categories are not given negative points.[261] For residence, the federal PTRA only gives two options—owning a home or being in the process of buying one—whereas renting a home is considered in other tools.[262] Given that each tool considers a relatively small number of factors and that the answer to each one is usually "yes" or "no," providing several options can improve the accuracy of the final prediction about the defendant.

Finally, as was previously mentioned, all the existing algorithms use factors that can serve as proxies for race.[263] One potential improvement is to use machine learning to count the race in the design of the algorithm but not in the prediction phase, as suggested by other researchers.[264] It is important to mention that the use of race here is to "fix" prior discrimination. Although it might be constitutionally challenging to implement such an approach, on the theoretical front, researchers are increasingly showing that explicit use of race does not harm equal protection, and can actually help racial minorities.[265]

B. Source of the Information: Interview or Only Databases

Table 2. Summary of the Source of Information Used in Pretrial Tools

Tool	**Source of Information**	
	Interview	**Database**
Pretrial Risk Assessment (PTRA): the federal tool	✓	✓
Public Safety Assessment (PSA): Arnold Foundation		✓
Virginia Pretrial Risk Assessment (VPRAI)	✓	✓
Colorado Pretrial Assessment Tool (CPAT)	✓	✓
Ohio Pretrial Assessment Tool (PAT)	✓	✓
COMPAS	✓	✓
Kleinberg et. al. tool		✓

[261] *See supra* Table 1.

[262] *See supra* Table 1.

[263] *See supra* Section 3.

[264] Yang & Dobbie, *supra* note 137, at 50.

[265] Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. (forthcoming 2020).

All of the tools examined rely on a combination of data collected from interview and databases, except for the PSA and the Kleinberg tool. There is a long-standing debate within the criminal justice literature about the type of factors that should be included in risk assessment tools. Studies show that criminal history is the factor with the highest correlation with recidivism, and it can easily be obtained and verified through criminal records.[266] A study using data from Kentucky concluded that the same level of predictive accuracy could be maintained in pretrial risk assessment tools that are based only on criminal justice data as compared to those also using interview data.[267]

Yet as mentioned before, some criminal justice experts point out that criminal history could be a flawed proxy for race, because minorities who have been historically disadvantaged and discriminated against often have lengthier criminal histories.[268] Black defendants are over-represented in the criminal justice system because they are over-policed, over-prosecuted, and over-convicted. Thus, if criminal history is included in the risk assessment tool, it will increase the ratchet effect.[269] In other words, if the efforts of the police and the judges are focused on minority groups, they will find more crime among members of those groups, and the balance between the offending population and the "carceral" population will be skewed.[270]

Other scholars argue against the use of dynamic factors like demographic and socioeconomic factors, or other information that is typically obtained in an interview. They claim that it is not fair to base predictions on items over which individuals have no control, such as the neighborhood they were born in and their gender, or on items for which they have little control, like mental and physical health status.[271] Thus, linking poverty with higher risk of pretrial failure disadvantages and punishes members of vulnerable communities who will be flagged as high risk because of factors outside of their control. In addition, considering socioeconomic

[266] Kristin Bechtel et al., *Identifying the Predictors of Pretrial Failure: A Meta-Analysis*, 75 FED. PROB. 78, 85 (2011).

[267] Bechtel et al., *A Meta-Analytic Review*, *supra* note 65, at 446.

[268] Bernard Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT'G REP. 237, 238 (2015).

[269] HARCOURT, AGAINST PREDICTION, *supra* note 69, at 145–72.

[270] *Id.*

[271] CHRISTOPHER SLOBOGIN, PROVING THE UNPROVABLE: THE ROLE OF LAW, SCIENCE, AND SPECULATION IN ADJUDICATING CULPABILITY AND DANGEROUSNESS 113 (2006); Thomas Nilsson et al., *The Precarious Practice of Forensic Psychiatric Risk Assessments*, 32 INT'L J. L. & PSYCHIATRY 400, 406 (2009); Skeem et al., *supra* note 76, at 25–26.

factors could distance the focus of the decision from the facts of criminal conduct and the law.[272]

As mentioned earlier, it would be interesting to compare the list of factors generated by machine learning with those developed by regression analysis to see if the focus is on static or dynamic factors. Although the Kleinberg machine learning tool does not rely on dynamic factors, this is because it was designed to be in line with the types of factors judges were considering in New York City under the current pretrial system. Regardless of how the factors were identified, the debate about the type of factors included in the algorithm centers on the issue of discretion. The majority of the tools relying on dynamic factors leave a great deal of room for pretrial officers to exercise their discretion and judgment. For example, in Colorado, eight of the twelve items are based on an interview, but staff have considerable discretion in marking responses as "yes" or "no." Take, for instance, Item 4: "Do you believe that you currently have or have ever had a problem with your use of alcohol?" The word "problem" is not defined in the question, so it is up to each defendant to characterize his or her drinking habits as problematic or not.[273] In addition, the revalidation study conducted in Ohio that rated the agreement among officers in scoring certain factors showed how much officers can vary in their scoring, from evaluating employment status at the time of the arrest, to determining whether or not respondents suffer from severe drug use problems.[274]

The key questions under debate are whether to leave room for discretion and judgment and, if so, how much room. The answers depend on the way pretrial officers are viewed: as expert personnel whose opinion matters or as administrators who are charged with filling out the questionnaire consistently and efficiently. None of the manuals accompanying the tools address these questions. It is also possible that the answers will change depending on the jurisdiction. Even if officers do not have discretion in calculating a score, most jurisdictions allow officers to include their recommendation (detain or release on what condition) along with the score, so their professional opinions are still taken into account.

