Automating Judicial Discretion: How Algorithmic Risk Assessments in Pretrial Adjudications Violate Equal Protection Rights on the Basis of Race

Christopher Thomas
John Jay College of Criminal Justice/CUNY Graduate Center

Antonio Pontón-Núñez
John Jay College of Criminal Justice/CUNY Graduate Center

Follow this and additional works at: https://lawandinequality.org/

Automating Judicial Discretion: How Algorithmic Risk Assessments in Pretrial Adjudications Violate Equal Protection Rights on the Basis of Race

Christopher Thomas† and Antonio Pontón-Núñez††

Abstract

Many American jurisdictions use algorithmic risk assessments when setting bail or deciding whether to detain criminal defendants before trial. Although the use of risk assessments has been touted as a reform to protect public safety and reduce bias against defendants, algorithmic risk assessments' opacity and racialized recommendations present serious concerns. This Article examines whether algorithmic risk assessments used during pretrial adjudications violate Fourteenth Amendment Equal Protection rights on the basis of race. The Article begins with an overview of algorithmic risk assessments in the pretrial justice system, focusing on the history of their implementation and how they work. The Article then examines the limited judicial opinions on the constitutionality of these risk assessments. Next, the Article analyzes pretrial algorithmic risk assessments with respect to Equal Protection rights, arguing that they facially discriminate on the basis of race. Additionally, the Article argues that these risk assessments result in disparate treatment of members of this protected class because of one of three types of intentional discrimination: deliberate indifference to racial targeting, discriminatory animus from algorithm designers, or discriminatory intent from the algorithm itself under a proposed theory of partial legal capacity for artificial intelligences. Finally, the Article contends that the use of algorithmic risk assessments is not narrowly tailored, and in many pretrial contexts the state cannot meet its burden of proving that the algorithms are narrowly tailored, due to their opacity. The Article concludes with a discussion of promising and more equitable alternative approaches to pretrial justice.

---

†. Department of Criminal Justice, John Jay College of Criminal Justice/CUNY Graduate Center.

††. J.D. City University of New York School of Law; B.A. Universidad de Puerto Rico.

Table of Contents

Introduction

Each year, over half a million people in the United States are held in local jails before trial, even though they have not been convicted of a crime.[1] Although pretrial preventive detention for public safety has been legally sanctioned since *United States v. Salerno* was decided in 1987, most legally innocent people in pretrial detention are held not for public safety reasons, but due to a racially differential inability to make cash bail.[2] These stints in

---

1. *See* Wendy Sawyer & Peter Wagner, *Mass Incarceration: The Whole Pie 2020*, PRISON POL'Y INITIATIVE (Mar. 24, 2020), https://www.prisonpolicy.org/reports/pie2020.html [https://perma.cc/EN4P-BVWW].

2. *See* United States v. Salerno, 481 U.S. 739 (1987); Christine Scott-Hayward

pretrial detention, however brief, have been found to worsen case outcomes and lead to job losses, housing disruptions, family problems, or other damages.[3] In that context, many state and local jurisdictions have adopted the use of predictive analytics as part of pretrial justice reform in recent years.[4] These tools use computational algorithms[5] to evaluate a criminal defendant's risk of rearrest before trial or failure to appear in court. Defendants are assigned a "risk score" ranging from low to high that judges use

& Sarah Ottone, *Punishing Poverty: California's Unconstitutional Bail System*, 70 STAN. L. REV. ONLINE 167, 168 n.6, 170 (2018) [https://perma.cc/C9AS-3DRR]; Will Dobbie, Jacob Goldin & Crystal S. Yang, *The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges*, 108 AM. ECON. REV. 201, 201–02 (2018), https://www.princeton.edu/~wdob bie/files/bail.pdf [https://perma.cc/KYP6-F9Y9]; *see generally* MIKAELA RABINOWITZ, INCARCERATION WITHOUT CONVICTION: PRETRIAL DETENTION AND THE EROSION OF INNOCENCE IN AMERICAN CRIMINAL JUSTICE (2021) (highlighting how not being able to make cash bail leads to guilty pleas and in-turn mass incarceration); CHRISTINE S. SCOTT-HAYWARD & HENRY F. FRADELLA, PUNISHING POVERTY: HOW BAIL AND PRETRIAL DETENTION FUEL INEQUALITIES IN THE CRIMINAL JUSTICE SYSTEM (2019) (discussing the social and economic ramifications of not being able to make bail).

3. *See* Natalie Goulette & John Wooldredge, *Collateral Consequences of Pretrial Detention*, *in* HANDBOOK ON THE CONSEQUENCES OF SENTENCING AND PUNISHMENT DECISIONS 271, 278–81 (Beth M. Huebner & Natasha A. Frost eds., 2018) (reviewing prior research on the effects of pretrial detention); *see also* Sara Wakefield & Lars Højsgaard Andersen, *Pretrial Detention and the Costs of System Overreach for Employment and Family Life*, 7 SOCIO. SCI. 342 (2020) (demonstrating the effect of pretrial detention on jobs and family); Christopher Thomas, *The Racialized Consequences of Jail Incarceration on Local Labor Markets*, RACE & JUST., May 2022, at 1, 11–14 (demonstrating that pretrial detention has racialized negative effects on local labor markets); Christopher M. Campbell, Ryan M. Labrecque, Michael Weinerman & Ken Sanchagrin, *Gauging Detention Dosage: Assessing the Impact of Pretrial Detention on Sentencing Outcomes Using Propensity Score Modeling*, J. CRIM. JUST., Aug. 2020, at 1, 9–10 (2020) (finding that people detained pretrial are about twice as likely to be sentenced to prison as people released pretrial).

4. *See* Sarah L. Desmarais, Samantha A. Zottola, Sarah E. Duhart Clarke & Evan M. Lowder, *Predictive Validity of Pretrial Risk Assessments: A Systematic Review of the Literature*, 48 CRIM. JUST. & BEHAV. 398, 398 (2021); Sharad Goel, Ravi Shroff, Jennifer Skeem & Christopher Slobogin, *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment*, *in* RESEARCH HANDBOOK ON BIG DATA LAW 9, 9 (2021); Lila Kazemian, Candace McCoy & Meghan Sacks, *Does Law Matter? An Old Bail Law Confronts the New Penology*, 15 PUNISHMENT & SOC'Y 43, 45–46 (2013).

5. *See* THOMAS H. CORMEN, CHARLES E. LEISERSON, RONALD L. RIVEST & CLIFFORD STEIN, INTRODUCTION TO ALGORITHMS 5 (3d ed. 2009) (defining algorithms broadly as "any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output"); *see also* Mariano-Florentino Cuéllar & Aziz Z. Huq, *Toward the Democratic Regulation of AI Systems: A Prolegomenon* 1, 5 (Univ. of Chi. Pub. Law Working Paper No. 753, 2020) (distinguishing between AI technology, defined as "technology relying on computing algorithms to discern patterns in data, and then trigger actions or recommendations in response," and the more legally pertinent concept of AI systems, defined as "a sociotechnical embodiment of public policy codified in an appropriate computational learning tool and embedded in a specific institutional context").

when making release or bail decisions. Theoretically, these algorithmic risk assessment tools (hereinafter RATs) reduce the burden of work for courts, reduce biases among judges and court officials, and make more accurate predictions about defendant "riskiness."[6] However, compelling objections have been raised about the use of these algorithmic RATs in the criminal justice system generally, calling into question whether these tools are the best way to achieve these goals.[7]

Almost all states have adopted the use of RATs at some stage of the criminal justice process, ranging from arrest to parole.[8] Currently over eighteen states and dozens of other local jurisdictions have enacted legislation mandating the use of crime RATs in pre- and post-trial stages.[9] Yet, an extremely limited number of courts have opined on the novel issue of whether algorithm-based RATs in the criminal justice system violate a defendant's constitutional rights.[10]

The rapid adoption of algorithmic RATs in the criminal justice system has already prompted legal and policy debate over issues of

6. *See* Evan M. Lowder, Carmen L. Diaz, Eric Grommon & Bradley R. Ray, *Effects of Pretrial Risk Assessments on Release Decisions and Misconduct Outcomes Relative to Practice as Usual*, 73 J. OF CRIM. JUST. 1, 1–2 (2021); N.Y. Crim. Pro. Law § 510.30; Michael Rempel & Tia Pooler, *Reducing Pretrial Detention in New York City*, 23 SISTEMAS JUDICIALES 1, 3 (2020) (noting that riskiness is usually conceptualized with respect to public safety, but a few states, such as New York, only allow one legal justification for pretrial detention: risk of not attending future court appearances).

7. *See* Aziz Z. Huq, *Artificial Intelligence and the Rule of Law* (U. Chi., Pub. L. Working Paper No. 764, 2021); Tim O'Brien, *Compounding Injustice: The Cascading Effect of Algorithmic Bias in Risk Assessments*, 13 GEO. J. L. & MOD. CRITICAL RACE PERSP. 39, 41 (2021); Gina M. Vincent & Jodi L. Viljoen, *Racist Algorithms or Systemic Problems? Risk Assessments and Racial Disparities*, 47 CRIM. JUST. & BEHAV. 1576, 1577 (2020).

8. Melissa Hamilton, *We Use Big Data to Sentence Criminals. But Can the Algorithms Really Tell Us What We Need to Know?*, GOV'T TECH. (Jun. 6, 2017), https://www.govtech.com/data/We-Use-Big-Data-to-Sentence-Criminals-But-Can-the-Algorithms-Really-Tell-Us-What-We-Need-to-Know.html   [https://perma.cc/MYF3-MZRJ].

9. *See* Brief for the United States as Amici Curiae, Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (No. 16-6387) 2017 WL 2333897; *AI in the Criminal Justice System*, ELEC. PRIV. INFO. CTR., https://epic.org/algorithmic-transparency/crim-justice/ [https://perma.cc/L5D7-S9WD] (detailing the use of artificial intelligence in the criminal justice system by state); John Villasenor & Virginia Foggo, *Algorithms and Sentencing: What Does Due Process Require?*, BROOKINGS (Mar. 21, 2019) https://www.brookings.edu/blog/techtank/2019/03/21/algorithms-and-sentencing-what-does-due-process-require/ [https://perma.cc/MS65-AG74] (quoting MODEL PENAL CODE: SENTENCING (AM. L. INST, Proposed Final Draft 2017)).

10. Erin Harbinson, *Understanding 'Risk Assessment' Tools What They Are and the Role They Play in the Criminal Justice System: A Primer*, 75 BENCH & BAR MINN. 14, 16 (2018) (stating only Indiana and Wisconsin "have considered th[e] issue directly . . .").

race and ethnicity and RATs' impact on people of color. For example, some scholars and legal actors assert that these new RATs present a more inclusive, objective, and complete report on defendants, thus reducing potential racial and ethnic biases from judges.[11] Conversely, other experts have raised serious concerns about the use of these tools in the legal field because they are opaque, operate at a massive scale "to sort, target, or 'optimize' millions of people" in racialized ways, and are reinforced by bias-multiplying feedback loops.[12] This raises the question, does the use of RATs in the criminal justice system in fact violate a defendant's constitutional rights? As these tools become more widely adopted in jurisdictions across the United States, critical examination of these nuanced and complex systems needs to guide this new regime of algorithmic pretrial justice that could be imperiling the fundamental rights of people of color in particular.

