ALGORITHMIC JUDICIAL ETHICS

KEITH SWISHER*

Judges have a brand-new bag—an algorithmic accessory in criminal adjudication. It scores criminal defendants, aiming to inform judges which defendants are likely reoffenders or flight risks and which ones are not. The downsides, however, include that the algorithms score defendants primarily on the basis of other defendants' (mis)conduct and that certain races effectively score lower than other races. This article explores these algorithmic developments in criminal courts across the country and makes four contributions: (1) a survey and preliminary application of judicial ethics to this development; (2) a preliminary moral argument, informed by related judicial ethics and legal standards, suggesting that judges should use these algorithmic tools only to help, not hurt, individual defendants; (3) an approach to judicial decision-making in the shadow of structural injustice that promises to deal less algorithmic damage to defendants and their family members; and (4) a technical constraint on algorithmic design that ensures equal (indeed, better than equal) protection on the basis of race.

* Professor, University of Arizona James E. Rogers College of Law. I owe many thanks to Marcello Di Bello, Charles Gardner Geyh, Chris Griffin, Cas Laskowski, Barbara Atwood, Shalev Roisman, and Shefali Milczarek-Desai for insightful comments on drafts or presentations of this work. The editors of the *Wisconsin Law Review* also significantly improved this work. Any errors are mine.

https://doi.org/10.59015/wlr.VLLW7889

INTRODUCTION

> *[T]here is an eerie similarity between the storyline in Minority Report and judicial reliance on AI at sentencing.*
>
> —Judge Herbert B. Dixon, Jr.[1]

Judging is less of a human endeavor than it once was. In criminal courts in the United States, algorithms now participate by rating the defendant as more or less likely to skip trial or to commit a future crime.[2] For example, when defendant Darrell Brooks, while out on bail, killed six people and injured dozens in a Christmas parade, the tragedy drew attention to both the human and algorithmic components in criminal justice.[3] The algorithm apparently got it right, having predicted that Brooks was highly likely to commit another offense.[4] At the time of Brooks's bail hearing, however, the human prosecutor did not have access to the algorithm's prediction and recommended an "'inappropriately' low" cash bail amount.[5] The judge agreed with the low amount, Brooks's mother posted it, and the parade killings sadly happened just days later.[6]

---

1. Herbert B. Dixon Jr., Foreword, Noel L. Hillman, *The Use of Artificial Intelligence in Gauging the Risk of Recidivism*, JUDGES' J., Winter 2019, at 36.

2. *See* Clare Amari, *Wisconsin Debates Cash Bail Changes in the Wake of Waukesha Parade Tragedy — As Some States Ditch System Entirely*, WISC. PUB. RADIO (Jan. 23, 2022), https://www.wpr.org/justice/wisconsin-debates-cash-bail-changes-wake-waukesha-parade-tragedy-some-states-ditch-system-entirely [https://perma.cc/E6YM-Z2DR].

3. *Id.*

4. *Id.*

5. Minyvonne Burke, *Suspect in Waukesha Holiday Parade Crash Was Released on Low Bail Because of 'Human Error,' DA Says*, NBC NEWS (Dec. 3, 2021, 10:42 AM), https://www.nbcnews.com/news/us-news/suspect-waukesha-holiday-parade-crash-was-released-low-bail-human-erro-rcna7515# (quoting John Chisholm, Milwaukee County District Attorney).

6. *See id.*; Amari, *supra* note 2; Shaun Gallagher, *Court Commissioner Who Set Low Bail for Waukesha Parade Suspect Reassigned Indefinitely*, NBC TMJ4 (Dec. 3,

Brooks's case is anomalous, in no small part because the algorithm was missed. The future is instead marked with the use, not the omission, of algorithms and artificial intelligence in criminal justice. Eric Loomis, as just one example, received six years in prison in part because a sophisticated algorithm classified him as "a high risk to the community."[7] Algorithms have thus started to influence judicial decision-making across the country. These predictive or actuarial models influence judicial rulings on, for example, pretrial release and sentencing of criminal defendants.[8] Indeed, many jurisdictions now require the use of these new models in criminal courts.[9] Yet the literature to date has all but ignored the significant implication for judges and judicial ethics. Judges looking for guidance on how to ethically interact with these algorithmic models are left in the lurch, potentially violating the ethical rules and harming defendants. This Article provides this missing guidance and hopefully

---

2021, 6:01 PM), https://www.tmj4.com/news/waukesha-christmas-parade/court-commissioner-who-set-low-bail-for-waukesha-parade-suspect-reassigned-indefinitely [https://perma.cc/C2RG-QEFZ] (noting that the judicial commissioner, who had set Brooks's cash bail at $1,000 when Brooks was accused of having run over the mother of his child, was thereafter "reassigned to non-criminal matters indefinitely"). To view Brooks's algorithmic prediction, see MILWAUKEE CNTY., PRETRIAL RISK ASSESSMENT REPORT: DARRELL BROOKS (2021), https://wisconsinwatch.org/wp-content/uploads/2021/06/Brooks_pretrial_assessment.pdf [https://perma.cc/GP9P-SUFL].

