

Reimagining Prediction Algorithms in the Criminal Justice System: Machine Learning Contributing to Mass Incarceration

Daveon Lilly*

ABSTRACT

Machine learning, better known as artificial intelligence, is more than Alexa adding milk to your grocery list and Chat GPT writing an essay. Machine learning also predicts the likelihood of an incarcerated individual reoffending before they are released. Probation and parole officers use prediction algorithms to help persuade the judge to go with their probation, release, or sentencing term recommendation.

This Note's central claim is that these prediction algorithms contribute to mass incarceration because the algorithms' biases keep Black men incarcerated while releasing white men who have committed similar or worse crimes. The prediction algorithms should be used to help incarcerated people re-enter society rather than decreasing their chances of re-entering altogether.

This Note will clarify the criminal justice system for outsiders and help outsiders understand machine learning. Second, we will discuss the history of prediction algorithms and an algorithm in current use in the criminal justice system. Third, consider the groups that depend heavily on these algorithms and the reasons behind their trust. Fourth, the connection between COMPAS and mass incarceration will be made. Fifth, there are ways to fix the algorithm as a prediction tool by being more transparent and reliable. Finally, a way to improve the issues within the algorithm and how the prediction tool can help incarcerated people re-enter society.

*B.S., 2019, and B.A., 2019, Arizona State University; J.D. Candidate, 2025, West Virginia University College of Law. I thank the Mid-Atlantic Black Law Students Association (MABLSA) Journal Review Team for their thoughtful feedback and support.

TABLE OF CONTENTS

INTRODUCTION	3
I BACKGROUND	3
A <i>Background Information on Mass Incarceration</i>	4
B <i>Background Information on Machine Learning</i>	5
II PREDICTION ALGORITHMS USED IN THE CRIMINAL JUSTICE SYSTEM	6
A <i>History of Prediction Algorithms</i>	6
B <i>Current Use of Prediction Algorithms</i>	8
C <i>COMPAS: An Algorithm Used in the Criminal Justice System for Sentencing</i>	8
III OFFICERS OF THE COURT’S DEPENDABILITY ON PREDICTION ALGORITHMS	10
A <i>Probation Officers’ Dependability on Algorithms</i>	10
B <i>Judges’ Dependability on Algorithms</i>	11
C <i>Attorneys’ Dependability on Algorithms</i>	12
IV PREDICTION ALGORITHMS CONTRIBUTING TO MASS INCARCERATION BECAUSE OF BIAS	13
V PRINCIPLES GUIDING PREDICTION ALGORITHMS’ FAIRNESS AND ACCURACY	14
A <i>Transparency</i>	15
B <i>Reliability</i>	17
VI RECOMMENDATIONS ON HOW PREDICTION ALGORITHMS CAN BE LESS HARMFUL IN THE CRIMINAL JUSTICE SYSTEM	18
A <i>Incorporating Unlikely People in the Development of Algorithms</i> .	18
B <i>Machine Learning Helping Incarcerated People Reenter Society</i> . .	19
CONCLUSION	19

INTRODUCTION

There are various artificial intelligence algorithms we all use every day. Two algorithms that use prediction algorithms are Facebook and Netflix. Facebook uses machine learning to “generate the estimated action rate” that predicts a person’s likelihood of seeing a specific ad, visiting the website, or purchasing something from the website.¹ Their machine prediction algorithm uses a user’s behavior on and off Facebook to personalize ads for them.²

Netflix provides personalized content recommendations to every user by analyzing one’s “viewing history, ratings, and other data” to predict movies and television shows the user may want to watch.³ Those outstanding and personalized algorithms are similar to the recidivism prediction algorithms used in the criminal justice system. Instead of showing you a pair of shoes you have been thinking about buying or a show similar to something you watched months ago, the algorithm predicts the likelihood of an incarcerated individual committing a crime after release.

This Note argues that as remarkable a tool as machine learning can be, it significantly contributes to mass incarceration—particularly through prediction tools used during sentencing hearings. To decrease machine learning’s contribution to mass incarceration, we should limit algorithmic biases and help incarcerated people prepare to re-enter society because the system is working against Black men more aggressively now than ever, with algorithms doing much of the work.

Part I of this Note describes the background knowledge of mass incarceration and machine learning needed to fully understand the prediction algorithms used in the criminal justice system. Part II examines the history and current use of specific prediction algorithms utilized in courts today. Part III details the dependability that probation officers, judges, and attorneys place on prediction algorithms. Part IV analyzes the role of machine learning in contributing to mass incarceration, emphasizing the adverse impacts these tools have on the criminal justice system. Part V focuses on solutions to improve recidivism prediction algorithms so they can be used accurately, fairly, and without contributing to mass incarceration. Part VI provides recommendations for limiting or ending the use of recidivism prediction algorithms and restructuring them for purposes that genuinely support successful re-entry for incarcerated people.

I. BACKGROUND

Two concepts need to be understood at a basic level before understanding the prediction algorithms used in the criminal justice system. The two concepts are mass incarceration and machine learning.

¹Meta, *Good Questions, Real Answers: How Does Facebook Use Machine Learning to Deliver Ads?*, <https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads> (last visited Dec. 15, 2023).

²*Id.*

³Netflix, *Machine Learning Research*, <https://research.netflix.com/research-area/machine-learning> (last visited Dec. 15, 2023).