This Article examines whether algorithmic RATs used during pretrial adjudications violate a criminal defendant's Fourteenth Amendment Equal Protection rights. Particularly, the Article argues that these tools impermissibly use race and ethnicity to calculate a defendant's risk score. Part I presents an overview of algorithmic RATs in the pretrial justice system, focusing on the history of their implementation, how they work, and which jurisdictions have adopted their use so far. Part II examines the limited judicial opinions on the constitutionality of pretrial algorithmic RATs. Part III analyzes the use of pretrial RATs vis-à-vis Fourteenth Amendment Equal Protection rights. This Section argues that the instruments facially discriminate on the basis of suspect classifications. This Section alternatively argues that pretrial RATs result in disparate treatment of members of these protected classes due to one of three forms of discriminatory intent:

---

11. *See, e.g.*, Adam Neufeld, *In Defense of Risk-Assessment Tools*, MARSHALL PROJECT (Oct. 22, 2017), https://www.themarshallproject.org/2017/10/22/in-defense-of-risk-assessment-tools [https://perma.cc/EP4U-5TAB] (defending the use of algorithmic risk assessment tools and discussing what needs to be done to make them successful); *but cf. More than 100 Civil Rights, Digital Justice, and Community-Based Organizations Raise Concerns About Pretrial Risk Assessment*, LEADERSHIP CONF. ON CIV. & HUM. RTS. (Jul. 30, 2018), https://civilrights.org/2018/07/30/more-than-100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/ [https://perma.cc/FU5T-YJH7] (emphasizing that ending money bail requirements does not necessarily mean pretrial risk assessments are the more equitable solution).

12. CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 12 (2016); *see also* Ben Grunwald & Jeffrey Fagan, *The End of Intuition-Based High-Crime Areas*, 107 CAL. L. REV. 345, 398 (2019) (reviewing how "high-crime feedback loops" are created by racially biased policing).

deliberate indifference to racial targeting, discriminatory animus from algorithm designers, or discriminatory intent from the algorithm itself under a novel theory of partial legal capacity for artificial intelligences. Further, this Section contends that the tools' use is not narrowly tailored, and the state cannot meet its burden of proof of narrow tailoring due to the tools' opacity. Part IV lists important limitations and considerations of an Equal Protection challenge to algorithmic RATs. Lastly, Part V concludes that the tools violate Equal Protection rights and should be banned from use in pretrial adjudications.

I. Overview of Algorithmic Risk Assessments Used in the Pretrial Justice System

Predictive analytic systems have permeated throughout many steps of the criminal justice processes. For instance, many police departments use "hot-spot maps" based on algorithmic risk assessments to strategize deployment of officers and surveillance of specific neighborhoods.[13] Some police departments also use "focused deterrence" approaches to algorithmically identify "high-risk" potential reoffenders within "risky" social networks, who the police then target with either social services or, most commonly, police contact or arrest on low-level crimes as a way to purportedly deter them from committing future crimes.[14] In courts, algorithms are commonly used to create risk assessments of individuals accused of committing a crime.[15] It is contended that these risk assessments help judges and other court officials make important determinations, including whether defendants are "dangerous" to

---

13. *See* Lyria Bennett Moses & Janet Chan, *Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability*, 28 POLICING AND SOC'Y 806, 808 (2018); Laura Myers, Allen Parrish & Alexis Williams, *Big Data and the Fourth Amendment: Reducing Overreliance on the Objectivity of Predictive Policing*, 8 FED. CTS. L. REV. 231, 236 (2015); *see also* Sarah Brayne, *Big Data Surveillance: The Case of Policing*, 82 AM. SOCIO. REV. 977 (2017) (describing how police use algorithmic predictions in practice).

14. *See* Anthony A. Braga, David Weisburd & Brandon Turchan, *Focused Deterrence Strategies and Crime Control: An Updated Systematic Review and Meta-Analysis of the Empirical Evidence*, 17 CRIMINOLOGY & PUB. POL'Y 205, 208–09 (2018); *see generally* ANDREW FERGUSON, THE RISE OF BIG DATA POLICING (2017) (analyzing the effect of big data on policing).

15. *See* Vienna Thompkins, *What Are Risk Assessments — and How Do They Advance Criminal Justice Reform?*, BRENNAN CTR. FOR JUST. (Aug. 23, 2018) https://www.brennancenter.org/our-work/analysis-opinion/what-are-risk-assessments-and-how-do-they-advance-criminal-justice-reform [https://perma.cc/7EDC-TFPU]; CYNTHIA A. MAMALIAN, STATE OF THE SCIENCE OF PRETRIAL RISK ASSESSMENT 10 (2011), https://bja.ojp.gov/sites/g/files/xyckuh186/files/Publications/PJI_PretrialRiskAssessment.pdf [https://perma.cc/5F6T-PVEX].

the community or whether they are likely to reoffend.[16] How do these RATs work? And how did they become a standard practice in the pretrial justice system?

In 1961, a young journalist started the Manhattan Bail Project (which grew into the organization now known as the Vera Institute of Justice), pioneering the use of simple pretrial risk assessments in an effort to reduce discriminatory bias among judges and release more criminal defendants pretrial.[17] The project's "Vera Point Scale" involved an interviewer at arraignment ticking off a checklist of five weighted static factors purportedly associated with failing to appear in court among prior defendants (since flight risk was historically the only legally permissible risk to consider for pretrial detention).[18] Only 1.6% of defendants released using the Vera Point Scale failed to appear, compared to 3% of those released on bail without the Scale.[19] Despite this success, the first wave of the checklist was found to focus too much on local ties to the community, so to accommodate people without such ties who were nonetheless low flight risks, the Vera Point Scale was modified.[20] The subsequent tool was so successful that it quickly spread from New York to other jurisdictions and was used widely for decades.[21]

However, after the punitive turn of the late 1970s and throughout the 1980s,[22] many judges started to explicitly consider

16. *See* MAMALIAN, *supra* note 15, at 18–20. Note that reoffending is usually operationalized as being arrested again before trial, despite the extensive literature documenting racial disparities in arrests, holding constant levels of committing crimes, as discussed below.

17. *See* Scott Kohler, *Vera Institute of Justice: Manhattan Bail Project*, *in* CASEBOOK FOR THE FOUNDATION: A GREAT AMERICAN SECRET 81, 81–82 (2007); *see also* SCOTT-HAYWARD & FRADELLA, *supra* note 2, at 95–96 (explaining standardization promotes both consistency and transparency concerning pretrial release).

18. *See* MARION C. KATSIVE, NEW AREAS FOR BAIL REFORM: A REPORT ON THE MANHATTAN BAIL REEVALUATION PROJECT 32–33 (1968) (noting in the first Vera Point Scale "[t]he defendant is evaluated on the basis of five factors - length of residence in jurisdiction, length of time at present employment, source of support, ties to family in the area in terms of frequency of contact, and prior conviction record").

19. *See* Kohler, *supra* note 17.

20. *See* KATSIVE, *supra* note 18, at app. 3 (showing the later modified checklist focused on a more inclusive set of facts that could be cited in support of bail reduction, which one checklist from that period listed as "family ties verified in court[,]" "[h]as job to return to[,]" "[r]eturn date more than a week away[,]" "[n]o prior record[,]" "[l]ast conviction more than 4 years earlier[,]" "[e]vidence probably won't support conviction[,]" "[a]ge (if over 50)[,]" "[f]emale with dependent children[,]" and "[i]llness or pregnancy").

21. Kohler, *supra* note 17, at 82.

22. *See generally* DAVID GARLAND, THE CULTURE OF CONTROL: CRIME AND SOCIAL ORDER IN CONTEMPORARY SOCIETY (2012) (explaining how social, economic,

"danger" to the community in their pretrial detention determinations.[23] This punitive turn toward a new type of risk became institutionalized after the *Salerno* decision in 1987.[24] In that decade and the 1990s, there was an explosion of criminological research on risk assessments generally and in the pretrial context specifically.[25] Most RATs in this period were simple clinical weighted checklists like the Vera Point Scale, though more complex actuarial pretrial risk assessments were beginning to get developed; yet, by the turn of the twenty-first century, only twelve local jurisdictions were using formal algorithmic RATs in pretrial hearings.[26] Since then, algorithmic pretrial RATs have proliferated.

Most of today's pretrial algorithmic RATs are regression-based—that is, designed to statistically analyze complex interactions of variables to predict how likely a defendant is to either get rearrested before trial or fail to appear in court.[27] Some of them combine administrative data from "court and demographic records with some sort of questionnaire administered by a court official, such as a pretrial services officer . . . ."[28] The tools assign a numerical value and weight based on considerations of static and dynamic factors such as demographic data, criminal history, employment status, level of education, and family background.[29] Weighting of these interacting factors is a particularly important part of the models because it influences the output variable that the model predicts. Nevertheless, some commercial RATs keep this weighting information private due to the proprietary nature of their products.[30] Most of the current pretrial RATs are regression-based,

---

and political forces in the late twentieth century reshaped criminological thought).

23. *See* PRETRIAL JUST. INST., OVERVIEW OF RESEARCH FINDINGS ON PRETRIAL RISK ASSESSMENT AND PRETRIAL SUPERVISION 2–3 (2008).

24. United States v. Salerno, 481 U.S. 739 (1987).

25. *See* PRETRIAL JUST. INST., *supra* note 23, at 5 (finding that "study after study has failed to replicate the findings of previous studies," suggesting that researchers were not converging on a reliable set of predictors).

26. *See* SCOTT-HAYWARD & FRADELLA, *supra* note 2, at 95–96.

27. *See* Matt Henry, *Risk Assessment: Explained*, APPEAL (Dec. 14, 2019), https://theappeal.org/risk-assessment-explained/ [https://perma.cc/QL6Z-FUCU].

28. *See id.*

29. *See id.*; Anna Maria Barry-Jester, Ben Casselman & Dana Goldstein, *The New Science of Sentencing*, MARSHALL PROJECT (Aug. 4, 2015), https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing [https://perma.cc/MAL2-73SG].

30. *See* Henry, *supra* note 27 ("The process of setting the weights is known as training the model. In the machine learning context, this is known as supervised learning. A data set is fed into the model that contains a set of features, such as age, number of prior arrests, etc. . . . . [I]f the model is supposed to predict likelihood of rearrest within two years (as it is in many risk assessment tools), the data used to train the model would have a big data set formatted much like a spreadsheet that

such as COMPAS and the more transparent Arnold Venture's Public Safety Assessment (hereinafter the PSA).[31]

The newest generation of RATs involve machine learning, making them a type of artificial intelligence.[32] Unlike earlier generations of pretrial risk assessments, which relied to varying extents on expert judgment from psychologists, social workers, probation officers, or other justice system actors, machine learning algorithmic RATs do not depend on human judgment. Instead, these algorithms are designed to mimic how humans learn how to solve complex tasks, changing on their own to learn new rules and rationales for decision-making.[33] Designers identify a particular outcome of interest (such as likelihood of arrest before trial), then design algorithms that explore a given dataset and identify complex patterns to make predictions, evolving as they work through more data to get closer to the desired outcome; in supervised machine learning, the algorithms learn how to use training data to replicate a human-identified pattern, whereas in unsupervised machine learning, the algorithms are even more divorced from human oversight, instead teaching themselves some inherent structure in the unlabeled data.[34] The key aspect of machine learning RATs for the purposes of this paper's argument is that the precise ways the algorithms use data points such as race and ethnicity are inherently unknowable because they are not programmed directly by humans. In short, both types of pretrial RATs are "designed to do one thing: take in the details of a defendant's profile and spit out a recidivism

---

associated values of the input variables with a value for the output variable (i.e., prediction).").

31. *See* Tim Brennan, William Dieterich & Beate Ehret, *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAV. 21, 24 (2009) (describing how COMPAS uses "logistic regression, survival analysis, and bootstrap classification methods . . .").