Another way to eliminate discretion without discarding dynamic factors completely is to rely on a self-report questionnaire completed by the defendant. This method is part of the Ohio PAT, but it is used in addition to the face-to-face interview and does not replace it.[275] It would be interesting to analyze the impact of self-reporting on the final score.

[272] Sonja Starr, *The Risk Assessment Era: An Overdue Debate*, 27 FED. SENT'G REP. 205, 205 (2015).

[273] TERRANOVA & WARD., *supra* note 205, at 5.

[274] Latessa et al., *ORAS Final Report*, *supra* note 214, at 29.

[275] Latessa et al., *supra* note 214, at 5–6.

C. Data Quality

Table 3. Summary of Data Quality Assessment in Pretrial Tools

Tool	Data Quality Assessment
Pretrial Risk Assessment (PTRA): the federal tool	565,178 pretrial cases collected from all federal districts except Washington, DC, 2001–2007
Public Safety Assessment (PSA): Arnold Foundation	More than 750,000 cases drawn from more than 300 U.S. jurisdictions, 2001–2011
Virginia Pretrial Risk Assessment (VPRAI)	14,383 cases of defendants arrested in Virginia's seven localities
Colorado Pretrial Assessment Tool (CPAT)	1,315 cases collected during 16 months from 10 Colorado counties that represent 81% of Colorado's population
Ohio Pretrial Assessment Tool (PAT)	452 cases collected from seven Ohio counties, 2006–2009
COMPAS (pretrial component)	2,831 felony defendants arrested in Kent County, Michigan. 2005–2008
Kleinberg et. al. tool	Approximately 750,000 pretrial release decisions from New York City, 2008–2013

As Table 3 shows, the number of cases used to develop each tool varies dramatically, ranging between a few hundred (Ohio) to nearly a million cases (the Arnold Foundation tool and the Kleinberg tool). To be reliable, an actuarial pretrial risk assessment tool requires access to large volumes of high-quality data about as many cases as possible.[276] However, because the use of actuarial tools is relatively new, criminal justice agencies have generally not yet implemented protocols for good data collection. In addition, developing those protocols and adopting technological tools for collecting, storing, maintaining, and analyzing data are very costly, which may constrain jurisdictions.[277] In any event, criminal justice data are known to be notoriously poor.[278] If high-quality data are

[276] *See* MAMALIAN, *supra* note 66, at 7; *see also* PRETRIAL JUSTICE INST., *supra* note 66, at 5.

[277] SCHWARTZTOL ET AL., *supra* note 23, at 19–20.

[278] EXEC. OFFICE OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS 21 (May 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

not available, jurisdictions risk implementing a tool that is not suitable for their populations.[279]

The data collected need to be accurate, complete, inputted consistently, and come from the same (or at least similar) population for which the tool will be used. The PSA, for example, was based on a diverse dataset of about 750,000 cases collected from more than 300 U.S. jurisdictions. However, there were many inconsistencies in the data because each jurisdiction was collecting data on different attributes.[280] Thus, it is not just the quantity but also the quality and the consistency of the data that makes a difference in the performance of the algorithm. Another problem is that if the data used to train the algorithm was collected years ago, it may not reflect recent legislative changes. Since 2012, more than 500 bills related to the pretrial phase have been enacted, of which nearly 120 laws related to pretrial administration were passed in 2015 alone.[281] The tools based on data that do not reflect these regulations run the risk of generating "zombie predictions" that will revive old practices that are no longer legal.[282]

Opponents of machine learning algorithms raise the concern that those tools will reinforce traditional biases because the tools are trained on discriminatory data. These opponents argue that the claim that neutrality and color-blindness are associated with sophisticated algorithms is nothing more than a myth because the underlying data are already biased.[283] The ratchet effect, mentioned in the previous section, can lead to a focus on minority members who are already over-represented in the system as criminals. This reinforces the self-fulfilling prophecy of arresting more minorities, scoring them as higher risk, and detaining and convicting them.[284]

It is important to realize that judges who made pretrial decisions based on their own judgment were using the same datasets that were used to train the algorithm. Thus, the algorithm is nothing more than a mirror that reflects our own human biases and practices.[285] Several solutions have been developed to better address bias as a result of the data quality.

First, as noted in the earlier section about fairness, computer scientists have been working on different ways to deal with groups

[279] *See, e.g.*, Christin et al., *supra* note 9, at 6–8.