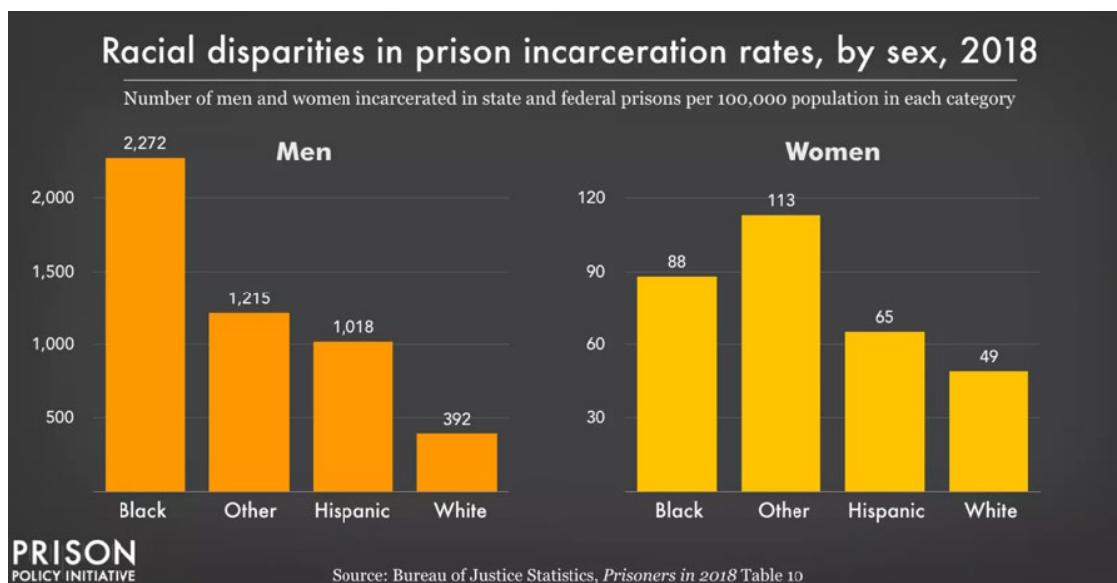
A. *Background Information on Mass Incarceration*

The first concept is mass incarceration and who it impacts within the United States. In the 1970s, the incarceration rate climbed rapidly after remaining around 100 per 100,000 people for decades.⁴ In 2008, “nearly 1 in 100 adults” in the United States were imprisoned.⁵ Nearly eight years later, by 2016, there was a fourteen percent decline, but mass incarceration remains a persistent problem today.

Mass incarceration refers to the extreme and excessive rates of imprisonment within the United States, disproportionately affecting people of color— specifically Black men. The term extends beyond “policing to prosecutorial decisions, pretrial release processes, sentencing, correctional discipline, and even reentry,” affecting both children and adults.⁶

There are several theories for the cause of mass incarceration. One theory is that mass incarceration is the product of institutional racism.⁷ A second theory points to the influential role lawmakers played in driving mass incarceration by passing harsh sentencing laws.⁸ Regardless of the theory, far too many children, men, and women are behind bars when they do not have to be.

Figure 1.⁹



⁴Katherine Beckett & Megan Ming Francis, *The Origins of Mass Incarceration: The Racial Politics of Crime and Punishment in the Post-Civil Rights Era*, 16 ANN. REV. L. & SOC. SCI. 433, 434 (2020).

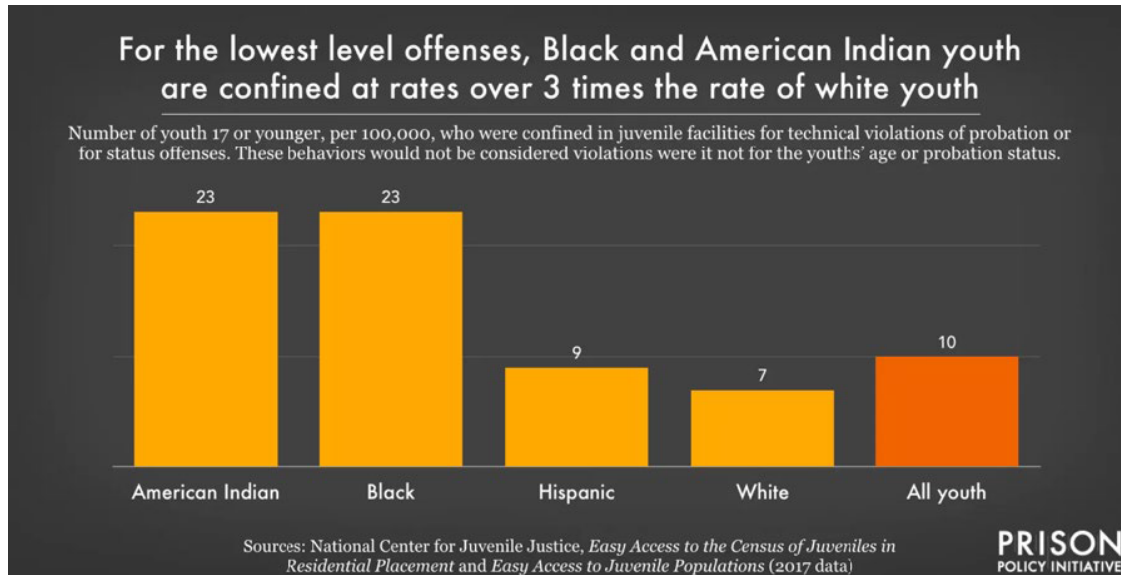
⁵*Id.*

⁶Wendy Sawyer, *Visualizing the Racial Disparities in Mass Incarceration* (July 27, 2020), <https://www.prisonpolicy.org/blog/2020/07/27/disparities/>.

⁷MICHELLE ALEXANDER, *THE NEW JIM CROW: MASS INCARCERATION IN THE AGE OF COLORBLINDNESS* (2012).

⁸Jeffery Bellin, *Reassessing Prosecutorial Power Through the Lens of Mass Incarceration*, 116 MICH. L. REV. 835, 840 (2018).

⁹Sawyer, *supra* note 7.

Figure 2.¹⁰

Both figures show that people of color, specifically Black men, women, and youth, and American Indian youth and women, are incarcerated at drastically higher rates than white and Hispanic men, women, and youth. Mass incarceration is not a new topic and will probably be prevalent for the rest of our lives. The systemic racism embedded in mass incarceration and a lot of institutional structures is too overwhelming. Change needs to come from somewhere, and this Note will propose a solution to help decrease mass incarceration.

B. Background Information on Machine Learning

The second concept is an extension of artificial intelligence known as machine learning. John McCarthy originated artificial intelligence (AI) to “describe a new field of computer science” and create algorithms that could teach themselves how to carry out specific tasks.¹¹

Machine learning is a class of AI techniques.¹² It can be defined as “a set of methods that allow computers to learn from data to make and improve predictions.”¹³ Put differently, machine learning involves an automated process for discovering correlations between variables in a dataset to predict or estimate an outcome.¹⁴ Machine

¹⁰*Id.*

¹¹Chris Meserole, *What Is Machine Learning?* (Oct. 4, 2018), <https://www.brookings.edu/articles/what-is-machine-learning/> (explaining McCarthy’s commitment to creating computers that “could observe the world and then make decisions based on those observations—to demonstrate, that is, an innate intelligence”).