32. *See* Doaa Abu Elyounes, *Bail or Jail? Judicial Versus Algorithmic Decision-Making in the Pretrial System*, 21 COLUM. SCI. & TECH. L. REV. 376, 389 (2020) (discussing a machine learning RAT developed by Professor Jon Kleinberg at Cornell University); *see also* Mariano-Florentino Cuéllar & Aziz Z. Huq, *Privacy's Political Economy and the State of Machine Learning: An Essay in Honor of Stephen J. Schulhofer*, N.Y.U. ANN. SURV. AM. L. 2 (forthcoming) (explaining the increase of machine learning in America after 9/11); Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 89 (2014) (discussing the interplay between artificial intelligence and law).

33. *See* L. Karl Branting, *Artificial Intelligence and the Law from a Research Perspective*, 14 SCITECH LAW. 32, 34–35 (2018).

34. *See* STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE 657 (4th ed. 2021); THOMAS W. MALONE, DANIELA RUS & ROBERT LAUBACHER, ARTIFICIAL INTELLIGENCE AND THE FUTURE OF WORK, MIT Research Brief 17 (Dec. 2020), https://workofthefuture.mit.edu/wp-content/uploads/2020/12/2020-Research-Brief-Malone-Rus-Laubacher2.pdf [https://perma.cc/L23Z-62W5].

score—a single number estimating the likelihood that [the defendant] will reoffend" before trial (or, in some jurisdictions, not show up to court).[35] As explained below, the tools have been criticized for being biased, treating defendants differently on the basis of their race or ethnicity, and heavily influencing a judge's decision-making.

II. Judicial Interpretations Directly Addressing Algorithms Used in Criminal Justice and Other Relevant Settings

Few courts have opined on the novel issue of whether algorithm-based RATs in the criminal justice system violate a defendant's constitutional rights. At the state level, Wisconsin and Indiana are the only two states where their highest courts have addressed the issue directly, generally affirming the use of algorithmic risk assessments for sentencing determinations. Conversely, federal courts have declined to rule on whether the use of predictive analytics violates a criminal defendant's constitutional protections. There is, however, a recent civil case in a Texas district court that may shed some light on the issue. There, a teachers' union brought an action against a school district, alleging that the district's use of an algorithmic evaluation system used to terminate teachers for ineffective performance violated their due process and equal protection rights. The following sections summarize these three cases.

A. State of Wisconsin v. Loomis

In 2013, the State of Wisconsin charged defendant Eric Loomis with five criminal counts for allegedly participating as the driver in a drive-by shooting.[36] While Loomis denied his involvement in that shooting, he ultimately accepted a guilty plea to only two of the lesser charges: "attempting to flee a traffic officer and operating a motor vehicle without the owner's consent."[37] Subsequently, the circuit court ordered a pre-sentence investigation, resulting in a report (hereinafter PSI) that included a risk assessment prepared by COMPAS, a privately-owned algorithmic tool.[38] COMPAS reports only present "risk scores displayed in the form of a bar chart,

---

35. *See* 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/ [https://perma.cc/39FA-A67E].

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

37. *Id.*

38. *Id.*

with three bars that represent pretrial recidivism risk, general recidivism risk, and violent recidivism risk."[39]

These scores are based on the defendant's criminal history and an interview conducted with the defendant. However, the scores are not individualized; they are a standardized prediction of recidivism "based on a comparison of information about the individual to a similar data group."[40] Based in part on these scores, Loomis was sentenced by the trial court to six years in prison followed by five years of extended supervision.[41]

On appeal, Loomis challenged the use of COMPAS at sentencing, alleging it violated his right to due process.[42] Specifically, Loomis argued that (1) "the proprietary nature of COMPAS prevent[ed him] from challenging the COMPAS assessment's scientific validity," (2) COMPAS risk assessments impermissibly take gender into account, and (3) the use of aggregate data to calculate risk scores violated his right "to an individualized sentence."[43] The Wisconsin Supreme Court ultimately affirmed his sentence.[44]

First, the court found that Loomis did not meet "his burden of showing that the circuit court actually relied on gender as a factor in sentencing."[45] Moreover, even if COMPAS did consider gender, the court determined that such a factor is necessary to promote statistical accuracy.[46] The State specifically argued in this regard that "because men and women have different rates of recidivism and different rehabilitation potential, a gender neutral risk assessment would provide inaccurate results for both men and women."[47] Second, the court found that the proprietary nature of the COMPAS algorithm did not infringe upon Loomis's due process rights because COMPAS largely relies on reviewable public data.[48] A practitioner's guide to COMPAS explained that "the risk scores are based largely on static information (criminal history), with limited use of some dynamic variables (i.e., criminal associates,

---

39. *Id.*
40. *Id.*
41. *Id.* at 756 n.18.
42. *Id.* at 757.
43. *Id.* at 753, 757. The Wisconsin Court of Appeals certified these questions to the State's Supreme Court.
44. *Id.* at 754.
45. *Id.* at 767.
46. *Id.*
47. *Id.* at 765.
48. *Id.* at 761.

substance abuse)" and a questionnaire filled by the defendant.[49] In other words, "to the extent that [his] risk assessment is based upon his answers to questions and publicly available data," Loomis "had the opportunity to verify that the questions and answers listed on the COMPAS report were accurate," even though the algorithmic formula, which predicts the score, is unavailable for review.[50] Lastly, the court agreed that COMPAS did use aggregate, unvalidated data to calculate his risk score.[51] Nevertheless, COMPAS risk assessments were not a "determinative factor" in his sentencing.[52] As such, sentencing ultimately relies on the discretion of a judge, which is informed by many factors included in the PSI.

The Wisconsin Supreme Court acknowledged that while sentencing courts may consider COMPAS RATs for sentencing determinations, they may not use risk scores to determine "whether an offender is incarcerated"; "the severity of the sentence"; or "whether an offender can be supervised safely and effectively in the community."[53] In addition, sentencing courts were required to generally explain the factors used to make sentencing decisions. The court further mandated that PSIs containing a COMPAS assessment include a "written advisement listing [its] limitations."[54] The five limitations were: (1) "[t]he proprietary nature of COMPAS . . . prevent[s] disclosure of . . . how risk scores are determined"; (2) because COMPAS only relies on aggregate data, it was unable to identify "a particular high-risk individual"; (3) "[s]ome 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"; and (5) COMPAS was originally intended "for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole."[55]

The Wisconsin Supreme Court, however, acknowledged the possibility of an equal protection challenge based on the use of gender in statistical generalizations. In its reasoning, the court specifically referenced *Craig v. Boren*, where an Oklahoma law was

---

49. *Id.*
50. *Id.*
51. *Id.* at 765.
52. *Id.* at 764–65.
53. *Id.* at 769.
54. *Id.*
55. *Id.* at 769–70.

challenged for prohibiting the sale of 3.2% beer to men under twenty-one years of age and women under eighteen years of age.[56] There, the United States Supreme Court declared that "classifications by gender must serve important governmental objectives and must be substantially related to achievement of those objectives"—a standard that was not met in *Craig*.[57] The *Loomis* court specifically noted the Supreme Court's explanation that sociological and empirical justifications for gender-based classifications may not pass judicial scrutiny because "the principles embodied in the Equal Protection Clause are not to be rendered inapplicable by statistically measured but loose-fitting generalities concerning the drinking tendencies of aggregate groups."[58] Notwithstanding the Wisconsin court's analogy of *Craig v. Boren* to the *Loomis* facts, the court refused to entertain the equal protection challenge because Loomis failed to directly raise it.[59] Accordingly, the court only focused on his due process claims.

In a concurrence, Justice Shirley Abrahamson agreed with the judgment, but stated she would have required sentencing courts to specifically "evaluate on the record the strengths, weaknesses, and relevance to the individualized sentence . . . ."[60] Such explanation was necessary because COMPAS risk assessment had "garnered mixed reviews in the scholarly literature and in popular commentary and analysis."[61] In addition, Justice Abrahamson raised a concern with the "court's lack of understanding of COMPAS . . . ."[62] She took issue with the court's denial of "[COMPAS' then owner] Northpointe's motion to file an amicus brief," since it could have provided critical information about COMPAS.[63]

B. Malenchik v. State

Unlike the *Loomis* court's cautionary allowance of algorithmic risk assessments in sentencing, the Supreme Court of Indiana

---

56. *See id.* at 766; *see also* Craig v. Boren, 429 U.S. 190 (1976).

57. *Craig*, 429 U.S. at 197.

58. *Loomis*, 881 N.W.2d at 766 (quoting *Craig*, 429 U.S. at 208–09).

59. *Id.* ("Notably, however, Loomis does not bring an equal protection challenge in this case. Thus, we address . . . Loomis's constitutional due process right [claims] . . . .").

60. *Id.* at 774 (Abrahamson, J., concurring).

61. *Id.* at 774–75.

62. *Id.* at 774 ("At oral argument, the court repeatedly questioned both the State's and defendant's counsel about how COMPAS works. Few answers were available.").

63. *Id.*

enthusiastically affirmed their use in *Malenchik v. State*.[64] In late 2008, defendant Anthony Malenchik was convicted and sentenced to six years in prison, pursuant to his guilty plea to theft and his admission to being a habitual offender.[65] In preparation for sentencing, the trial court was presented with a PSI indicating that Malenchik "f[ell] into the High Risk/Needs category" and "ha[d] a high probability of having a Substance Dependence Disorder," based on reports created by algorithmic risk assessment instruments, including one named Level of Service Inventory– Revised (hereinafter LSI–R).[66] LSI–R generally measures recidivism by taking into consideration a defendant's "areas of Criminal History, Education and Employment, Financial, Family, Accommodations, Leisure and Recreation, Companions, Alcohol and Drugs, Emotional and Personal Issues, and Attitudes and Orientation," combined with other demographic information.[67] LSI– R is a privately owned algorithmic tool.[68]

The Supreme Court of Indiana granted transfer from the appellate court to resolve the specific issue of whether a trial court may consider, and to what extent, reports from algorithmic risk assessment instruments when making sentencing determinations.

---

64. *See* Malenchik v. State, 928 N.E.2d 564 (Ind. 2010).

65. *See id.* at 566; *see also* Malenchik v. State, 908 N.E.2d 710 (Table), 2009 WL 1577832, *3 (Ind. Ct. App. 2009) (affirming the conviction after Malenchik appealed his sentence, arguing the trial court abused its discretion).

66. *Malenchik*, 928 N.E.2d at 567.

67. *Id.*; *see also* Anthony W. Flores, Christopher T. Lowenkamp, Paula Smith & Edward J. Latessa, *Validating the Level of Service Inventory—Revised on a Sample of Federal Probationers*, 70 FED. PROB. 44, 45 (2006) (citations omitted).

> The LSI-R measures 54 risk and need factors about 10 criminogenic domains that are designed to inform correctional decisions of custody, supervision, and service provision. The theoretically informed predictor domains measured by the LSI-R include criminal history, education/employment, financial situation, family/marital relationships, accommodation, leisure and recreation, companions, alcohol or drug use, emotional/mental health, and attitudes and orientations.
> The LSI-R assessment is administered through a structured interview between the assessor and offender, with the recommendation that supporting documentation be collected from family members, employers, case files, drug tests, and other relevant sources as needed. The total risk/need score produced by the LSI-R is indicative of the number of predictor items (out of 54) scored as currently present for the offender. The LSI-R score is then actuarially associated with a likelihood of recidivism that was derived from the observed recidivism rates of previously assessed offenders. Last, domain scores of the LSI-R are used to identify an offender's most promising treatment targets.

Id.

68. *See* MEGAN E. COLLINS, EMILY M. GLAZENER, CHRISTINA D. STEWART & JAMES P. LYNCH, FOLLOW-UP REPORT TO THE MSCCSP: USING ASSESSMENT INSTRUMENTS DURING CRIMINAL SENTENCING 10 (2015) ("[T]he LSI-R and LS/CMI are proprietary tools offered by Multi-Health Systems Inc.").