[280] Telephone Interview with Marie VanNostrand, Justice Project Manager, Luminosity (Oct. 13, 2017).

[281] John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725, 1729 n.6 (2018).

[282] *Id.* at 1755–56.

[283] Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947, 980 (2016).

[284] HARCOURT, AGAINST PREDICTION, *supra* note 69, at 145–72.

[285] Rahul Bhargava, *The Algorithms Aren't Biased, We Are*, MEDIUM: MIT MEDIA LAB (Jan. 3, 2018), https://medium.com/mit-media-lab/the-algorithms-arent-biased-we-are-a691f5f6f6f2.

that have different base rates of criminality. Some solutions are focused on ensuring that all groups are equally represented in a certain dataset, other solutions are focused on "favoring" one group over the others to compensate for previous discrimination, and others focus on equalizing the types of error an algorithm makes across groups.[286]

Second, special attention needs to be given to the potential impact of bias on each factor in the tool. For example, all the tools include prior arrests, even if they did not lead to conviction, as a factor. This factor likely has the highest potential of any item to reinforce biases, because the legal standard of proof (reasonable suspicion or probable cause) that the police need to establish for arresting a person is much lower than the burden of proof for conviction (beyond a reasonable doubt). Therefore, it is possible that some arrests are based on the police officers' prejudice or negative previous encounters with other members of the group to which the defendant belongs.[287]

Third, it is very important to ask the algorithm the right questions in order to produce a particular targeted outcome. Attention to mitigating bias and the trade-off between fairness and accuracy should be embodied in the design of the algorithm in advance.[288] Thus, policy makers could decide on the explicit goal that they wish to achieve using the tool—keeping risky defendants away from society or making sure that low risk defendants are not spending unnecessary time in jail—and could tune the algorithm differently given that goal.

Fourth, data included in the algorithm should be collected from diverse sources. In addition to that provided by the police and the court, data could also be collected from the Bureau of Justice Statistics National Crime Victimization Survey, which tracks crimes based on victim reports.[289]

[286] PARTNERSHIP ON AI, REPORT ON ALGORITHMIC RISK ASSESSMENT TOOLS IN THE U.S. CRIMINAL JUSTICE SYSTEM 1, 19-20 (Apr. 2019), https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/.

[287] DOYLE ET AL., *supra* note 31, at 16.

[288] Nicol Turner Lee et al., *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, BROOKINGS INST. (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/#cancel.

[289] Simmons, *supra* note 284, at 983.

D. Periodic Validation

Table 4. Summary of Revalidation Studies Made for Pretrial Tools

Tool	Revalidation Study
Pretrial Risk Assessment (PTRA): the federal tool	Validated twice, in **2010** and in **2019**.
Public Safety Assessment (PSA): Arnold Foundation	The PSA is being used in more than 38 jurisdictions, so different validation studies involving more than **650,000 cases** have been conducted. In addition, the LJAF is funding **external organizations** such as the Access to Justice Lab at Harvard, for conducting revalidation studies.
Virginia Pretrial Risk Assessment (VPRAI)	Revalidated twice, in **2007** and in **2014**. The results of the studies were taken seriously and led to major revisions in the tool.
Colorado Pretrial Assessment Tool (CPAT)	Revalidation study began in 2018 and is expected to be completed in **2020**.
Ohio Pretrial Assessment Tool (PAT)	Validated upon the implementation in 2009 and again in **2018**. The 2018 study examined an important aspect: the **inter-rater reliability**.
COMPAS	Northpointe has been tracking and validating the pretrial tool internally, but **no external studies** have been conducted. Many revalidation studies were conducted for the general recidivism tool based on the same data analyzed by ProPublica; however, those studies are **partial** because the tool is **proprietary**.
Kleinberg et al. tool	The tool has not been used in practice yet, so no additional revalidation study was conducted based on actual outcomes. However, the researchers revalidated their findings on a **national dataset**.

As can be seen from Table 4, the tools vary in regard to how often they are validated, what exactly is assessed by the revalidation process, and who conducts the validation. The validation process is essential both to (1) determine that the risk assessment tool reflects current regulations and social and technological trends and (2) to evaluate its performance vis-a-vis the local population as a whole, as well as specific minority groups. For example, a study in New York City twenty years ago showed that having a landline phone in the defendant's house was a good predictor for showing up at trial.[290] However, given the increased reliance on cellphones and other connective devices today, it is unlikely that the landline factor is still a good predictor. Periodical validation is also useful for building trust in the tool among the public, litigants, and judges.[291]

Some jurisdictions have unique needs that require them to develop their own risk assessment tools. For example, the federal courts are the only courts that have the authority to deal with immigration-related cases; thus, immigration-related factors need to be included in any federal pretrial risk assessment tool.[292] However, most jurisdictions adopt tools used elsewhere because of the high costs of developing their own tools. A survey conducted by the U.S. Department of Justice found that 39% of agencies using pretrial risk assessment tools adapted them from another jurisdiction.[293] Of those agencies, only 25% validated the tool for use on their own populations.[294] Because the risk assessment tool provides an estimation that is relative to the group of people that the defendants is compared with, if the tool is not validated to each defendant's jurisdiction, the rate will not mean the same thing.[295]