¹²Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 88 (2014).

¹³Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* 2.3 (2d ed. 2023).

¹⁴David Lehr & Paul Ohm, *Playing with the Data: What Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 671 (2017).

learning enables algorithms to improve their performance on a task through experience by adjusting their behavior.¹⁵

The machine-learning process relies heavily on statistics and has become increasingly popular as the term has become synonymous with AI.¹⁶ As a result, it is difficult to “tease out the implications of AI without understanding how machine learning works.”¹⁷ In machine learning, intelligence depends on probability rather than abstract reasoning or logic.¹⁸

For example, when we play Monopoly, we estimate what number we need to roll to avoid landing on another player’s property or to reach an unowned property to purchase. Our estimations depend not on high-level reasoning but on our “ability to accurately assess how likely something is.”¹⁹ Computers operate similarly—they compute probabilities when provided with sufficient data.

Machine-learning prediction algorithms are used in numerous everyday applications, such as Internet search results, facial recognition systems, fraud detection, and email spam filters.²⁰ In simpler terms, machine learning—as a subcategory of AI—teaches machines how to perform specific tasks and generate accurate results by identifying patterns after being trained on massive amounts of data.

II. PREDICTION ALGORITHMS USED IN THE CRIMINAL JUSTICE SYSTEM

AI can be used in many ways in the criminal justice system. One way AI has become prominent is through risk-needs assessments (RNAs), particularly within the past decade.²¹ Various RNA prediction algorithms are used in courts across the United States. First, this Note reviews the history of RNA prediction algorithms. Second, it examines the current use of prediction algorithms in courts, probation offices, and other settings. Finally, it evaluates a specific prediction algorithm—COMPAS—to explore its role within the criminal justice system.

A. *History of Prediction Algorithms*

Crime prediction has been a feature of the criminal justice system for centuries.²² During the 1960s and 1970s, crime-prediction research focused on identifying dangerousness in individuals who committed violent crimes.²³ The prediction of dangerousness, promoted during the “selective incapacitation” movement, proved complex,

¹⁵Surden, *supra* note 13, at 89.

¹⁶Meserole, *supra* note 12.

¹⁷*Id.*

¹⁸*Id.*

¹⁹*Id.*

²⁰Meserole, *supra* note 12.

²¹Danielle Kehl, Priscilla Guo & Samuel Kessler, *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School 2–3 (2017).

²²Thomas Mathiesen, *Selective Incapacitation Revisited*, 22 *LAW & HUM. BEHAV.* 455, 458 (1998).

²³Kehl, *supra* note 22, at 4.

producing a substantial number of false positives and resulting in many individuals being mistakenly labeled as dangerous.²⁴

A major shift in sentencing theory occurred during this era, advancing the idea of punishing individuals not solely for past conduct but for what they might do in the future—a concept now known as recidivism prediction.²⁵ Peter Greenwood and Allan Abrahamse conducted a large study of 2,100 male prison and jail inmates in California, Michigan, and Texas to compile self-reports intended to predict future behavior.²⁶

Because no formal test existed, Greenwood and Abrahamse constructed a “predictive scale” that classified individuals based on risk: “low-rate,” “medium-rate,” and “high-rate.”²⁷ The scale was moderately accurate for predicting low-rate offenders but highly inaccurate for high-rate offenders, generating both false positives and false negatives.²⁸ False negatives occurred when individuals predicted to be low-rate offenders reoffended, while false positives occurred when individuals inaccurately labeled as high-risk were incarcerated for offenses they did not commit—errors that threatened individual liberty.²⁹

Modern RNA prediction algorithms emerged alongside a broader American shift from capital and corporal punishment to rehabilitation.³⁰ Under rehabilitative sentencing, individuals received individualized sentences and treatment, with the goal of preparing them for safe reentry into society.³¹ Judges had wide discretion when imposing sentences, which produced disparate sentencing outcomes that disproportionately harmed minority communities, especially because courts lacked uniform guidelines.³²

This broad discretion contributed significantly to sentencing disparities and, by extension, to mass incarceration. For instance, two individuals with nearly identical backgrounds and both convicted of armed robbery could receive starkly different sentences solely because of race.

In response, sentencing reforms in the 1970s and 1980s sought to shift back toward retributivism—the idea that criminal punishment should be based primarily on the crime committed, rather than predictions of future behavior.³³ This shift increased the use of structured sentencing guidelines and created more consistent sentencing practices.³⁴

Retributivism itself, however, has been linked to the rise of mass incarceration because of its punitive structure and its disproportionate effects on minority communities. Today, courts continue to rely on both judicial discretion and retributivist

²⁴*Id.*

²⁵*Id.* at 4.

²⁶*Id.* at 4–5.

²⁷*Id.* at 5.

²⁸*Id.*

²⁹*Id.* at 5.

³⁰*Id.* at 6.

³¹*Id.*

³²*Id.* at 6–7.

³³*Id.* at 6–7.

³⁴*Id.* at 7.

principles, though they increasingly incorporate evidence-based practices—such as RNA algorithms—to reduce discriminatory outcomes.

B. Current Use of Prediction Algorithms

More courts are moving toward evidence-based practices and away from judicial discretion and pure retributivism. Evidence-based practice (EBP) incorporates scientific and quantitative methods to improve sentencing decisions.³⁵ EBP takes an actuarial approach to assessing and treating risk, using the scientific method to predict future behavior.³⁶ In this sense, EBP operates as a risk-prediction tool, evaluating an individual’s likelihood of recidivism to inform release and detention recommendations.³⁷

EBP has faced some of the same criticisms as judicial discretion and retributivism—namely, that it may contribute to mass incarceration. But it is often praised because, unlike earlier models, EBP weighs factors that increase recidivism risk alongside factors that reduce that risk.³⁸ Treatment considerations are integrated as part of the assessment.³⁹

Today’s RNA tools are considered “fourth-generation” instruments.⁴⁰ They incorporate both static and dynamic factors to evaluate individuals according to their specific risk characteristics while using a systematic and comprehensive method to measure recidivism.⁴¹ Dynamic factors are those an incarcerated individual can change; static factors cannot be changed.⁴² Modern RNA instruments incorporate machine learning to generate recidivism predictions and are used in prison rehabilitation programs, pretrial risk assessments, and—most relevant here—sentencing.⁴³

C. COMPAS: An Algorithm Used in the Criminal Justice System for Sentencing

Judges generally use a two-step process to determine a sentence. First, a judge decides what form of punishment to impose.⁴⁴ If the judge chooses incarceration or probation, they must then determine the appropriate duration of that punishment.⁴⁵

³⁵Kehl et al., *supra* note 22, at 7.