Malenchik argued, as relevant here, (1) that "such models have not been recognized as scientifically reliable so as to qualify for admissibility under Indiana Evidence Rule[s]"; (2) "that the scoring models lack objective reliability"; (3) "they are not relevant to statutory aggravating circumstances"; (4) "they are unfairly discriminatory"; (5) "the use of the LSI–R test in this case impinged upon his right to counsel"; (6) "the use of scoring models conflicts with Indiana's constitutional requirement that the penal code be founded on principles of reformation and not vindictive justice"; and (7) "using such scores may lead to an unwise fundamental change in Indiana's sentencing system."[69] The State countered that the algorithmic tools were permissible because they were "employed consistently with [their] proper purposes and limitations."[70] Ultimately, the court found that the use of algorithmic RATs was not unlawful for sentencing decisions because the tools enhance and supplement considerations for judges making such determinations, as opposed to deciding on their own a defendants' sentencing outcome.[71]

As to the objective reliability of the algorithmic instruments, the court repeatedly asserted that scoring models, particularly LSI–R, have "widespread acceptance" and are "widely recognized as valid and reliable" by governmental and scholarly communities.[72] The court assured that these algorithmic tools do not constitute aggravating circumstances, but rather help judges make comprehensive sentencing evaluations.[73] Although Malenchik argued that LSI–R was discriminatory because "a person's family disharmony, economic status, personal preferences, or social circumstances should never bear any weight with a sentencing judge,"[74] the court disagreed. The court instead reasoned that sentencing courts were statutorily mandated to consider these

---

69. *Malenchik*, 928 N.E.2d at 567–68.

70. *Id.* at 568.

71. *Id.* at 573–74.

72. *Id.* at 568–71 (finding that "academic literature has demonstrated for decades [that] objective actuarial risk/needs instruments more accurately predict risk and identify criminogenic needs than the clinical judgment of officers," and these models "are well supported by empirical data and provide target areas to change an individual's criminal behavior, thereby enhancing public safety").

73. *Id.* at 572 ("The nature of the LSI–R is not to function as a basis for finding aggravating circumstances, nor does an LSI–R score constitute such a circumstance. But LSI–R scores are highly useful and important for trial courts to consider as a broad statistical tool to supplement and inform the judge's evaluation of information and sentencing formulation in individual cases.").

74. *Id.* at 574 (internal quotation marks omitted).

factors in PSIs.[75] Addressing the overarching goals and purposes of the algorithmic tools, the court asserted that the tools did not violate the Indiana Constitution because they "provide usable information based on extensive penal and sociological research to assist the trial judge in crafting individualized sentencing schemes with a maximum potential for reformation."[76] As such, the court concluded that algorithmic risk assessments serve an appropriate purpose in line with the current prescribed sentencing objectives and limitations, and would not significantly change the sentencing system.[77]

C. Houston Federation of Teachers, Local 2415 v. Houston Independent School District

No federal court has opined on the issue of algorithmic risk assessments used in the criminal justice system. However, a teacher's union representing over 6,000 members filed a federal civil suit alleging that a privately owned algorithmic tool used by the Houston Independent School District (hereinafter the School District) to terminate teachers for ineffective performance during the 2011–2015 school years violated, in part, their constitutional right to equal protection.[78] The tool, Educational Value–Added Assessment System (hereinafter EVAAS), generally "compar[es] the average test score growth of students taught by the teacher compared to the statewide average for students in that grade or

---

75. *See id.*; IND. CODE § 35-38-1-9(b)(2).

76. *Malenchik*, 928 N.E.2d at 575.

77. Notably, in 2015, 30,347 people were incarcerated in Indiana prisons, where "Black people constituted 10% of state residents, but . . . 34% of people in prison" and "Black people were incarcerated at 2.7 times the rate of [W]hite people . . . ." VERA INST. OF JUST., INCARCERATION TRENDS IN INDIANA 1–2 (2019), https://www.vera.org/downloads/pdfdownloads/state-incarceration-trends-indiana.pdf [https://perma.cc/S5MT-YNZE]. "In 2018, there were 23,844 people in the Wisconsin prison system," where "Black people constituted 7% of state residents, but . . . 41% of people in prison" and "[i]n 2017, Black people were incarcerated at 10.9 times the rate of [W]hite people . . . ." VERA INST. OF JUST., INCARCERATION TRENDS IN WISCONSIN 1–2 (2019), https://www.vera.org/downloads/pdfdownloads/state-incarceration-trends-wisconsin.pdf [https://perma.cc/6WMD-7M2G].

78. Hous. Fed'n of Teachers, Loc. 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1171 (S.D. Tex. 2017). The plaintiffs also claimed procedural and substantive due process violations, which will not be discussed in this Article. *See id.* at 1173 (asserting that plaintiffs raised violations of "1. procedural due process, due to lack of sufficient information to meaningfully challenge terminations based on low EVAAS scores; 2. substantive due process, because there is no rational relationship between EVAAS scores and HISD's goal of employing effective teachers; 3. substantive due process, because the EVAAS system is too vague to provide notice to teachers of how to achieve higher ratings and avoid adverse employment consequences").

course."[79] Specifically, plaintiffs argued that the School District's policy of aligning teachers' instructional performance ratings with EVAAS scores, which "subverts the independence of the instructional practice score," wrongly classified teachers with no rational explanation.[80]

The District Court for the Southern District of Texas admitted that plaintiffs presented a "novel claim" with no controlling precedent in an analogous context.[81] Nevertheless, the court rejected plaintiff's argument, finding that the termination policy was not a classification system.[82] Assuming there was a classification, the court found that EVAAS passed rational basis review under a substantive due process claim—the same standard it would have applied to an equal protection claim. In analyzing the due process claim, the court found that even if the algorithmic tool was imperfect, "the loose constitutional standard of rationality allows governments to use blunt tools which may produce only marginal results."[83] As such, the district court denied summary judgment on the substantive due process claim.[84]

III. An Equal Protection Analysis of Algorithmic Risk Assessments

The Equal Protection Clause of the Fourteenth Amendment declares that "[no] State shall . . . deny to any person within its jurisdiction the equal protection of the laws," which is essentially a direction that "all persons similarly circumstanced shall be treated alike."[85] An Equal Protection claim may be raised when the state facially classifies individuals or when it acts discriminatorily "as applied."[86] Facial classifications are reviewed under tiered levels of scrutiny.[87] "As applied" classifications are reviewed under the same scheme, but claimants must also prove there was a discriminatory

---

79. *Id.* at 1172.

80. *Id.* at 1183.

81. *Id.* ("This appears to be a novel claim, and the court has found no authority addressing an equal protection claim in an analogous context.").

82. *Id.* at 1175.

83. *Id.* at 1182.

84. *Id.* at 1183.

85. U.S. CONST. amend. XIV, § 1; Plyler v. Doe, 457 U.S. 202, 216 (1982) (quoting F.S. Royster Guano Co. v. Virginia, 253 U.S. 412 (1920)).

86. *See* Washington v. Davis, 426 U.S. 229 (1976); Romer v. Evans, 517 U.S. 620 (1996).

87. *Facially Neutral Laws Implicating a Racial Minority*, LIBRARY CONG.: CONST. ANNOTATED, https://constitution.congress.gov/browse/essay/amdt14-S1-4-1-3-1-4/ALDE_00000825/ [https://perma.cc/6A5K-LXYQ].

impact and intent by the state.[88] Race and ethnicity are considered suspect classifications reviewed under strict scrutiny.[89] To trigger this level of scrutiny, members of the class must be treated categorically differently, which this Article argues is the case in the pretrial detention context.[90] Such classifications must be narrowly tailored to meet a compelling government interest. [91] If a court finds that the use of race or ethnicity does not pass muster under its appropriate level of scrutiny, the law or policy is declared unconstitutional.

Here, the overarching question is whether the government is violating criminal defendants' Equal Protection rights by using algorithmic assessments that include race and ethnicity to calculate risk scores used for pretrial determinations. From the *Loomis* and *Houston Federation of Teachers, Local 2415* opinions, it seems very likely that courts will treat algorithm-based classifications within the existing Equal Protection tiered-scrutiny framework.[92] First, we argue there is significant evidence to show that the government facially classifies individuals impermissibly. However, even assuming, *arguendo*, that the risk assessment classifications are facially neutral, we then argue that they have a disparate impact, and that the government intentionally discriminated on the basis of race under one of three types of legal intentionality (deliberate indifference to racial targeting, discriminatory animus from algorithm designers, or discriminatory intent from the algorithm itself). Lastly, we show how the use of algorithmic assessments is not narrowly tailored to meet the government's purported goal of reducing bias in pretrial adjudications and how the government cannot meet its burden of proving RATs are narrowly tailored, due to the opacity of the algorithms' black box mechanisms.

A. Algorithmic Risk Assessments Explicitly Use Suspect Classifications

Race and ethnicity are suspect classifications.[93] An Equal Protection Clause challenge based on these classifications must be

---

88. *Id.*

89. *Id.*

90. *See* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1192–93 (2017); *see, e.g.*, Gratz v. Bollinger, 539 U.S. 244, 251–57 (2003); Grutter v. Bollinger, 539 U.S. 306, 312–16 (2003).

91. *Gratz*, 539 U.S. at 246.

92. *See* State v. Loomis, 881 N.W.2d 749, 754 (Wis. 2016); Hous. Fed'n of Teachers, Loc. 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168 (S.D. Tex. 2017).

93. *See* Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 216 (1995); Regents of

reviewed under strict scrutiny because they "are simply too pernicious to permit any but the most exact connection between justification and classification."[94] Generally, suspect classifications, especially racial classifications, must be used as a "last resort."[95] And in such cases, they must be narrowly tailored to meet the government's stated compelling interest.[96]

Creators of algorithmic RATs deny using suspect classifications in their calculations.[97] For instance, one of the major market competitors selling regression-based RATs is Northpointe, Inc. (now doing business as Equivant). Their algorithmic risk assessment tool COMPAS, the one at issue in Loomis, is used by many states, including New York and California, both of which rank in the top five states with the largest pretrial detainee population.[98] Northpointe firmly denies that COMPAS uses race as a variable, but due to the proprietary nature of their algorithm, Northpointe refuses to reveal its variables.[99] However, there is strong scholarly consensus that algorithmic risk assessments almost all use static factors like race, either explicitly or in other ways.[100]

Univ. of Cal. v. Bakke, 438 U.S. 265, 291 (1978).

94. Wygant v. Jackson Bd. of Educ., 476 U.S. 267, 280 (1986) (quoting Fullilove v. Klutznick, 448 U.S. 448, 537 (1980) (Stevens, J., dissenting)).

95. Bartlett v. Strickland, 556 U.S. 1, 21 (2009) (quoting Richmond v. J.A. Croson Co., 488 U.S. 469, 519 (1989) (Kennedy, J., concurring in part and concurring in judgment)); *see* Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198, 2237 (2016) (quoting *Croson*, 488 U.S. at 519).

96. *Bartlett*, 556 U.S. at 21; *see also Grutter*, 539 U.S. at 326.

97. *See* Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67, 80 (2017) ("[O]rganizations do not intend to discriminate and do not use sensitive classifiers like race and gender."); *see also* SARAH PICARD, MATT WATKINS, MICHAEL REMPEL & ASHMINI KERODAL, BEYOND THE ALGORITHM: PRETRIAL REFORM, RISK ASSESSMENT, AND RACIAL FAIRNESS 5 (2018) (stating "[t]he tool did not explicitly use race or ethnicity in calculating risk scores").