Several jurisdictions have conducted high-quality, thorough revalidation studies and can serve as models for the field. The VPRAI has been validated three times since it was developed; the last study was in 2014 and led to major revisions of the tool. That study examined the predictive accuracy of each of its factors for the general population, across race and gender. In addition, new potential factors were included to assess whether they would increase the predictive accuracy of the tool. The revised VPRAI includes more response options for several factors; for example, the employment factor now includes student, part-time worker, retired, or primary caregiver as forms of employment.[296]

[290] QUDSIA SIDDIQI ET AL., PREDICTION OF PRETRIAL FAILURE TO APPEAR AND AN ALTERNATIVE PRETRIAL RELEASE RISK-CLASSIFICATION SCHEME IN NEW YORK CITY: A REASSESSMENT STUDY, NEW YORK CITY CRIMINAL JUSTICE AGENCY 22 (2002).

[291] Joint Tech. Comm., *supra* note 261, at 20–21.

[292] Cadigan et al., *The Re-Validation of the Federal Pretrial Services Risk Assessment*, *supra* note 150, at 6.

[293] Christopher Lowenkamp et al., *The Development and Validation of a Pretrial Screening Tool*, 72 FED. PROB. 2, 3 (2008).

[294] *Id.*

[295] DESMARAIS & LOWDER, PRETRIAL RISK ASSESSMENT TOOLS, *supra* note 71, at 4-5.

[296] DANNER ET AL., *supra* note 182, at 12-13.

The Arnold Foundation is funding highly regarded criminal justice research institutes, such as the Access to Justice Lab of Harvard Law School, to examine the validity of the PSA.[297] The Access to Justice Lab is currently conducting a well-designed study that can shed light on the true predictability of PSA compared to human judgment and decision making.

The 2018 validation study of the Ohio PAT examined the inter-rater reliability of the tool, an important aspect that is rarely researched in the context of risk assessment tools.[298] The research examined whether there was consistency between different officers in scoring each factor and ultimately each case.[299] The results can be used to design the training that officers receive to enhance consistency and can also aid in the revision of the wording of each question to make them easier to score consistently. Inter-rater reliability is very important because if the staff do not apply the tool consistently, its utility will be reduced even if its predictive validity is very high.[300]

If data about the performance of the algorithm are collected in a coherent and organized way, machine learning can make conducting revalidation studies faster and more efficient. In addition, as illustrated by the Kleinberg study, machine learning methods can train the algorithm and then validate it. For example, the *K*-fold cross-validation technique, described earlier, makes it possible to use the whole dataset for both training and validation, which is especially useful when the dataset is small.

[297] Access to Justice Lab, *supra* note 179.

[298] Desmarais et al., *supra* note 71, at 6.

[299] Latessa et al., *ORAS Final Report*, *supra* note 206, at 29.

[300] Faye S. Taxman, *Risk Assessment: Where Do We Go From Here?*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 271, 274 (Jay P. Singh et al. eds., 2018).

E. Ways of Implementing Risk Assessment Tools

Table 5. Summary of Implementation Mechanisms in Pretrial Tools

Tool	Implementation
Pretrial Risk Assessment (PTRA): the federal tool	Implemented in **all federal districts**. The score, along with the pretrial officer's recommendation, is provided to the judge as part of a report. Information about training was not available at the time of publication.
Public Safety Assessment (PSA): Arnold Foundation	Implemented so far in more than **38 jurisdictions,** but the process of the implementation varies. In general, the LJAF provides technical support and training for jurisdictions.
Virginia Pretrial Risk Assessment (VPRAI)	Implemented **state-wide in Virginia and in different jurisdictions across 12 states**. In Virginia, it is part of the Pretrial and Community Corrections (PTCC) case management system and is managed by the Virginia Department of Criminal Justice Services.
Colorado Pretrial Assessment Tool (CPAT)	Implemented so far in **22 counties across Colorado**. Training for using the tool is advised and could be done through the Colorado Association of Pretrial Services (CAPS), which also published a publicly-available training manual.
Ohio Pretrial Assessment Tool (PAT)	Implemented in **different counties in Ohio as well as counties in Indiana, Texas, Massachusetts and California**. Training is required for jurisdictions that purchase more than one risk assessment tool; it is provided by the University of Cincinnati Correction Institute.
COMPAS	The pretrial tool has been implemented in **two counties in California**. The General Recidivism tool and the Violent Recidivism tool are used in different counties in Florida, Wisconsin and California. Training is required and provided by Northpointe.
Kleinberg et. al. tool	A tool based on the study has not been implemented yet, but it is being considered in New York City.