³⁶*Id.*; see also U.S. Dep’t of Just., *What Is Risk Assessment?*, <https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment#gcov4e> (last visited Nov. 8, 2023) (explaining that actuarial instruments “systematically quantify an individual’s risk of reoffending”).

³⁷Christopher T. Lowenkamp & Jay Whetzel, *The Development of an Actuarial Risk Assessment for U.S. Pretrial Services*, 73 FED. PROB. J. 70, 71 (2009).

³⁸Kehl et al., *supra* note 22, at 8.

³⁹*Id.*

⁴⁰*Id.* at 9.

⁴¹*Id.*

⁴²Amy B. Cyphert, *Programming Recidivism: The First Step Act and Algorithmic Prediction of Risk*, 51 SETON HALL L. REV. 331, 349 (2020).

⁴³*Id.*

⁴⁴Kehl et al., *supra* note 22, at 13.

⁴⁵*Id.*

In making this decision, judges often consider deterrence, incapacitation, punishment, and rehabilitation.⁴⁶

Individuals who are predicted to have a high likelihood of reoffending are generally viewed as poor candidates for rehabilitation and may be considered better suited for incarceration to protect public safety.⁴⁷ Conversely, individuals assessed as low-risk are more likely to receive less-severe punishments and be deemed appropriate for rehabilitation.⁴⁸ But it remains unclear how a judge should use a risk assessment score when making a sentencing decision.⁴⁹

RNA scores provide clearer guidance in the pretrial context—where the only question is whether an individual should be released pending trial.⁵⁰ Sentencing is more complex: it involves both selecting the type of punishment and determining its length.⁵¹ Longer sentences, however, do not reduce a person’s likelihood of recidivism—they simply keep the individual incarcerated longer, eliminating the opportunity to reoffend while imprisoned.

RNA tools are encouraged—and even endorsed—by the Model Penal Code.⁵² The American Bar Association likewise encourages states to use RNAs to promote public safety and reduce recidivism.⁵³ The U.S. Department of Justice, in contrast, has expressed concerns that risk assessments used in sentencing may produce disparate impacts on poor and marginalized communities, even where well-intentioned.⁵⁴

One widely used risk-need assessment tool is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). Developed by Equivant, COMPAS is designed to provide decisional support to criminal justice practitioners.⁵⁵ COMPAS can be useful when supported by independent factors but should not be the sole basis for a sentencing determination.⁵⁶

The COMPAS RNA is typically included in the Presentence Investigation Report (PSI) prepared by the probation department and shared with the judge and counsel.⁵⁷ Individuals receive three COMPAS scores—pretrial recidivism risk, general recidivism risk, and violent recidivism risk—typically presented on a bar chart.⁵⁸ COMPAS provides a prediction by comparing an individual to a broader data group—not

⁴⁶Model Penal Code § 1.02(2) (describing the purposes of sentencing, including “prevent[ing] the commission of offenses,” promoting “correction and rehabilitation,” and differentiating among offenders for just individualized treatment”); *Model Penal Code*, BLACK’S LAW DICTIONARY (11th ed. 2019).

⁴⁷Kehl et al., *supra* note 22, at 13.

⁴⁸*Id.*

⁴⁹*Id.* at 14.

⁵⁰*Id.*

⁵¹Kehl et al., *supra* note 22, at 14.

⁵²*Id.* at 13.

⁵³*State v. Loomis*, 881 N.W.2d 749, 752 (Wis. 2016), *aff’d*, 137 S. Ct. 2290 (2017).

⁵⁴Letter from Jonathan J. Wroblewski, Dir., Office of Policy & Legislation, U.S. Dep’t of Justice, to Hon. Patti B. Saris, Chair, U.S. Sentencing Comm’n 7 (July 29, 2014), <https://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annualletter-final-072814.pdf> [hereinafter DOJ Letter].

⁵⁵Equivant, *Practitioners Guide to COMPAS Core 1* (Apr. 4, 2019), <https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf>.

⁵⁶See *Loomis*, 881 N.W.2d at 753.

⁵⁷*Id.* at 754.

⁵⁸Equivant, *supra* note 56, at 3, 31.

by determining whether individuals with similar histories are more or less likely to reoffend after release.⁵⁹

In *State v. Loomis*, the PSI explicitly cautioned the judge not to rely solely on the COMPAS score when deciding whether to incarcerate or determine sentence length.⁶⁰ If caution is required every time a COMPAS score is used, how reliable can the information truly be?

III. OFFICERS OF THE COURT'S DEPENDABILITY ON PREDICTION ALGORITHMS

Probation officers, judges, and attorneys depend on RNA tools. All three officers of the court use the outcome from COMPAS or another RNA tool to reach a recommendation or conclusion for an offender's sentencing hearing. Even though each officer uses the tool differently, the impact on the offender can be harmful. Part A discusses probation officers' dependability on RNA prediction tools and how they may be overly reliant on them. Part B examines judges' dependability on RNA prediction tools and why they should be more skeptical. Part C considers attorneys' use of RNA prediction tools and how that use differs for prosecutors and defense counsel.

A. Probation Officers' Dependability on Algorithms

Probation officers interview offending individuals, their families, and employers on a wide range of topics; create and submit reports to the court; and supervise individuals who have recently been released.⁶¹ Within this role, they can "conduct a standardized risk and needs assessment" to help develop a recommendation submitted to the court for the judge, prosecution, and defense counsel to review.⁶² This Note focuses on the presentence investigation report (PSI) that probation officers prepare.⁶³

The PSI has been the "central source of information to judges since the 1920s."⁶⁴ It is likely to remain an essential component of the criminal justice system, even as its purpose has shifted from providing information about a defendant's personal history and criminal conduct for individualized sentencing to being more offense-focused and less individualized.⁶⁵ This shift stems from increasingly punitive sentencing policies and the adoption of sentencing guidelines.⁶⁶

Typical PSI elements include: the offense, criminal history, sentencing options, offender characteristics, fines and restitution, factors that might warrant departure

⁵⁹See Northpointe COMPAS Risk Assessment, <https://www.equivant.com> (providing overview of COMPAS scoring).