98. *See* Jason Tashea, *Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions*, A.B.A. J. (2017), http://www.abajournal.com/magazine/article/algorithm_bail_sentencing_parole [https://perma.cc/P6FY-FFBJ] (identifying California, Michigan, New York, and Wisconsin as states that have adopted the use of COMPAS); *see also* Wanda Bertman & Alexi Jones, *How Many People in Your State Go to Local Jails Every Year?*, PRISON POL'Y INITIATIVE (Sept. 18, 2019), https://www.prisonpolicy.org/blog/2019/09/18/state-jail-bookings/ [https://perma.cc/R6DA-U6RS] (comparing the number of jailed individuals in various states and the seriousness of their offenses).

99. Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES, 1, 1 (2018), https://advances.sciencemag.org/content/4/1/eaao5580 [https://perma.cc/EE4B-27MV] ("[T]he data used by COMPAS do not include an individual's race . . . .").

100. *See, e.g.*, Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 270–96 (2013); Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT'G REP. 237, 237–43 (2015);

Even assuming that race is not explicitly used in algorithmic RATs, there is substantial evidence proving that many of these tools use other variables as proxies for race.[101] In an American Society of Criminology handbook on risk assessments, for example, the risk assessment scholar Robert Brame concluded that "one of the important lessons of the methodological literature on risk assessment is that leaving variables like race and ethnicity out of [the] recidivism risk assessments guarantees that they will still be there."[102] Similarly, an analysis of an algorithmic risk assessment designed to replicate the PSA (which is used in more than forty jurisdictions) found that the PSA algorithm included information on detainee race via proxy variables, concluding that "there are likely *no* truly [racially] uncorrelated input variables in real-world data, and, as a result, that likely *all* of the commonly used algorithms may violate core principles underlying antidiscrimination law by allowing race to contaminate predictions of risk."[103] The consensus is strong that risk assessments use race either explicitly or implicitly through proxies.

Even if racial proxies are used, laws and policies that employ proxies are commonplace, so the question is whether the proxy acts as a means to an impermissible end.[104] Equal Protection doctrine requires that the government state a legitimate purpose for non-suspect classifications.[105] However, a claimant may challenge the

---

Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2251–54 (2019).

101. *See* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 804–05 (2014) (arguing that, by directing judges to use these algorithmic risk assessments, they are directed to "explicitly consider a variety of variables . . . not just in special contexts in which one of those variables might be particularly relevant (for instance, ability to pay in cases involving fines), but routinely, in all cases. This is not a fringe development"); *see also* Sonja B. Starr, *The Risk Assessment Era: An Overdue Debate*, 27 FED. SENT'G REP. 205, 205–06 (2015) (drawing on scholarship that argues risk factors like prior arrests become proxies for race).

102. Robert Brame, *Static Risk Factors and Criminal Recidivism*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 67, 82 (Faye S. Taxman ed., 2016) (providing a generalized test of this finding using a simulated dataset with known covariance between race and other factors).

103. Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 371 (2020).

104. *See* Deborah Hellman, *Two Types of Discrimination: The Familiar and the Forgotten*, 86 CAL. L. REV. 315, 328 (1998) ("The dominant inquiry of Equal Protection case law is about fit: How tight is the correlation between the trait used in the statute and its purported target?").

105. *See, e.g.*, Bush v. Vera, 517 U.S. 952, 968, 985 (1996) (stating, "to the extent that race is used as a proxy for political characteristics, a racial stereotype requiring strict scrutiny is in operation[,]" and "[o]ur Fourteenth Amendment jurisprudence evinces a commitment to eliminate unnecessary and excessive governmental use and reinforcement of racial stereotypes").

state's purported purpose if the non-suspect classification ultimately serves a non-legitimate end or as a stand-in for a suspect classification, and courts are likely to strike them as unlawful.[106] The Supreme Court has also applied this reasoning within the Fourteenth Amendment Due Process framework.[107]

Algorithmic assessment tools have been found to use some variables as stand-ins for suspect classifications, as data scientists are generally sanctioned from using race and ethnicity altogether. The nonprofit coalition Partnership on AI found that these assessment tools use "imperfect proxies such as crime reports or arrests" to calculate the *likely possibility* of recidivism.[108] Recidivism is measured by these algorithms as whether the defendant is likely to get arrested before trial, rather than whether the defendant will commit a crime, per se.[109] This definition of recidivism, which does not narrowly capture the "public safety" objective in pretrial determinations, is chosen by data scientists because "the target for prediction (having actually committed a crime) is unavailable" as a variable.[110] The choice to define recidivism this way, however, presents a significant problem, considering contacts with the criminal justice system are not equally distributed, particularly around racial groups.[111] In essence,

---

106. *See, e.g.*, Mhany Mgmt., Inc. v. Cnty. of Nassau, 819 F.3d 581, 609 (2d Cir. 2016) (explaining that terms like "affordable housing" served as "[r]acially charged code words [which] may provide evidence of discriminatory intent") (quoting Smith v. Fairview Ridges Hosp., 625 F.3d 1076, 1085 (8th Cir. 2010)); Floyd v. City of New York, 959 F. Supp. 2d 540, 586 (S.D.N.Y. 2013) ("Crime suspect data may serve as a reliable proxy for the pool of *criminals* exhibiting suspicious behavior. But there is no reason to believe that crime suspect data provides a reliable proxy for the pool of *non-criminals* exhibiting suspicious behavior. Because the overwhelming majority of people stopped fell into the latter category, there is no support for the City's position that crime suspect data provides a reliable proxy for the pool of people exhibiting suspicious behavior.").

107. *See, e.g.*, J.E.B. v. Alabama ex rel. T.B., 511 U.S. 127, 143 (1994) (explaining that gender "may not serve as a proxy for bias" for removing jurors through peremptory strikes).

108. P'SHIP ON AI, REPORT ON ALGORITHMIC RISK ASSESSMENT TOOLS IN THE U.S. CRIMINAL JUSTICE SYSTEM 16 (2019), https://www.partnershiponai.org/wp-content/uploads/2019/04/Report-on-Algorithmic-Risk-Assessment-Tools.pdf [https://perma.cc/A3E2-CD3M].

109. *Id.* n.14.

110. *Id.*

111. *See, e.g.*, Ellen A. Donnelly & John M. MacDonald, *The Downstream Effects of Bail and Pretrial Detention on Racial Disparities in Incarceration*, 108 J. CRIM. L. & CRIMINOLOGY 775, 801 (2018); David S. Kirk, *The Neighborhood Context of Racial and Ethnic Disparities in Arrest*, 45 DEMOGRAPHY 55, 73–74 (2008); Rory Kramer & Brianna Remster, *Stop, Frisk, and Assault? Racial Disparities in Police Use of Force During Investigatory Stops*, 52 LAW & SOC'Y REV. 960, 986–88 (2018); Sandra G. Mayson & Megan T. Stevenson, *Misdemeanors by the Numbers*, 61 B.C. L. REV. 971, 1016–17 (2020); Carlos Berdejó, *Criminalizing Race: Racial Disparities in Plea-*

policy-salient variables like criminality and arrest become proxies for race.

Partnership for AI also determined that "in complex settings like criminal justice, virtually all statistical predictions will be biased even if the data was accurate, and even if variables such as race are excluded, unless specific steps are taken to measure and mitigate bias."[112] To do so, the data are trained by inputting variables that mimic omitted variables that are relevant causal factors. But these variables may be highly correlated with race or explicitly serve as proxies for race.[113] The ACLU has argued that data like a defendant's age, substance use, family relationships, and community ties can serve, alone and together, as proxies for race.[114] These variables are clearly legally permissible when employed for legitimate purposes, but in this context, they serve as stand-ins for race and ethnicity.

In machine learning risk assessments, race or its proxies are also used in a slightly different way—that is, in the training data through which the artificial intelligence learns about the world and how to make predictions about recidivism. A 2017 study published in *Science* found that "standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture."[115] Researchers found that historic biases and stereotyped attitudes involving race can permeate the training data used by algorithms, even if training data explicitly exclude race and ethnicity as variables.[116] While the algorithm may or may not itself expressly use race in the black box decision-making of its

---

Bargaining, 59 B.C. L. REV. 1187, 1194–1200 (2018).

112. P'SHIP ON AI, *supra* note 108, at 18.

113. *Id.*

114. ACLU OF KANSAS, CHALLENGING PRE-TRIAL RISK ASSESSMENT TOOLS (2019), https://www.kscourts.org/KSCourts/media/KsCourts/court%20administration/Pretri al_Justice_Task_Force/PJTFReporttoKansasSupremeCourt.pdf [https://perma.cc/ HA3T-J7YR].

115. Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-like Biases*, 356 SCIENCE 183, 183 (2017).

116. *Id.* at 185; *see also* David Arnold, Will Dobbie & Peter Hull, *Measuring Racial Discrimination in Algorithms* 2 (Becker Friedman Inst. Working Paper No. 2020-184, 2020) (finding that "a sophisticated machine learning algorithm discriminates against Black defendants, even though defendant race and ethnicity are not included in the training data. The algorithm recommends releasing [W]hite defendants before trial at an 8 percentage point (11 percent) higher rate than Black defendants with identical potential for pretrial misconduct, with this unwarranted disparity explaining 77 percent of the observed racial disparity in algorithmic recommendations. We find a similar level of algorithmic discrimination with regression-based recommendations, using a model inspired by a widely used pretrial risk assessment tool").

predictions, the algorithm's construction and training may be racialized because of the initial use of deeply racialized data. In other words, the criminal justice system is so deeply racist that by using criminal justice data to train algorithms, developers are creating naively racist artificial intelligence. Artificial intelligences are tasked with figuring out how to predict future rearrest before trial as their singular focus without legal restrictions on how to approach this goal.[117] These artificial intelligences are therefore dutifully examining patterns in the data and accurately detecting that in the recent past, at least, one of the best ways to predict who will be arrested in the future is to consider either the color of their skin or closely correlated proxies for race.

In addition, training an algorithm to make decisions may inadvertently create feedback loops that ultimately classify people based on their race and ethnicity. For example, the Netflix movie-streaming algorithm presents users with many options, and the user ultimately makes a choice that is then introduced as new knowledge that trains the algorithm to choose other movies.[118] The algorithm, however, does not consider that the user's choice was originally shown by the algorithm. As a result, a user receives recommendations similar to the choice the user initially made. Similarly, in the criminal justice context, poor minority groups are more likely to score higher in risk assessment predictions because the tools have large amounts of their data, which puts them at risk of more policing and indictments (which creates more data), ultimately reinforcing the systems' biases towards these groups.[119] In other words, the outcomes of predictions unjustly influence future predictions.

One objection that has been raised is that algorithmic risk assessments might not trigger strict scrutiny because they do not consistently and categorically disadvantage members of the suspect class. With machine-learning algorithms in particular, it has been argued that "consideration of class membership will not necessarily, or even often, give rise to categorically different treatment . . . [because] . . . most machine-learning applications will be used to forecast complex phenomena . . . that are not easily

---

117. *See generally* RUSSELL & NORVIG, *supra* note 34 (explaining various forms and methods of machine learning relevant to algorithms).

118. David Chong, *Deep Dive into Netflix's Recommender System*, TOWARDS DATA SCI. (Apr. 30, 2020), https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48 [https://perma.cc/RS4G-7HZ2].