As can be seen from Table 5, the implementation of the tools varies widely. The VPRAI, the Ohio tool, and the federal PTRA are similar in that they have been implemented for the entire population for which they were developed: statewide in Virginia and Ohio; and countrywide in the federal system. The Arnold Foundation PAT has also been implemented statewide in Arizona, Kentucky, Utah, and New Jersey. Given the advantages described below, it is recommended that Colorado pretrial assessment tool be implemented statewide, because it was created for and validated on its own population. Adopting the same tool statewide has many advantages: it facilitates the implementation process, fosters more uniformity among law enforcement in the state, enables pretrial officers to share knowledge gained from their experience with the same tool, and enhances the creation of guidelines and detailed manuals for using it.[301]

Another important implementation issue is how the judges make use of the score produced by the tool. A 2017 study of usage of a risk assessment tool in Kentucky showed that it produced different results at the county and the state-wide level. Within each county, Black and White defendants scored similarly on the risk assessment tool and their outcomes were similar; however, state-wide, White defendants were released during pretrial at much higher rates than were Black defendants. The researcher attributed this difference to the way judges interacted with and were influenced by the tool. In counties with more White defendants, the judges liberalized bail practices compared to judges in counties that had predominantly Black defendants.[302] Thus, it is very important to collect data about the percentage of cases on which judges agree or diverge from the risk score produced by the tool and why they do so.[303]

The implementation process should also take into account specific factors that are relevant to each jurisdiction, such as current and anticipated jail density, attitude towards incarceration, and tolerance of misbehavior.[304]

The level of training that pretrial officers and judges receive in using the tool also varies. Only COMPAS and the Ohio PAT require training, and only require it if more than one risk assessment tool has been purchased from the package. The other tools merely

[301] PRETRIAL JUSTICE INST., THE STATE OF PRETRIAL JUSTICE IN AMERICA, *supra* note 17, at 7.

[302] Christopher Bavitz et al., *Assessing the Assessments: Lessons from Early State Experiences in the Procurement and Implementation of Risk Assessment Tools*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y 1, 5 (2018).

[303] Koepke & Robinson, *Danger Ahead*, *supra* note 282, at 1796.

[304] Access to Justice Lab, *supra* note 179.

advise training.[305] It would be very beneficial if training be mandatory for all officers and judges using all available tools. Training is particularly important when implementing a machine learning based tool to clarify its capabilities and limitations. In addition to providing extensive training for judges and pretrial officers who work with the tool daily, it is important also to provide defense lawyers, prosecutors, and the public with sufficient knowledge so that they understand how a certain score was calculated and can challenge its use when its results seem arbitrary or unfair.

The support offered by the Arnold Foundation for jurisdictions implementing the PSA illustrates the importance of training. Its tailored training focuses on how to collect the needed data for implementing the tool and setting guidelines for communicating the score to judges, prosecutors, and defense lawyers.[306] This detailed implementation package is provided by the Arnold Foundation because the PSA was designed from the start to be implemented on a nationwide basis.

F. Double Counting

Table 6. Do Pretrial Tools Contain Double-Counted Factors?

Tool	Implementation
Pretrial Risk Assessment (PTRA): the federal tool	*Potentially yes*: There may be overlap between item 1, pending charges, and items 2–3, previous misdemeanor and felony arrest; and between item 7, substance abuse, and item 10, alcohol consumption.
Public Safety Assessment (PSA): Arnold Foundation	*Potentially yes*: Several items in the criminal history can be counted more than once; for example, item 6 counts prior convictions for violent crime and item 5 counts prior felony convictions.
Virginia Pretrial Risk Assessment (VPRAI)	*Yes*: Item 4 that refers generally to criminal history could overlap with items 3, 5, and 6 that examine different aspects of criminal history.
Colorado Pretrial Assessment Tool (CPAT)	*Potentially yes*: Although the factors are quite distinct, items 1-3, having a home or cell phone, owning or renting one's residence, and contributing to residential payments, might

[305] *See supra* Table 5.

[306] Mathew Demichele et al., *The Intuitive-Override Model: Nudging Judges Toward Pretrial Risk Assessment Instruments* 11 (Apr. 25, 2018) (unpublished article) (available at https://ssrn.com/abstract=3168500).

	slightly overlap. Item 10, other pending charges, could be a good attempt to deal with double counting.
Ohio Pretrial Assessment Tool (PAT)	*Most likely no*: The only two factors that are closely related but still quite distinct are items 6–7, illegal drug use during past six months and severe drug use problem.
COMPAS	*Yes*: All the tools use different components of criminal history multiple times, and if jurisdictions are using more than one instrument, then double counting is unavoidable because there is overlap between the tools.
Kleinberg et. al. tool	*No:* The number of factors used in the tool is low.