⁶⁰*Loomis*, 881 N.W.2d at 754.

⁶¹Probation Officer EDU, <https://www.probationofficeredu.org/probation-officer/what-is-a-probation-officer/> (last visited Dec. 12, 2023).

⁶²*Id.*; Powers and Duties of Probation Officers, W. Va. Code § 62-12-6(a)(2) (2022).

⁶³See *State v. Loomis*, 881 N.W.2d 749, 754 (Wis. 2016).

⁶⁴Center on Juvenile & Criminal Justice, *The History of the Presentence Investigation Report*, https://www.cjcj.org/media/import/documents/the_history.pdf.

⁶⁵*Id.*

⁶⁶*Id.*

from sentencing guidelines, and sentencing recommendations.⁶⁷ Probation officers interview the offender; their family, employer, and associates; the prosecutor; treatment providers; and others who can attest to the offender's character.⁶⁸ They also review court dockets, criminal history records, indictments, employment records, financial records, and other documents to develop a complete picture of the offender.⁶⁹ The PSI then synthesizes information from these interviews and documents into one report that aims to offer a holistic view of the offender.

The entire PSI is important, but this Note argues that the sentencing recommendation is the most crucial section. The recommendation portion provides the sentencing guidelines applicable to the offense(s) and offers a recommendation for whether to depart based on what has been learned about the offender. Probation officers process an enormous amount of information to develop this recommendation, much of which overlaps with the information that feeds COMPAS's recidivism predictions.

This overlap raises a key question: if a probation officer already conducts such intensive preparation and has direct, qualitative insight into the offender's life, could that officer independently reach a prediction about the offender's likelihood of reoffending—perhaps as accurately, or even more accurately, than COMPAS? If so, the necessity of a COMPAS assessment becomes doubtful. The more probation officers can, through their training and experience, approximate or improve upon the algorithm's predictions, the less justification there is for judges or probation officers to depend on COMPAS.

B. Judges' Dependability on Algorithms

There are countless ways that AI and machine learning can enter the courtroom and require judicial involvement.⁷⁰ This Note focuses on the third role: judges deciding whether to receive or rely on AI-generated outputs—such as COMPAS scores—to inform sentencing decisions.⁷¹

⁶⁷*Id.*

⁶⁸See *Presentence Investigation*, U.S. Probation & Pretrial Servs., N. Dist. of W. Va., <https://www.wvnp.uscourts.gov/presentence-investigation>.

⁶⁹*Id.*

⁷⁰The countless roles judges play when AI and machine learning are brought into the courtroom are:

First, they will serve as evidentiary gatekeepers, applying the Federal Rules of Evidence (or state equivalents) to proffers of testimonial and documentary evidence, including and perhaps especially Rules 401, 402, and 403. Second, judges will serve as guardians of the law, specifically the values embedded in the Bill of Rights as well as statutes and rules of procedure and evidence. Third, judges may serve as potential AI consumers who need to decide whether to receive or rely on AI-generated outputs to inform bail, probation, and sentencing decisions. Fourth, judges will serve as communicators, translating the sometimes complex inputs behind AI into plain-language instructions for jurors and case law precedent for lawyers.

James R. Baker, Laurie N. Hobart & Matthew Mittelsteadt, *An Introduction to Artificial Intelligence for Federal Judges*, Fed. Jud. Ctr. 5, 23 (2023).

⁷¹*Id.*

When a probation officer includes a COMPAS RNA in the PSI, the report contains limitations and written advisements stressing caution.⁷² A judge may not rely solely on COMPAS risk scores to determine whether an offender should be incarcerated or to decide the severity of the sentence.⁷³ The written advisement in *Loomis* explained that:

1. The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighted or how risk scores are determined.
2. Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders—not a particular high-risk individual.
3. Several studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism.
4. A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed. Risk assessment tools must be constantly monitored and reformed for accuracy due to changing populations and subpopulations.
5. COMPAS was not developed for use at sentencing, but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole.

74

These limitations raise serious questions about why any judge would use COMPAS in sentencing at all. The primary argument for using COMPAS and similar tools is that they are “data driven,” and courts sometimes treat errors involving them as subject to harmless-error analysis.⁷⁵ But unless judges understand COMPAS well enough to know how the tool moves from input to output, what data it was trained on, and how it weights each factor, they should hesitate to use it to reach any conclusion—especially a conclusion about whether and how long to incarcerate someone.⁷⁶

C. Attorneys’ Dependability on Algorithms

The roles of prosecutors and defense attorneys are generally adverse at all stages of a criminal case. Occasionally, they work collaboratively—most commonly when negotiating and finalizing a plea agreement. Both sides may invoke a COMPAS assessment to persuade the judge to deviate from the sentencing guidelines.

Absent an agreement, prosecutors are more likely to use COMPAS results to argue for a sentence at the higher or middle end of the guideline range, emphasizing

⁷²See *Loomis*, 881 N.W.2d at 769–70.

⁷³*Id.* at 769.

⁷⁴*Id.* at 769–70.

⁷⁵Baker et al., *supra* note 72, at 76.

⁷⁶*Id.* at 77.

public safety and perceived risk. Defense attorneys, by contrast, tend to use COMPAS to argue for a sentence at the lower end of the range or for alternatives to incarceration, highlighting any factors that suggest a lower risk of recidivism.

Unlike probation officers and judges, attorneys are not required to depend on COMPAS in making their recommendations. Nonetheless, they still use its outputs as advocacy tools—even though those outputs may be biased and may reinforce the same racial and socioeconomic disparities this Note seeks to critique.

IV. PREDICTION ALGORITHMS CONTRIBUTING TO MASS INCARCERATION BECAUSE OF BIAS

Racial disparities are reproduced within prediction algorithms because past data show Black people being incarcerated at higher rates than white people. Algorithmic prediction is only as sound as the data on which the algorithm is trained—a concern captured by the computer-science idiom “garbage in, garbage out.”⁷⁷ The problem begins with biased data being fed into the algorithm, but the resulting disparities reveal significant and measurable racial bias.