119. *See* VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018).

predicted by standard, less powerful, statistical techniques."[120] However, even if that were the case more generally for algorithmic risk assessments, in the pretrial detention context, and particularly with respect to regression-based RATs, there is compelling evidence that algorithmic RATs treat racial groups differently, as will be further discussed below.[121]

In short, there is significant proof that algorithmic RATs classify individuals based on their race and ethnicity. These algorithms either explicitly use racial assumptions or impermissibly use variables as proxies for race. Algorithms can also engage in feedback loops, where racial biases are reinforced through the dynamism between inputs and outputs of data. Most importantly, how could a court know whether a privately owned algorithm actually uses suspect classifications as variables if they are not reviewable due to the proprietary nature of the tool? And similarly, how can courts examine whether proxy variables are legally permissible because they purportedly serve a legitimate purpose? The assertion of opacity of algorithms is not a valid argument of constitutional soundness.

B. Substantial Evidence of Discriminatory Intent and Disparate Impact

A court may find that state actors are not explicitly classifying individuals based on their race or ethnicity. However, a criminal defendant may still raise an Equal Protection claim by showing that algorithmic risk assessments result in racially disparate treatment of individuals, so long as it was motivated by racial animus. In *Washington v. Davis*, the Supreme Court held that a claimant may use racial impact as a relevant fact that bears on the question of racial intent—the key element.[122] The Court has also clarified that disparate treatment must be "'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group."[123] Therefore, an Equal Protection challenge of this nature must necessarily include proof of disparate treatment and discriminatory intent.

---

120. *See* Coglianese & Lehr, *supra* note 90, at 1196.

121. *See, e.g.*, *id.*, n.232 ("Regression analysis is more susceptible [than machine learning] to tacit bias because it is driven by theories about how individuals are likely to behave."); *see also* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043 (2019) (demonstrating that constitutional law is unsuited to correct racial discrimination resulting from using RATs in the criminal justice system).

122. *See* Washington v. Davis, 426 U.S. 229 (1976).

123. Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 279 (1979).

1.   Racial and Ethnic Disparate Impact

There is ample evidence that many algorithmic RATs used in criminal adjudications impact defendants differently based on their race or ethnicity. For instance, a ProPublica study analyzed the COMPAS risk score assessments for more than seven thousand people arrested in Broward County, Florida between 2013 and 2014.[124] They concluded that predictions were biased against Black defendants.[125] The analysis showed that while the overall accuracy of risk predictions for both Black and White defendants were very similar (61%), "[B]lack [individuals] are almost twice as likely as [W]hite [individuals] to be labeled a higher risk but not actually re-offend."[126] Conversely, White defendants received false negatives almost twice as often as their Black counterparts.[127] Similarly, University of Texas, Austin Law Professor Melissa Hamilton's study, which used the same dataset as ProPublica, found that COMPAS "is not well calibrated for Hispanics" in almost identical ways.[128] Put differently, COMPAS risk scores favor White defendants with both false positives and negatives.[129]

---

124. Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/QES8-KCKE].

125. *See id.*; *see also* Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 161 (2017) (using the ProPublica data, researchers "demonstrate[d]" how using a recidivism prediction instrument that has "different false postive [sic] and false negative rates between groups can lead to disparate impact when individuals assessed as high risk receive stricter penalties"); *see also* Melissa Hamilton, *Investigating Algorithmic Risk and Race*, 5 UCLA CRIM. JUST. L. REV. 53, 97–98 (2021) (finding that the PSA RAT has racially disparate impacts, demonstrates racialized group bias, and inconsistently classifies and predicts White and Black outcomes in ways that are not consistent with prevailing theories of algorithmic fairness). *But see* WILLIAM DIETERICH, CHRISTINA MENDOZA & TIM BRENNAN, NORTHPOINTE, INC., RSCH. DEP'T, COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY (2016) (disputing ProPublica's allegations); *but cf.* Jeff Larson & Julia Angwin, *Technical Response to Northpointe*, PROPUBLICA (July 29, 2016), https://www.propublica.org/article/technical-response-to-northpointe [https://perma.cc/N9N2-LJ9V].

126. *See* Angwin et al., *supra* note 124 (finding the "Labeled Higher Risk, But Didn't Re-Offend" rates were 44.9% for Black defendants and 23.5% for White defendants).

127. *Id.* (finding "Labeled Lower Risk, Yet Did Re-Offend" rates were 28.0% for Black defendants versus 47.7% for White defendants).

128. Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 AM. CRIM. L. REV. 1553, 1577 (2019).

129. *See generally* Goel et al., *supra* note 4, at 6 (providing an overview of recent research on this issue); *see also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016) (arguing that algorithms inherit racial biases in the data they rely on).

Another factor unique to RATs and disparate impact is that due to the scale and objective consistency of RATs, some classes of people could be likely to always get classified as high risk and in need of incarceration. Advocates of RATs tout the absence of individual bias or inconsistency in determinations of riskiness, as compared to fallible human judges spitballing riskiness using their subjective discretion.[130] However, the downside of that consistency is that any error the algorithm makes is repeated mercilessly every single time. Compared to humans, there is much less stochastic variation in the algorithms.[131] What this means is that if COMPAS or the PSA, for instance, determine that someone with a combination of some particular factors is at high risk of rearrest before trial, every member of that suspect classification will also be rated high risk. With judicial discretion, there is always room for the statistical error of mercy or of considering the particularities of a person's life that do not show up in models that by design simplify the messiness of the real world. The algorithms lack any such unexpected divergence from their predictions, since unconstrained algorithms are designed to objectively maximize predictive validity as best as possible, without subjective mercy or distraction.[132] When combined with the scale of their use, where every judge in a state might be relying on the exact same RAT, the potential for pretrial release recommendation becomes very difficult for someone who is a member of a group identified as high risk by the algorithm. For defendants with an unlucky combination of variables, it could be akin to not having any alternative to one particular judge's idiosyncratic biases. Any racial or ethnic biases within the algorithms are multiplied and compounded at scale, relentlessly.[133]

---

130. *See* Joseph J. Avery & Joel Cooper, *Racial Bias in Post-Arrest and Pretrial Decision Making: The Problem and a Solution*, 29 CORNELL J. L. & PUB. POL'Y 257, 270–71, 283–85 (2019); *see* Mayson, *supra* note 100, at 2278 ("Subjective prediction is vulnerable to irrational bias."); *see also* Cass R. Sunstein, *Algorithms, Correcting Biases*, 86 SOC. RSCH.: AN INT'L Q. 499, 502 (2019) (arguing pretrial judges "suffer from a cognitive bias that produces severe and systematic errors").

131. *See generally* RUSSELL & NORVIG, *supra* note 34 (explaining how machine learning causes algorithms to repeat information with near-perfect consistency).

132. *See* Sam Corbett-Davies, Sharad Goel, Emma Pierson, Aziz Z. Huq & Avi Feller, *Algorithmic Decision Making and the Cost of Fairness* 1–10 (Stan. Univ., Working Paper, Feb. 17, 2017). *See generally* O'NEIL, *supra* note 12 (arguing for a similar problem of scale in another context).

133. *See generally* O'NEIL, *supra* note 12, at 124 (explaining the process by which biases are replicated by algorithms).

2. Discriminatory Intent

We argue below that discriminatory intent may be inferred from: (1) deliberate indifference to racial targeting; (2) discriminatory animus from the algorithm's designer; and (3) discriminatory intent from the machine.[134]

First, in *Floyd v. City of New York*, the district court determined that "the use of a facially neutral policy applied in a discriminatory manner, or through express racial profiling, targeting [minority populations] violates bedrock principles of equality."[135] At issue in this case was whether the New York Police Department's stop-and-frisk policy violated Fourteenth Amendment protections of Black and Latino individuals. The court reasoned that plaintiffs there showed a state "*policy* of indirect racial profiling" where the state acted "*deliberately indifferent* to the intentionally discriminatory application" of that policy.[136] According to the court, a state policy includes "the decisions of a government's lawmakers, the acts of its policymaking officials, and practices so persistent and widespread as to practically have the force of law."[137] In *Floyd*, a policy directing officers to target young Black and Latino men "based on local crime suspect data" and racial animosity by the police commissioner were sufficient to prove intent.[138]

Regarding algorithmic RATs, states have clearly ignored these tools' discriminatory impact on Black and Latino defendants. In fact, states in the last decade have aggressively enacted legislation and executive policies mandating the use of these tools, despite criticism from communities, experts, and advocacy organizations.[139]

---

134. *See* Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 911 (2018); *see also* Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 826 (2020) (explaining how instructions to computers can inject race into the algorithm); *see also* Coglianese & Lehr, *supra* 90, at 1198 (acknowledging that some opponents of algorithms argue the inclusion of a race variable itself shows discriminatory intent); *see also* Floyd v. City of New York, 959 F. Supp. 2d 540, 664 (S.D.N.Y. 2013) (finding a facially neutral police policy failed strict scrutiny where it resulted in higher levels of stops among non-White drivers).

135. *Floyd*, 959 F. Supp. 2d at 664. Companies like Equivant, which owns COMPAS, claim they do not engage in express racial profiling, but, as argued above, that is either false or they use proxies impermissibly to racially profile.

136. *Id.* at 660.

137. *Id.* at 558, 564 (quoting Connick v. Thompson, 563 U.S. 51, 61 (2011)).

138. *Id.* at 660.

139. *See, e.g.*, Tom Simonite, *Algorithms Should've Made Courts More Fair. What Went Wrong?*, WIRED (Sept. 5, 2019), https://www.wired.com/story/algorithms-shouldve-made-courts-more-fair-what-went-wrong/ [https://perma.cc/B9K4-7RAS] (explaining that a 2011 Kentucky law requires judges consider an algorithmic risk

Moreover, judges have used algorithmic risk assessments in ways that disadvantage Black defendants, and which a reasonable person would expect them to be aware disadvantage Black defendants. A 2019 study by the Harvard John M. Olin Center for Law, Economics, and Business found that "judges were more likely to override the [bail] recommended default for moderate risk [B]lack defendants than similar moderate risk [W]hite defendants," likely "suggest[ing] that interaction with the same predictive score may lead to different predictions by race."[140] The study further argued that such results may be caused by judges being unresponsive to policy changes or acting with racial animosity.[141]

As such, a *Floyd* intent framework could be applied to algorithmic RATs because state actors have both deliberately ignored the adverse effects on Black and Latino defendants, as well as mandated their use without consideration of scientific studies warning against their use.[142]

Second, human bias from data scientists creating and training the algorithms may encroach into the data.[143] A data scientist makes a series of choices when designing the formulas to be used by the algorithmic tool. As University of Chicago Law School Professor Aziz Z. Huq explains: "an algorithm's designer might be motivated by either an animosity toward a racial group, or else a prior belief that race correlates with criminality, and then deliberately design the algorithm on that basis."[144] Such design-making "might occur through either a choice to use polluted

---

assessment when posting bail); *see also* Elizabeth Hardison, *After Nearly a Decade, Pa. Sentencing Commission Adopts Risk Assessment Tool Over Objections of Critics*, PA. CAPITAL-STAR (Sept. 5, 2019), https://www.penncapital-star.com/criminal-justice/after-nearly-a-decade-pa-sentencing-commission-adopts-risk-assessment-tool-over-objections-of-critics/ [https://perma.cc/RPC7-9VK3] (illustrating a 2019 Pennsylvania law adopting the use of algorithmic risk assessment for sentencing determinations).

140. Alex Albright, *If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions* 1 (Harv. John M. Olin Ctr. for L., Econ., & Bus. Fellows' Discussion Paper Series, Discussion Paper No. 85, 2019), http://www.law.harvard.edu/programs/olin_center/Prizes/2019-1.pdf [https://perma.cc/VA78-EZHC].

141. *Id.* at 25.

142. *See* Angwin et al., *supra* note 124 (finding that from a sample of seven thousand criminal defendants in Broward County, Florida, Black defendants were "77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind" than their White counterparts, controlling for race, gender, age, criminal history, and recidivism).