The issue of double counting, summarized in Table 6, occurs when the same item is scored more than once, which could increase the final score of a defendant unjustifiably.[307] When judges were making those decisions without the aid of a risk assessment tool, they could read the entire file of the defendant, evaluate it as a whole, and decide accordingly. When actuarial tools are involved, the risk is that the impact of some factors will be magnified, which will have a negative consequence on the final outcome.[308]

As seen in Table 6, nearly all the tools include overlapping factors, and the most frequently double-counted factor is criminal history. For example, item 5 in the PSA refers to prior felony convictions, and item 6 refers to prior convictions for violent crime. Presumably most violent crimes are felonies, so the same offense could be counted twice unless the manual specifies for the examiners that only a ratio of the same offense should be counted toward the other predictor. In the VRPAI, item 4 that refers to overall criminal history could overlap with items 3, 5, and 6, which refer to other particular aspects of one's criminal history (pending charge, failures to appear, and violent conviction). Item 10 in the Colorado PAT attempts to deal with the issue of double counting: it refers to "other pending charges." The use of the word "other" implies that this item deals with aspects of criminal history not already accounted for in the other items. This could be a way to word the items such that that no item in the criminal history is uncounted, but also ensures that each item is not counted more than once. The tool in use that is the least prone to double counting, when used properly, is the Ohio

[307] Melissa Hamilton, *Back to the Future: The Influence of Criminal History on Risk Assessments*, 20 BERKELEY J. CRIM. L. 75, 96–97 (2015).

[308] *Id.* at 98.

PAT. The seven factors used in the tool are quite distinct, with each one touching on a different aspect of the defendant's history.

Double counting can also occur when the jurisdiction is using a tool that is part of a suite or system of tools: this is the case for the Ohio PAT, which is part of the Ohio Risk Assessment System, and for COMPAS, which is part of the Northpointe Suite. In this case, jurisdictions might be using two tools in the same phase of the criminal justice system, such as for predicting a general pretrial score and a score for violent crime. This issue is not addressed in the manuals of the risk assessment tools, despite its clear importance.

G. *The Meaning of the Predicted Score*

Table 7. Summary of Outcome Being Predicted by Pretrial Tools

Tool	**Outcome**
Pretrial Risk Assessment (PTRA): the federal tool	**One score on a scale of 1-5.** Each risk score represents an X% risk of failure to appear, Y% risk of new criminal arrest, and Z% risk for pretrial revocation.
Public Safety Assessment (PSA): Arnold Foundation	**Two separate scores on a scale of 1-6**, one for failure to appear and one for new criminal arrest; + a raw score (yes/ no) for the risk to commit a new violent crime.
Virginia Pretrial Risk Assessment (VPRAI)	**One score on a scale of 1-6** that compounds failure to appear and new criminal arrest. The possibility to separate the scores and add a score for violent crime is being examined.
Colorado Pretrial Assessment Tool (CPAT)	**One score on a scale of 1-4**. Reporting is done in the following format "[Defendant's name] has a CPAT risk score consistent with other Colorado defendants whose average public safety rate is [##]% and whose average court appearance rate is [##]%."
Ohio Pretrial Assessment Tool (PAT)	**One score on a scale of 0-9**.
COMPAS	The pretrial tool provides **one score on a scale of 1-10** for pretrial misconduct which includes failure to appear and arrest for a new felony offence while on pretrial release.
Kleinberg et al. tool	The tool determines the likelihood that the defendant will **fail to appear in percentage**. The tool only considers failure to appear as an outcome because this is the only factor that judges in New York City are allowed to consider.

As observed in Table 7, the final output and meaning of each given score varies. Only the PSA separates the two outcomes and provides a different score on a scale of 1–6 for new crime arrest and for failure to appear. Although the COMPAS pretrial tool does not separate the outcomes, another tool that is part of the Northpointe suite—the General Recidivism Tool—focuses only on new crime arrests. The federal PTRA and the Colorado PAT, within the single score that they provide, estimate the probability, when compared with other defendants, for each defendant to commit a new crime or to be rearrested.

Four of the seven tools generate a combined score for both failure to appear and the risk of committing a crime while awaiting trial. Although both risks are important and judges have to take them into account, they have completely different meanings and opportunities for mitigation.[309] There are many effective ways to reduce the risk of a failure to appear—for example, sending reminders of the court date, providing community supervision, and giving transportation vouchers valid for the date of the hearing to low-income defendants. Therefore, the combined score could miss the important distinction between the two behaviors and flag defendants as high or low risk without providing judges the requisite information to understand what exactly this score means.

There are both legal and policy considerations that support separating the likelihood of failure to appear and the likelihood of committing a new crime into two scores. From a legal perspective, the types of evidence and the government's burden of proof for danger-based detention (i.e., preventative detention to protect the public) are higher than the evidentiary standard adopted by the courts in regard to flight risk-based detention.[310] From a policy perspective, separating the scores could reduce judges' reliance on their intuition, because they would have to explain, for example, why they decided to detain a defendant who was at low risk for a new crime arrest but at high risk for a failure to appear. Separate scores would make the link between the statistical probability and the actual outcome clearer, which should be reflected in judges' opinions.[311]

In addition, separating flight risk and public safety will improve the ability to impose conditions of release that are more closely aligned with defendants' needs. Defendants may fail to appear for their pretrial hearings for a variety of reasons, ranging from leaving the country to escape their sentences, to a lack of money to commute to court, to the need to work and support their

[309] Lauryn P. Gouldin, *Disentangling Flight Risk from Dangerousness*, 2016 B.Y.U. L. REV. 839, 844.