There are several forms of potential bias in machine learning. This Note focuses on two: training-data bias and unwritten human bias.⁷⁸ Training-data bias occurs when an algorithm learns from data that already reflect biased outcomes, leading the algorithm to reproduce and reinforce those same patterns.⁷⁹ Inputs can embed bias associated with socioeconomic status, neighborhood location, historical crime statistics, and prosecutorial decisionmaking—all of which distort recidivism predictions.⁸⁰ Unwritten human bias arises when developers’ subjective judgments and implicit biases are baked into the design of the algorithm.⁸¹

If biased data go into the algorithm, biased outputs will follow. This reflects a broader historical pattern: the United States has a long record of biased treatment toward Black Americans. When predominantly white, non-diverse developers—who may not understand the lived realities of incarcerated individuals, racial bias in policing, or structural inequality—design algorithms without those perspectives, their underlying biases can filter into the system.⁸² This is a harm that can only be addressed by incorporating diverse developers and greater transparency into algorithmic processes.

In 2014, Professor Sonja Starr argued that the use of variables correlated with race, gender, or income in sentencing-related risk assessments is constitutionally suspect, and likely contributes to mass incarceration.⁸³ The Department of Justice like-

⁷⁷Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2122, 2224 (2019) (“bias in, bias out”).

⁷⁸James R. Baker et al., *An Introduction to Artificial Intelligence for Federal Judges*, Fed. Jud. Ctr. 32 (2023).

⁷⁹*Id.*

⁸⁰*Id.*

⁸¹*Id.* at 31.

⁸²Rachel Thomas, *Artificial Intelligence Needs All of Us*, TEDx Talk (Oct. 2018), https://www.ted.com/talks/rachel_thomas_artificial_intelligence_needs_all_of_us (“what’s dangerous is a homogenous and exclusive group creating technology that impacts us all”).

⁸³See generally Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014).

wise warned of “the promise and danger of data analytics in sentencing,” emphasizing that RNA prediction tools may “exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”⁸⁴

The United States’ history of racially discriminatory policing and criminal law enforcement has produced higher arrest, prosecution, conviction, and incarceration rates for Black Americans than for white Americans.⁸⁵ As a result, criminal history itself now correlates strongly with race.⁸⁶ Thus, any RNA tool—such as COMPAS—that relies heavily on criminal history will inevitably have a disparate impact on Black communities, particularly Black men. The algorithms are racially biased in the sense that they systematically overstate or understate the average risk for one racial group relative to another.⁸⁷

The bias embedded in the data thus spreads directly to the outcomes and then to the decisions based on those outcomes. In practical terms, COMPAS’s use of racially correlated variables often results in Black men being assigned higher recidivism scores—even when they are no more likely to reoffend than similarly situated white defendants.

The biased data feeding these prediction algorithms and the racially disparate outputs they produce are contributing to mass incarceration. Ensuring the fairness and accuracy of RNA prediction tools is crucial to preventing them from exacerbating racial disparities and deepening the harms already present in the criminal justice system.

V. PRINCIPLES GUIDING PREDICTION ALGORITHMS’ FAIRNESS AND ACCURACY

RNA prediction tools can, in theory, be made fairer and more accurate, which would reduce many of the criticisms they currently face. Countless reforms could advance fairness and accuracy, but two foundational principles are transparency and reliability. COMPAS, for example, has labeled Black defendants as more likely to reoffend than they actually were, while categorizing white defendants as lower risk even when they were more likely to commit new crimes when tested against real-world data.⁸⁸ If a tool like COMPAS were more transparent and more reliable, these errors

⁸⁴Letter from Jonathan J. Wroblewski, Dir., Office of Policy & Legislation, U.S. Dep’t of Justice, to Hon. Patti B. Saris, Chair, U.S. Sentencing Comm’n (July 29, 2014), <https://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annualletter-final-072814.pdf>; see also Eric H. Holder, Jr., Att’y Gen., Remarks at the National Association of Criminal Defense Lawyers 57th Annual Meeting (Aug. 1, 2014), <https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th>.

⁸⁵Mayson, *supra* note 78, at 2229.

⁸⁶*Id.*

⁸⁷William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, Northpointe Inc. (July 8, 2018), https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final.070616.pdf.

⁸⁸U.S. Comm’n on Civil Rights, Connecticut Advisory Comm., *The Civil Rights Implications of Algorithms* 6 (2023) [hereinafter *Civil Rights Implications*], <https://www.usccr.gov/files/2023-04/ct-sac-algorithm-report.pdf> (noting that “maximizing accuracy can lead to unfair or discriminatory predictions while maximizing fairness typically leads to less accuracy”).

and biases would not be as pervasive or as harmful. Transparency and reliability must work together to meaningfully improve predictive tools.

Part A discusses transparency and its importance in the design and use of prediction algorithms. Part B considers the significance of reliability and assesses whether COMPAS can be considered reliable enough to use at sentencing.

A. Transparency

Most machine-learning algorithms are “black boxes.”⁸⁹ Black-box systems lack transparency about how inputs are weighted and how outputs are generated, making it extremely difficult to test, challenge, or meaningfully audit them.⁹⁰ Transparency, by contrast, would allow courts, advocates, and affected individuals to see what information a risk-needs assessment uses and how it uses that information.⁹¹

Equivant’s decision in *State v. Loomis* to invoke trade-secret protection and refuse disclosure of its COMPAS algorithm is particularly troubling. By shielding core details of its proprietary model, Equivant made it impossible to fully evaluate COMPAS for fairness or accuracy.⁹² Any automated decisionmaking tool used in criminal cases should, at a minimum, permit the institutions relying on it to understand how the system reaches its conclusions.⁹³ Greater transparency would enable external audits, facilitate independent validation, and help improve accuracy over time.⁹⁴

Transparency is also a core value of “technological due process.”⁹⁵ Individuals should have the “right to inspect, correct, and dispute inaccurate data and to know the source of the data” used in scoring them.⁹⁶ While companies like Equivant have legitimate interests in protecting their trade secrets, those interests should not override defendants’ due process rights.⁹⁷

Protective orders offer one way to ease the tension between vendors’ trade secret claims and defendants’ rights. Courts could order disclosure of source code, documentation, or design details under strict conditions—such as limiting access to neutral

⁸⁹David Lehr & Paul Ohm, *Playing with the Data: What Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 656 (2017); see, e.g., Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 6 (2014) (defining black boxes as systems that convert inputs to outputs “without revealing how they do so”); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 660 (2017) (explaining that one approach to accountability treats the system as a black box whose inputs and outputs are visible while its inner workings remain unseen); John Villasenor & Virginia Foggo, *Artificial Intelligence, Due Process, and Criminal Sentencing*, 2020 MICH. ST. L. REV. 295, 339 (defining auditability as preserving information used in a risk assessment so it can be accessed if the assessment is challenged).