143. *See* P'SHIP ON AI, *supra* note 108, at 15–22; *see also* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018) (explaining how human bias encroaches into computer programs run by algorithms).

144. Huq, *supra* note 121, at 1089.

training data or the deliberate selection of some features but not others on racial grounds."[145] Some state courts have singled out algorithms' developers as legally responsible for the algorithms in some respects.[146] However, the opaque or proprietary nature of the algorithmic tools may prohibit defendants from determining how the data scientist designed the algorithm. If the algorithm is unreviewable, then it is challenging to directly detect the designer's motivation.

Indirect evidence of intentionality can be deduced from an important mathematical proof by the statisticians Jon Kleinberg and colleagues, which has since been replicated.[147] Analyzing the ProPublica COMPAS data, they found that there are three main ways to operationalize racial equality: racial equality of false negatives, racial equality of false positives, and racial parity of outcomes.[148] They proved that in a context of unequal initial conditions (i.e., racial disparity in recidivism rates), it is mathematically impossible for the three types of equality to be achieved simultaneously, so there is a necessary trade-off between the three forms of equality.[149] This trade-off implies that creators of the assessments are making choices about trade-offs, intentionally or unintentionally. Minimizing racial inequality in risk assessments became such a priority among legal and policy decision-makers that most current assessments include attempts to minimize racial disparities.[150] Kleinberg and colleagues' proof then implies that any intentional act of reducing inequality in one

---

145. *Id.*

146. *See, e.g.*, People v. Wakefield*, 175 A.D.3d 158, 169–70 (N.Y. App. Div. 2019) (holding that the "true accuser" within a Confrontation Clause challenge was the writer of the source code for an algorithm used in software that calculates the probability of a defendant's presence at the scene of the crime, considering that said algorithmic source code writer was "the declarant in the epistemological, existential and legal sense rather than the sophisticated and highly automated tool powered by electronics and source code that he created"), *lv denied*, 34 N.Y.3d 1083 (N.Y. App. Div. 2019), *lv granted*, 35 N.Y.3d 1097 (N.Y. App. Div. 2020).

147. JON KLEINBERG, SENDHIL MULLAINATHAN & MANISH RAGHAVAN, INHERENT TRADE-OFFS IN THE FAIR DETERMINATION OF RISK SCORES (2017), https://arxiv.org/pdf/1609.05807.pdf [https://perma.cc/9BGF-WZU5].

148. *See id.* at 4. There are other models of equality, but similar arguments hold for those models. *See* Huq, *supra* note 121, at 1053 (2019) (arguing that the law "provides no creditable guidance" about which model of fairness or equality to apply to risk assessments); Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kerns & Aaron Roth, *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOCIO. METHODS & RSCH., 3, 34–35 (2021).

149. KLEINBERG ET AL., *supra* note 147, at 17.

150. Joseph J. Avery & Joel Cooper, *Racial Bias in Post-Arrest and Pretrial Decision Making: The Problem and a Solution*, 29 CORNELL J.L. & PUB. POL'Y 257, 289 (2019).

dimension necessarily involves intentionally increasing inequality in one of the other dimensions. Assessment developers cannot argue that remaining inequalities were unintentional or incidental since the trade-offs force a developer to make a choice between inequalities.

Third, for machine learning algorithmic RATs, can intent also be inferred from decisions made by machines based on their "deep learning" and autonomous decision-making?[151] The Supreme Court has not yet ruled on whether machines that replace human decision-making should be treated like natural persons for Equal Protection intent purposes. Still, there is great interest in the question of legal personhood for artificial entities and autonomous devices.[152] For instance, judges are barred from considering race and ethnicity when making bail or sentencing determinations.[153] However, judges rely on an algorithmic assessment that, as mentioned above, directly or indirectly uses prohibited classifications. Furthermore, the machine is able to learn and apply racial biases and stereotypes (racial animosity), as in the case of the Netflix algorithm. The algorithm selects a defendant's features to make a choice of who the defendant is, without ever needing to use race or ethnicity as a factor. The machine then would be liable for discriminatory intent just like a court officer who created a bail determination report or PSI.[154] In other words, if one treats a

---

151. These arguments about AI's intermediate level of legal intentionality do not as clearly apply to regression-based RATs because their algorithms are not inherently opaque and independent like the machine learning RATs. However, that makes the regression-based RATs more likely to be found to use race facially (or some other elements discussed above), and machine learning RATs more likely to pass a facial discrimination Equal Protection Clause review, but fail a disparate impact plus intent review.

152. *See* Gerhard Wagner, *Robot, Inc.: Personhood for Autonomous Systems?*, 88 FORDHAM L. REV. 591, 593 (2019) (prepared for the symposium *Rise of the Machines: Artificial Intelligence, Robotics, and the Reprogramming of Law*); *see also* Citizens United v. Fed. Election Comm'n, 558 U.S. 310 (2010) (expanding personhood rights for artificial corporate entities based on an Equal Protection claim); Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043 (2020) (arguing for the application of strict liability to humans responsible for damages caused by AI entities acting as the human's agent).

153. *See* 28 U.S.C. § 994(d) ("The Commission shall assure that the [sentencing] guidelines and policy statements are entirely neutral as to the race, sex, national origin, creed, and socioeconomic status of offenders."); U.S. PROB. OFF. FOR THE W. DIST. OF N.C., THE PRESENTENCE INVESTIGATION REPORT: A GUIDE TO THE PRESENTENCE PROCESS 6 (2009) ("[C]ertain demographic data such as age, race and sex are precluded from consideration in the sentencing process both by statute and by the guidelines . . . .").

154. *See* Kimberly Mok, *Mathwashing: How Algorithms Can Hide Gender and Racial Biases*, NEW STACK (Dec. 8, 2017), https://thenewstack.io/hidden-gender-

machine as a human, would we permit the human to do this? Here, absent judicial precedent, the answer is *likely no*. Thus, while the *Loomis* and *Malenchik* courts held that algorithmic assessments are one of many factors considered by a judge in making sentencing determinations, treating these tools as human-like systems may alter judicial review of the intent issue.

A useful legal model has been developed in Germany[155] that could be applied to the most advanced RATs.[156] In the German model of *Teilrechtsfähigkeit*, or partial legal capacity, advanced machine-learning algorithms such as unsupervised machine learning RATs would be treated as legal subjects in some limited ways that entail some independent legal capacity under the indirect supervision of humans. In this partial legal capacity model, algorithms "are not legal persons with full legal capacity, they are still legal subjects, yet the range of their subjectivity is limited by their specific functions."[157] Some U.S. courts have already suggested that more independent AI systems could have something like *Teilrechtsfähigkeit* in, for example, the context of Sixth Amendment Confrontation Clause challenges.[158] In the pretrial risk

---

racial-biases-algorithms-can-big-deal/ [https://perma.cc/7343-PLYN] ("In one recent study which trained an off-the-shelf machine learning AI system on 2.2 million words, Princeton University researchers used a word-association technique to map out what kind of links the system would [make] between words and concepts. It found that the system would associate words such as 'flower' and 'music' as being more pleasant concepts than words like 'insects' and 'weapons.'").

155. Jan-Erik Schirmer, *Artificial Intelligence and Legal Personality: Introducing "Teilrechtsfähigkeit": A Partial Legal Status Made in Germany*, *in* REGULATING ARTIFICIAL INTELLIGENCE 123 (Thomas Wischmeyer & Timo Rademacher eds., 2020) ("[I]ntelligent agents would be treated as legal subjects as far as this status followed their function as sophisticated servants. This would both deflect the 'autonomy risk' and fill most of the 'responsibility gaps' without the negative side effects of full legal personhood."); *see also* Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 549 (proposing "a new category of a legal subject, halfway between person and object"); Wagner, *supra* note 152, at 608 (2019) (developing a comparable intermediary tort liability status for AI systems using "a functional explanation that is in tune with the general principles and goals of tort law, namely compensation and deterrence," which is particularly needed when "people injured by a robot may face serious difficulties in identifying the party who is responsible for the misbehavior of the device").

156. This analysis applies to machine learning RATs since they share important characteristics with human decision-making, such as processing information independently without direct human supervision, unlike regression-based RATs that require direct supervision and would be more comparable to very sophisticated tools.

157. Schirmer, *supra* note 155, at 135. This model has previously been applied in Germany to preliminary companies, homeowners' associations, and fetuses.

158. *See* People v. Wakefield, 175 A.D.3d 158, 169–70 (N.Y. App. Div. 2019) (finding that "an artificial intelligence-type system" involving "distributed cognition between technology and humans" could itself be a declarant in a Sixth Amendment challenge, depending on the level of human supervision and the totality of the

assessment context, machine learning RATs would be legal subjects only in the sense that they are responsible for independently performing functions for human subjects,[159] such as using race to accurately predict risk of recidivism. Legal questions about artificial intelligence will only become more common in the near future, and the model of partial legal capacity could resolve many pressing legal dilemmas, such as conflicts involving driverless cars.[160]

C. Algorithmic Risk Assessments Do Not Pass Judicial Strict Scrutiny

The last step of a suspect class Equal Protection analysis requires a showing that the means chosen to achieve a compelling government interest be narrowly tailored. Many states purportedly employ algorithmic RATs to eliminate or reduce racial disparities in the criminal justice system.[161] Proponents also advocate for their "potential to streamline inefficiencies, reduce costs, and provide rigor and reproducibility for life-critical decisions."[162] However, the use of algorithmic assessment tools is not narrowly tailored to meet those objectives because they are not the least restrictive means necessary to achieve those government interests—they do not produce considerably better assessments, and they negatively influence judges. Further, the opacity of many RATs makes it impossible for the government to meet its burden of proof that they are narrowly tailored.

Studies have found that algorithmic risk calculations for recidivism are no more accurate or less racially biased than human predictions. For example, a high-profile 2018 Dartmouth University study found that COMPAS risk calculations were "nearly identical"

---

circumstances); *see also* Itiel E. Dror & Jennifer L. Mnookin, *The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science*, 9 L., PROBABILITY & RISK 47, 48–49 (2010).

159. *See* Schirmer, *supra* note 155, at 136 (emphasizing that algorithmic partial legal capacity does not require complete intentional autonomy, since a "trading algorithm does not trade on its own account, but on the account of the person who deploys it. In other words, we are looking at the typical 'master-servant situation', in which the servant acts autonomously, but at the same time only on the master's behalf").

160. *See generally* Neal Katyal, *Disruptive Technologies and the Law*, 102 GEO. L.J. 1685 (2014) (discussing the potential problems arising from the development of mass surveillance, 3D printing, and driverless cars).

161. *Id.*; *see* Adam Neufeld, *Commentary: In Defense of Risk-Assessment Tools*, MARSHALL PROJECT (Oct. 22, 2017), https://www.themarshallproject.org/2017/10/22/in-defense-of-risk-assessment-tools [https://perma.cc/FU5T-YJH7].

162. P'SHIP ON AI, *supra* note 108, at 7.

to untrained humans at predicting recidivism.[163] The study also confirmed ProPublica's finding that COMPAS racially disproportionately assigns false positives and negatives to criminal defendants, showing that these tools are no better than judges overall.[164] Further, they found that, although COMPAS uses 137 variables in an opaque algorithm, the same accuracy could be achieved with a simple linear regression with only two variables: age and total number of previous convictions.[165] These two equivalent and more narrowly tailored alternatives suggest that risk assessments like COMPAS are not the least restrictive means necessary to achieve the state's objectives.