[310] *Id.* at 873.

[311] *Id.* at 886.

families. Each reason can be deterred by different methods. For example, the literature shows that it is much more effective to send reminders about hearing dates via text messages than by mailed postcards.[312] Another tool that could be particularly useful in reducing failures to appear is electronic monitoring.[313] Similarly, there should be a nuanced response to those at high risk of committing a new crime. The array of offenses in the criminal code is huge, and, certainly, the danger to the community from someone committing a murder is not the same as that from someone jaywalking.[314] Thus, policymakers need to calculate the risks that they are willing to take in balancing between false positives and false negatives.

Having separate scores for new crime arrests and for failures to appear is undoubtedly the first step, but the predictive accuracy of the risk assessment tools in terms of actual pretrial misconduct needs to increase. A helpful direction of research could be to predict more specifically the type of crime that the defendant is at risk to commit—in other words, to give a more tailored meaning to the score.[315] The PSA generates a separate score for the risk of committing a violent crime, and COMPAS has a tool designed for scoring violent crimes. Virginia is considering adding a score for violent crimes, and the recent revalidation studies of the federal tool and the Ohio PAT examined their ability to predict the risk to commit a violent crime. An updated tool could ask, for example, "What is the likelihood that the defendant will commit a felony?" The risks that the tools predict should be as specific as possible, and machine learning can play a key role in providing a more precise prediction for the risk level of a defendant to commit a certain type of crime and the risk level of a defendant for certain kinds of flight risk. Educating judges and all actors in the field about the meaning of such a score is very important because the statistical probability of the actual score may be lower than what judges think.[316]

[312] Koepke & Robinson, *Danger Ahead*, *supra* note 282, at 1765.

[313] *Id.* at 1767.

[314] Berk & Hyatt, *supra* note 93, at 223.

[315] Taxman, *supra* note 301, at 275–76.

[316] David G. Robinson et al., *Pretrial Risk Assessments: A Practical Guide for Judges*, 57 JUDGES J. 8, 9 (2018).

H. Possibility to Challenge the Outcome of the Tool

Table 8. Ability to Appeal the Outcome of Pretrial Tools

Tool	**Appealability**	
	Easy to Appeal	**Hard to Appeal**
Pretrial Risk Assessment (PTRA): the federal tool	✓	
Public Safety Assessment (PSA): Arnold Foundation	✓	
Virginia Pretrial Risk Assessment (VPRAI)	✓	
Colorado Pretrial Assessment Tool (CPAT)	✓	
Ohio Pretrial Assessment Tool (PAT)	✓	
COMPAS		✓
Kleinberg et al. tool	✓	

As shown in Table 8, it is easy in principle for defendants to appeal the scores given by every tool, except for COMPAS. Due process requires at a minimum that the decision to detain someone is made by a judge who assesses the evidence and the accuracy of the information brought in front of him or her, including the risk assessment tool, and gives a clear and detailed ruling. Then, in theory, the defendant is able to appeal the ruling and challenge what he or she perceives as an unfair or inaccurate judgment.[317] The meaning of due process differs between the pretrial and postconviction stages. In pretrial, imposing conditions for release or detention on the defendant is aligned with due process requirements so long as "they are reasonably related to a legitimate and non-punitive governmental purpose."[318]

Most criticisms of machine learning techniques center around the fact that the decision is untraceable and therefore unappealable. But pretrial hearings are already opaque, and using a machine learning-based tool, if done properly, is unlikely to cause more harm and may prove beneficial. Jurisdictions vary widely in the conditions of pretrial hearings, but typically they do not last more

[317] Frank I. Michelman, *Formal and Associational Aims in Procedural Due Process*, 18 NOMOS 126, 132–33 (1977).

[318] Hamilton, *supra* note 50, at 267–68.

than a few minutes, they often take place through video conference rather than in person, legal representation is not always provided, and the official who makes the decision is often a magistrate and not necessarily a judge. In addition, there is evidence showing that court officials spend very little time looking at each defendant's file and determining the release or detention conditions.[319] The procedure for appealing a pretrial decision varies across jurisdictions, but in general the decision is subject to a strict standard of review.[320] In any case, the defendant's ability to raise substantive claims about the weight that each factor is given is limited.

The scores generated by every tool used today in the pretrial phase—even the COMPAS and Kleinberg tools—are neither random nor completely understandable. Their lists of factors used to generate the scores are publicly available and can be used by defendants in appealing the outcomes. The hardest scores to appeal are those generated by COMPAS, not because of its machine learning components, but because of its proprietary nature. The other tools provide detailed manuals about their operation, which can provide solid grounds for defense lawyers to appeal a pretrial decision, if they think it was made because of an error in the scoring. Therefore, jurisdictions should use a tool whose operation is as transparent as possible. In this context, transparency refers not just to the inner workings of the algorithm but also to the procurement of those tools.[321] The only exception would be if the proprietary, nontransparent tool performed significantly better than the other tools, something that is yet to be proven. Thus, Northpointe and other private companies should be able to release more information and instructions about their tools without compromising their commercial advantage.