⁹⁰AI in the Criminal Justice System, Elec. Priv. Info. Ctr., <https://epic.org/issues/ai/ai-in-the-criminal-justice-system/> (last visited Nov. 10, 2023).

⁹¹Villasenor & Foggo, *supra* note 90, at 339.

⁹²Kehl et al., *supra* note 22, at 33.

⁹³Kroll et al., *supra* note 90, at 656.

⁹⁴Kehl et al., *supra* note 22, at 32.

⁹⁵Citron & Pasquale, *The Scored Society*, 90 WASH. U. L. REV. 1, 20 (2014) (explaining that technological due process seeks to ensure meaningful opportunities to challenge algorithmic decisions); Kehl et al., *supra* note 22, at 32 (describing audit trails as a key mechanism of technological due process).

⁹⁶Citron & Pasquale, *supra* note 96, at 20; Kehl et al., *supra* note 22, at 32.

⁹⁷Villasenor & Foggo, *supra* note 90, at 343–44.

experts and prohibiting copying or external distribution—similar to what occurs in patent litigation.⁹⁸ This approach would allow courts and experts to evaluate whether a tool is biased or unreliable without enabling competitors to freely copy the model.

A persistent concern is cost: protective-order litigation can be expensive. Rather than placing that burden solely on companies or on individual defendants, civil rights and criminal justice organizations could play a role in supporting or coordinating challenges to opaque tools.⁹⁹ Successful challenges could benefit not only a particular defendant but everyone subjected to the same algorithm.

Vendors often argue that transparency risks “gaming” the system—allowing individuals who understand the algorithm to strategically shape inputs to obtain more favorable scores.¹⁰⁰ In theory, a person who knows how an algorithm works could tailor responses to reduce their assessed risk. But this concern does not justify the current level of secrecy. Vendors can mitigate gaming by using verifiable data sources, minimizing overly subjective inputs, and designing models that emphasize stable, externally corroborated information.¹⁰¹

Transparency has limits. Even when source code and model logic are available, any decision process that incorporates randomness or complex statistical procedures may produce outcomes that cannot be easily replicated or fully explained.¹⁰² For example, a lottery based on a random-number generator is transparent in its design but yields nonreproducible and unpredictable results by construction.¹⁰³ Transparency alone thus cannot guarantee fairness or accuracy; it must be paired with auditability, accountability, and substantive evaluation standards.¹⁰⁴

A further limitation is enforcement. Even if an algorithm is transparent, it may remain difficult to hold its creators accountable without legal frameworks that impose obligations and consequences.¹⁰⁵ Government algorithms are increasingly subject to transparency through tools like the Freedom of Information Act, but many privately developed tools used in criminal cases face no comparable public-disclosure regime.¹⁰⁶

Despite these limitations, greater transparency—especially for courts, probation departments, and defense counsel—would significantly reduce the risks associated with RNA tools. If decisionmakers can understand how an algorithm works, what data it uses, and how it weighs that data, there will be fewer unresolved concerns about fairness, bias, and due process.

⁹⁸See Villasenor & Foggo, *supra* note 90, at 344; Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1349–50 (2018) (noting that plaintiffs commonly gain access to defendants’ source code and other trade-secret information in civil litigation).

⁹⁹Villasenor & Foggo, *supra* note 90, at 345.

¹⁰⁰Kroll et al., *supra* note 90, at 658.

¹⁰¹See Villasenor & Foggo, *supra* note 90, at 346.

¹⁰²Kroll et al., *supra* note 90, at 659.

¹⁰³*Id.* (using a lottery to illustrate how a transparent algorithm can still produce nonreproducible outputs).

¹⁰⁴Kehl et al., *supra* note 22, at 657–58.

¹⁰⁵*Civil Rights Implications*, *supra* note 89, at 12.

¹⁰⁶*Id.* (noting that although FOIA can shed light on government use of AI, sentencing decisions are often shielded from full transparency because judges retain formal discretion and algorithms are framed as merely advisory).

B. Reliability

Reliability addresses whether a prediction tool produces consistent, stable outputs when presented with similar inputs. In other words, reliability seeks to prevent substantially different predictions for similarly situated individuals simply because of when or how the tool is run.¹⁰⁷ Concerns arise when two defendants with nearly identical characteristics receive different risk scores because one was assessed months later, after the algorithm had been updated or retrained. Because many machine-learning systems are continuously updated as they ingest new data, their outputs can subtly shift over time.¹⁰⁸

Reliability is central to evaluating whether the principles and methods underlying expert testimony are sound.¹⁰⁹ Although the Federal Rules of Evidence do not formally apply at sentencing, their standards for reliability are instructive for evaluating prediction algorithms. *Daubert v. Merrell Dow Pharmaceuticals, Inc.* is particularly relevant. In *Daubert*, the Supreme Court rejected the older *Frye* test and adopted a more flexible reliability-centered approach for scientific expert testimony.¹¹⁰ Rule 702 allows an expert to testify when their “scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue,” and when the testimony is based on sufficient data, reliable methods, and a reliable application of those methods.¹¹¹

Daubert emphasized a two-part inquiry: whether the expert’s testimony constitutes “scientific knowledge,” and whether it will assist the trier of fact in resolving an issue in the case.¹¹² This requires a preliminary assessment of “whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue.”¹¹³ The Court identified several nonexclusive factors: (1) whether the theory or technique can be and has been tested; (2) whether it has been subjected to peer review and publication; (3) the known or potential error rate; (4) the existence and maintenance of standards controlling the technique’s operation; and (5) the degree of general acceptance in the relevant scientific community.¹¹⁴

Applying these factors to COMPAS yields a troubling picture. First, while COMPAS is testable in theory, its lack of transparency makes thorough, independent evaluation difficult.¹¹⁵ Second, COMPAS has been the subject of limited peer-reviewed research, and much of what is publicly available comes from outside analyses rather than from Equivant itself.¹¹⁶

¹⁰⁷Villasenor & Foggo, *supra* note 90, at 347; Molnar, *supra* note 14, at 3.1 (defining reliability as ensuring that “small changes in the input do not lead to large changes in the prediction”).