Judges are supposed to consider, but not rely on, algorithmic assessments in pretrial adjudications.[166] However, studies find that judges sometimes completely rely or are heavily influenced by these assessments. Also known as *automation bias*, cognitive biases may cause judges to over-rely on algorithmic assessments because of "the brain's natural tendency to rely on heuristics, or simple rules of thumb, when dealing with complicated mental tasks."[167] The empirical research on how judges use RATs is limited, but a 2019 Harvard University study simulated pretrial judicial discretion with respect to automated risk assessments using an online survey experiment to assess how people make predictions about pretrial risk, both with and without RATs.[168] The results were consistent with automation bias, with researchers finding that participants' behavior heavily mimicked that of the algorithms, "which can be racially biased even when race is not included as an explicit

---

163. Dressel & Farid, *supra* note 99, at 3; *see* Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel & Daniel G. Goldstein, *Simple Rules for Complex Decisions* 9 (Stan. Univ., Working Paper, Apr. 4, 2017) (demonstrating that humans using simple weighted checklists comparable to the early Vera Scale are as accurate as complex algorithmic risk assessments at predicting rearrest before trial).

164. Dressel & Farid, *supra* note 99, at 3.

165. *See id.*

166. *See* Brief for the U.S. as Amicus Curiae at ¶ 98, Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (No. 16-6387); Andrea Nishi, *Privatizing Sentencing: A Delegation Framework for Recidivism Risk Assessment*, 119 COLUM. L. REV. 1671, 1672 (2019).

167. Andrew Lee Park, *Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing*, UCLA L. REV.: L. MEETS WORLD (Feb. 19, 2019), https://www.uclalaw review.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/ [https://perma.cc/5Y4P-43BE] (citing Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124, 1125 (1974)).

168. *See* Ben Green & Yiling Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments* (Conf. on Fairness, Accountability, and Transparency, 2019), https://scholar.harvard.edu/files/19-fat.pdf [https://perma .cc/XBT4-VXRB].

factor."[169] In a related court context, an ongoing study of prosecutorial discretion suggests that prosecutors were strongly influenced by RATs, even though they were unaware of what elements went into the scores.[170] Prosecutors who had been prepared to offer defendants diversion programs were swayed to not do so "because the risk assessment showed too high of a risk," even though after being pressed the prosecutors could not explain the elements of the score or what determined the risk levels.[171]

Another weakness in the narrowly tailored step of the Equal Protection argument could be that some RATs are too opaque to prove that they are narrowly tailored. As discussed above, machine-learning RATs evolve specific processes on their own in response to real-world data, so their precise algorithms are not programmed or known by any human.[172] Although regression-based RATs are not inherently opaque in the same way, RATs like COMPAS are de facto opaque because their algorithms are protected as trade secrets. Yet according to the Supreme Court, "[u]nder strict scrutiny, the government has the burden of proving that racial classifications are narrowly tailored measures that further compelling government interests."[173] It is the burden of the state to prove that no other alternative that is less intrusive of the right could work to achieve those interests. If the black-box algorithms driving machine learning RATs are by nature too unidentifiable to prove that they are or are not narrowly tailored (or if corporations like the designers of COMPAS refuse to open the black box of the algorithm to prove it), then the government using these risk assessments would necessarily fail to meet their burden of proof.[174]

In sum, the government cannot meet its burden of proof that algorithmic assessment tools are narrowly tailored to meet the

---

169. *Id.* at 8.

170. Chiara C. Packard, *"The Question Is, Should You Charge?": A Multi-Site Case Study Exploring Prosecutor's Use of Discretion in Wisconsin* (Soc'y for the Study of Soc. Probs. Ann. Conf., 2021).

171. *Id.*

172. *See, e.g.*, Coglianese & Lehr, *supra* note 90, at 1199 ("Given how machine-learning analysis works on a black-box basis, it is virtually impossible for anyone to know a priori what a given variable's likely importance in the algorithm will be or what its ultimate effects will be on any disparities of predictions.").

173. Johnson v. California, 543 U.S. 499, 505 (2005) (internal quotation marks omitted) (quoting Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 227 (1995)).

174. This is less applicable to regression-based algorithms that are more transparent in their processes, but that transparency in turn makes those RATs more vulnerable to discriminatory intent claims. For instance, in a defense of machine learning RATs, Coglianese and Lehr admit that algorithmic "[r]egression analysis is more susceptible to tacit bias because it is driven by theories about how individuals are likely to behave." Coglianese & Lehr, *supra* note 90, at 1205 n.232.

government's purported goal of reducing bias in the criminal justice system. These tools do not perform better than untrained humans or judges, nor do they perform better than simple and utterly transparent regressions of two variables. In addition, they impact judicial discretion by pointing judges to ultimately make racially biased determinations. Furthermore, these algorithmic tools carry significant weight, if not complete weight, in a judge's determination of pretrial adjudications. Therefore, algorithmic assessments are not the least restrictive means necessary to achieve the state's purported compelling purpose of, among other things, reducing biases in judges and releasing more defendants pretrial.

IV. Limitations and Other Considerations

One major limitation of an Equal Protection challenge against privately owned RATs is that their algorithms are considered trade secrets, and therefore it would be hard for courts to evaluate the legally relevant processes. As a result of their trade secret status, the algorithms may not be evaluated by the general public or criminal defendants without consent of the company. Companies often do not grant consent because it may result in criticism and revelation of secret information, both of which could cut into corporate profit.[175] Courts have sided with companies on this issue. For example, in 2014, the Urban Justice Center filed a Freedom of Information Law (FOIL) request in New York State for COMPAS's "instruction manuals, training guides and information regarding scoring for the COMPAS Reentry Assessment tool," as part of an administrative challenge.[176] The request was denied because of the "trade secrets" exemption under FOIL, as "these materials are the sole property of Northpointe."[177] However, the trade secrecy argument may empower courts to ban privately-owned algorithms altogether, since they lack government and public review, as was mandated by the court in *Loomis*. In fact, courts could start reviewing these tools in camera or through protective orders.[178]

---

175. *See* Andrew A. Schwartz, *The Corporate Preference for Trade Secret*, 74 OHIO ST. L.J. 623 (2013) (discussing the benefits of trade secrets for corporations).

176. Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System* 14 n.51 (Apr. 14, 2017) (unpublished manuscript) (on file with author) ("This FOIL request was submitted in connection with an Article 78 ruling finding that the COMPAS tool was not adequately tailored for use on individuals with mental illness.").

177. *Id.*

178. *But see* Nishi, *supra* note 166, at 1682–83 n.70 ("Although in civil cases these protections can be overcome through protective orders or in camera review, the use

Another limitation is that the Supreme Court requires an individualized inquiry for Equal Protection challenges. In *McClesky v. Kemp*, the Court held that "[s]tatistics at most may show only a likelihood that a particular factor entered into some decisions," and are usually insufficient to show a particularized injury.[179] The Court reasoned that "the application of an inference drawn from the general statistics to a specific decision in a trial and sentencing" were permissible in jury selection claims, but not in reviewing judicial discretion in capital sentencing.[180] Pretrial determinations are closer in the procedural stage to jury selection, but are made by judges like in capital sentencing determinations. Therefore, in the event a defendant is unable to review their individualized assessment due to the algorithm's proprietary nature and corporate trade secrecy, it is unclear whether a court would accept statistical generalizations to find particularized harm of an individual defendant, particularly in the face of companies who refuse to reveal their algorithms.

Conclusion

Algorithmic RATs in pretrial adjudication are not constitutionally sound. Their opacity, biases, judicial influence, and racially disparate treatment of Black and Latino defendants, all of whom are legally innocent, likely do not pass muster under the Equal Protection framework. Nonetheless, many states continue to advocate for their implementation in the criminal justice system, especially with bail reform gaining traction in jurisdictions across the United States.

We reject the idea of modifying or improving these algorithms to make them marginally less discriminatory, since the constitutional problems with risk assessments are fundamental, not fixable at the margins. For example, there is simply no way to use arrest data algorithmically that is not discriminatory, since racial discrimination is always already baked into prior arrest data. Instead, many less racially discriminatory alternatives to pretrial risk assessments have been proposed, such as public health approaches to identifying pretrial needs of people charged with crimes.[181] Indeed, major organizations like the Pretrial Justice

---

of these techniques in the criminal context may conflict with a defendant's Sixth Amendment right to a public trial.").

179. McCleskey v. Kemp, 481 U.S. 279, 308 (1987).

180. *Id.* at 294.

181. *See, e.g.*, ALICIA VERANI, RODRIGO PADILLA-HERNANDEZ, TALI GIRES, KAITLYN FRYZEK, RACHEL PENDLETON, ETHAN VAN BUREN & MÁXIMO LANGER,

Institute have called for abolition of both bail and algorithmic risk assessments.[182] Similarly, as part of its recent pretrial reforms, New York City experimented successfully with a system of behavioral nudges in the form of phone calls and texts reminding people of court dates, which significantly reduced rates of failing to appear in court.[183]

Notwithstanding these promising alternatives, the focus of this Article is not to comprehensively assess alternatives to risk assessments, but rather to identify their unconstitutionality. As argued above, the use of risk assessments is legally impermissible because it violates the Equal Protection rights of people of color, who are too often doomed by these algorithms to be swept up into the system. Machine learning risk assessments in particular are not narrowly tailored to minimize discrimination; they are naively racist systems that are inscrutably tailored to maximize predictive accuracy by any means necessary. Yet no one should be subjected to the pains of pretrial incarceration because they are a member of a particular racial or ethnic class. It is time to think beyond algorithmic risk assessments and reimagine equitable alternatives to pretrial justice.

CREATING A NEEDS-BASED PRE-TRIAL RELEASE SYSTEM: THE FALSE DICHOTOMY OF MONEY BAIL VERSUS RISK ASSESSMENT TOOLS 1 (2019) https://law.ucla.edu/ sites/default/files/PDFs/Academics/CJP   Pretrial   Proposal   -   2020.pdf [https://perma.cc/WTG2-49C5] (proposing a public health approach to pretrial justice that eschews RATs and prioritizes support services and "a presumption of release").

182. *Updated Position on Pretrial Risk Assessment Tools*, PRETRIAL JUST. INST. (Feb. 7, 2020), https://static1.squarespace.com/static/61d1eb9e51ae915258ce573f/t/ 61df34bb945c52230a215be9/1642018002889/PJI+Statement+Against+Risk+Assess ments [https://perma.cc/538L-3HHM] (arguing that "[r]egardless of their science, brand, or age, these tools are derived from data reflecting structural racism and institutional inequity that impact our court and law enforcement policies and practices. Use of that data then deepens the inequity"); *see also The Case Against Pretrial Risk Assessment Instruments*, PRETRIAL JUST. INST. (Nov. 2020), https://static1.squarespace.com/static/61d1eb9e51ae915258ce573f/t/61df300e02183 57bb223d689/1642017935113/The+Case+Against+Pretrial+Risk+Assessment+Instr uments--PJI+2020.pdf [https://perma.cc/5SDC-5NDZ] (arguing that "[p]retrial risk assessment instruments (RAIs) are constructed from biased data, so the RAIs perpetuate racism[;] RAIs are not able to accurately predict whether someone will flee prosecution or commit a violent crime[;] RAIs label people as 'risky' even when their odds of success are high[;] [and] RAI scores inform conditions of release, but there is no proven connection between RAI scores, specific conditions, and pretrial success").

183. *See* Russell Ferri, *The Benefits of Live Court Date Reminder Phone Calls During Pretrial Case Processing*, 18 J. EXPERIMENTAL CRIMINOLOGY 149, 160 (2022) (finding that phone call reminders reduced failures to appear by thirty-seven percent); Alissa Fishbane, Aurelie Ouss & Anuj K. Shah, *Behavioral Nudges Reduce Failure to Appear For Court*, 370 SCIENCE 1 (2020) [https://perma.cc/8BNS-YWJ9] (finding that text message reminders reduced failures to appear by twenty-one percent, and redesigned forms reduced them by thirteen percent).