It would be interesting to investigate whether any difference exists between the acceptance rate of appeals when a risk assessment tool was used versus when it was not used in imposing conditions in the pretrial phase. However, given that most tools used these days are relatively easy to understand, it seems safe to assume that there is not a significant difference.

V. CONCLUSION

The pretrial phase is the "front door" of the criminal justice system, and any decision about the defendant is highly likely to affect the rest of the trial and the defendant's future. Therefore, it is crucial that decisions made at this stage are fair, legal and unbiased. This paper examines the implications of using machine learning to

[319] Stevenson & Mayson, *supra* note 2, at 25.

[320] Nick Oberheiden, *Appealing a Pre-Trial Detention Order (Bond)*, FED. LAW. (2016), https://federal-lawyer.com/appealing-a-pre-trial-detention-order-bond/.

[321] Bavitz et al., *Assessing the Assessments*, *supra* note 303, at 3.

develop risk assessment tools used in pretrial and to investigate whether these tools are a major problem in the criminal justice system, as portrayed by the media and some scholars. Machine learning has a set of unique strengths and weaknesses that challenge our commitment to human judgment and basic concepts of law. Because of the way machine learning algorithms operate, they require us to adopt new ways of understanding concepts such as transparency, explainability, and fairness. However, a comparison between machine learning and regression analysis shows us that there are more similarities than differences between the two, and the comparison of the seven tools presented in this article strengthens this conclusion. In regard to each policy consideration, the article concludes that adding a machine learning aspect to risk assessment tools will not worsen the outcome, and in many cases may improve it. Machine learning algorithms may speed up the revalidation process, allow each factor to be divided into subcategories, and produce a more personalized score that has interpretable meaning for the defendant before the court.

Successful implementation of machine learning algorithms that improve criminal justice outcomes requires that the following conditions be met. First, collaboration between the engineers who create the algorithms and the policymakers responsible for their implementation is needed to ensure that both groups have a comprehensive understanding of the capabilities and limitations of the algorithms. Second, consensus about the trade-offs between concepts such as fairness, accuracy, efficiency, transparency, justice, and equity needs to be part of the system design.[322] Third, proper safeguards are needed to ensure that machine learning algorithms comply with legal principles such as due process and equal protection. Such safeguards could include built-in accountability mechanisms that guarantee that the score is understandable and appealable. If a proprietary algorithm is used, this may require negotiating around proprietary clauses in the contract between the law enforcement agency and the producer of the algorithm.

The use of actuarial risk assessment tools alone cannot reverse centuries of racial injustice or gender inequality, but if used together with other means such as eliminating or restricting money bail, it can reduce it. The Pretrial Justice Institute, a leading nonprofit organization in the field, published a comprehensive report in 2017, *The State of Pretrial Justice in America*, which gives all fifty states a score on a scale of A–F, (A being the highest score and F the lowest), based on their success in implementing pretrial reforms. The score is based on the number of people per capita held in local jail awaiting trial, the percentage of the population living in an area where an actuarial risk assessment tool is being used, and

[322] Calo, *supra* note 86, at 415.

the percentage of people living in an area where money bond has been eliminated.[323]

The only state that received an "A" is New Jersey, where money bail has been eliminated, except in instances where no other condition is sufficient, and where the Arnold Foundation PSA has been implemented statewide. As a result of these reforms, pretrial detention has dropped by 34% and there has been a reduction in all types of crime.[324] Nine states, including Virginia, received a "B." In five of these, an actuarial risk assessment tool has been implemented statewide, and in the other four states, more than 80% of the population live in an area where a tool is being used. Ten states were classified as a "C" because they had implemented some pretrial reforms but had not completed the process. At the bottom of the list are 17 states classified in category "F": Alabama, Alaska, Arkansas, Georgia, Idaho, Indiana, Louisiana, Mississippi, Missouri, Montana, Nebraska, North Dakota, Oklahoma, South Carolina, Tennessee, West Virginia, and Wyoming.[325]

These findings show that 25% of the American people live in a jurisdiction that has implemented an actuarial tool, an improvement from four years ago when only 10% did so. This progress cannot be solely attributed to machine learning. However, as this paper argued, adopting more sophisticated algorithms will not reverse this trend. On the contrary, we could accelerate the rate of reform by taking the right precautionary measures to ensure that these algorithms are implemented properly and in a way that balances the interests of society, defendants, and the justice system.

[323] PRETRIAL JUSTICE INST., THE STATE OF PRETRIAL JUSTICE IN AMERICA, *supra* note 17, at 6–8.

[324] *Id.* at 4.

[325] *Id.* at 11–12.