¹⁰⁸Villasenor & Foggo, *supra* note 90, at 347.

¹⁰⁹See Fed. R. Evid. 702.

¹¹⁰*Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 586 (1993).

¹¹¹Fed. R. Evid. 702; see *Daubert*, 509 U.S. at 589.

¹¹²*Daubert*, 509 U.S. at 592.

¹¹³*Id.* at 592–93.

¹¹⁴*Id.* at 593–94.

¹¹⁵See *Civil Rights Implications*, *supra* note 89, at 6.

¹¹⁶See generally Han-Wei Liu, Ching-Fu Lin & Yu-Jie Chen, *Beyond State v. Loomis: Artificial Intelligence, Government Algorithmizing, and Accountability*, 27 INT’L J.L. & INFO. TECH. 122 (2019).

Third, the error rate associated with COMPAS is not fully known, but available evidence highlights troubling patterns of false positives and false negatives. Studies have shown, for example, that Black defendants are more likely to be labeled high-risk but not subsequently rearrested, while white defendants labeled low-risk are more likely to reoffend.¹¹⁷ As Professor Mayson notes, the result is a mathematically embedded error structure that disproportionately misclassifies Black defendants as high-risk.¹¹⁸

Fourth, there are no widely accepted external standards governing how a tool like COMPAS must operate. Although the model formally excludes race as an explicit input, the continued presence of racial disparities in its outputs signals deep structural problems in how it uses racially correlated variables.¹¹⁹ Fifth, while COMPAS has achieved some degree of real-world adoption, its limited transparency and well-documented disparities undermine its acceptance as a scientifically reliable tool.

Taken together, these concerns indicate that COMPAS and similar RNA tools currently fall short of the reliability that should be demanded of instruments used to inform sentencing decisions. Their error rates, lack of transparent standards, and racially disparate impacts counsel strongly against relying on them—especially when liberty is at stake.

VI. RECOMMENDATIONS ON HOW PREDICTION ALGORITHMS CAN BE LESS HARMFUL IN THE CRIMINAL JUSTICE SYSTEM

Prediction algorithms are very harmful to Black men by configuring outcomes that disproportionately state Black men are more likely to re-offend than their white counterparts who re-offend. Companies could do plenty of things to create less harmful algorithms, but this Note will only focus on two. Part A will discuss how the people working on the algorithms are just as important as the outcomes of the algorithms. Part B will discuss a possible solution of using the algorithm for another purpose, such as assisting incarcerated individuals to reenter society.

A. *Incorporating Unlikely People in the Development of Algorithms*

One way to step away from biased inputs and outcomes is to bring people of diverse backgrounds to help build and modify the algorithm. Rachel Thomas said it best: “we need people from communities that have been disproportionately targeted with harassment to help create our technology because of their understanding of how tech can be weaponized against the most vulnerable and of what safeguards we need to put in place.”¹²⁰

The current people working on COMPAS should keep working on it, while Black men and women who have and have not been incarcerated should join in as well. Also,

¹¹⁷*Civil Rights Implications*, supra note 89, at 6; 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>.

¹¹⁸Mayson, supra note 78, at 2234.

¹¹⁹Angwin et al., supra note 121.

¹²⁰Thomas, supra note 84.

white men and women who have been incarcerated should join. Incorporating the different perspectives and experiences will help input less biased data and produce fewer false positives and negatives across the board. It may be beneficial to have those who have worked as attorneys, probation officers, and anyone who has used the algorithm to get a prediction to identify gaps.

Incorporating these different perspectives and experiences will not fix all the bias, reliability, transparency, or racial disparity issues. Still, it will help fill the gaps to be more accurate and fairer so there can be fewer tradeoffs. No one expects all algorithms to be 100% transparent, bias-free, reliable, and free of racial disparities. Many people, especially offending individuals in their sentencing hearings, want to know they are not being judged by the color of their skin—one of many factors that cannot change.

B. Machine Learning Helping Incarcerated People Reenter Society

Bringing in more people who can help rectify the harm COMPAS has already done to many Black men is one way of combatting mass incarceration. Another is using the same or similar algorithm to assist incarcerated individuals in re-entering society. One reason incarcerated individuals re-offend is that they do not have a home, money, food, or other necessities everyone deserves.

If there were an algorithm that could help incarcerated individuals find a job, housing, get any form of identification, government assistance if needed—anything that is ten times more difficult to do once you are a convicted felon—there would be no need for a recidivism prediction algorithm if incarcerated individuals did not have to start at square one once released.

This algorithm would not necessarily be needed in the courtroom, but it could help judges decide sentencing because they know individuals will not be homeless or without a job once released. First, one of the outputs can be jobs that hire convicted felons within the area where their probation officer will be. This can alleviate the stress of finding a job after already being released.

Second, another output can be the likeliness of an incarcerated individual getting accepted for government assistance and a way to complete the process without getting the runaround. A potential benefit could be submitting the paperwork while being in custody so benefits could start once they are released or not too long after.

Finally, housing complexes that are willing to rent to convicted felons in various areas and the likelihood they will be accepted based on their crime and potential job can reduce the overpopulation of halfway houses.

I am not a developer, but I know there are plenty of ways companies, including COMPAS, can change their algorithm to be beneficial and combat mass incarceration rather than contributing to it as they are now.

CONCLUSION

AI is something that can genuinely help a lot of people in their daily lives, work, and even hobbies. The downside of AI is that subfields, such as machine learning within prediction algorithms, need to be more accurate, fairer, transparent, and reli-

able. Instead, they are constantly being used with those issues contributing to mass incarceration. We must decrease the biased data input into the algorithm to limit the biased outcomes contributing to mass incarceration. Probation officers, judges, and attorneys rely heavily on recidivism algorithms that could be better and are biased against some offending individuals more than others. I would not want a prediction algorithm to predict anything for my future. There are ways we can make the prediction algorithms better to be used in the courts and ways for them to be used to help incarcerated people reenter society.