7. *See* Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html (quoting sentencing judge). This is not the only case in which an algorithm or AI has led to increased incarceration or wrongful arrest. *See, e.g.*, Kashmir Hill, *Wrongly Accused by an Algorithm*, N.Y. TIMES, https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html (Aug. 3, 2020) (discussing facial recognition technology leading to a wrongful arrest); Garance Burke, Martha Mendoza, Juliet Linderman & Michael Tarm, *How AI-Powered Tech Landed Man in Jail with Scant Evidence*, ASSOCIATED PRESS (Mar. 5, 2022, 12:23 PM), https://apnews.com/article/artificial-intelligence-algorithm-technology-police-crime-7e3345485aa668c97606d4b54f9b6220 (discussing how gunshot detection technology installed in neighborhoods can lead to false positives and wrongful arrests).

8. *See, e.g.*, Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1068–76 (2019) (noting that "[a]lgorithmic tools are used now in three main criminal justice contexts: policing, bail decisions, and post-conviction matters" and discussing each context); Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439, 450 (2020) ("Risk assessments are now commonplace at each stage of the criminal process, from police investigations, pretrial settings, sentencing, corrections, during parole and community supervision . . . ."); *see also id.* at 452–53 ("There are many important legal and policy differences between the pretrial and sentencing contexts. In the pretrial context, the question is whether a person will appear in court and whether they might pose a danger of recidivism pretrial.").

9. *See, e.g.*, Huq, *supra* note 8, at 1075 ("In some jurisdictions, such as Pennsylvania, New Hampshire, Arkansas, and Vermont, state law even affirmatively mandates the use of predictive instruments in the sentencing phase.").

will spur judicial ethics authorities to provide additional guidance and oversight.

Part I below briefly highlights some of the key benefits and risks of algorithmic criminal justice, and Part II discusses the current judicial uses of algorithms in criminal justice, focusing on algorithms deployed in pretrial release, sentencing, and postconviction release determinations. Beginning the Article's unique contributions, Part III reviews the judicial ethics system to reveal what it forbids, requires, suggests, and completely misses in this new technological frontier. Finally, Part IV offers three distinct approaches to make judges and the judiciary more responsible in their current and future algorithmic usage. In light of the current limitations and risks of algorithmic-based decisions in criminal cases (*e.g.*, the risk of perpetuating or even amplifying racial prejudice), judicial ethics rules, due process, and equal protection suggest that judges generally (1) should not punish (but may help) a defendant based on the algorithm's output, (2) must exercise increased caution and competence about the algorithm's inputs and outputs, and (3) should consider demanding racial statistical parity or at least four-fifths of it from the criminal justice algorithms discussed below.

I. THE PROMISE AND PRECARIOUSNESS OF ALGORITHMIC JUSTICE

To set the stakes, we should briefly highlight the material benefits and detriments of algorithmic usage in criminal justice.

Pieces of the algorithmic turn appear laudable. Copious studies show human decision-makers' susceptibility to implicit and cognitive biases, and reliance on seemingly objective algorithms holds the promise to lessen or eliminate these biases and errors. For example, owing at least subconsciously to racial bias or to political pressure, judges might sentence a certain racial group more harshly on average than other groups.[10] In theory, an algorithm might not incorporate race at all in its model, and in practice, the models rarely explicitly incorporate race.

Moreover, relying on big data, a model might develop more accurate predictions of what factors lead to recidivism than a judge who

10. *See, e.g.*, Garrett & Monahan, *supra* note 8, at 478 ("Even judges who believe they rely on many types of information in fact rely 'almost exclusively on prosecutorial recommendation.' Studies have also found troubling evidence that judges rely on an offender's race when making decisions concerning sentencing." (quoting Jeffrey J. Rachlinski & Andrew J. Wistrich, *Judging the Judiciary by the Numbers: Empirical Research on Judges*, 13 ANN. REV. L. & SOC. SCI. 203, 213, 221 (2017))); *see generally* Keith Swisher, *Pro-Prosecution Judges: "Tough on Crime," Soft on Strategy, Ripe for Disqualification*, 52 ARIZ. L. REV. 317 (2010) (discussing pro-prosecution bias in state judiciaries).

has access to, or could process, far less data.[11] Furthermore, these algorithms have been rolled out primarily to lighten the overflowing jail and prison loads. Judges previously might have been detaining all or at least too many of a certain group of criminal defendants, while the new models might find that such detentions are unnecessary to protect the public or to assure appearance at trial.

Algorithmic use, in addition, could eventually eliminate or alleviate the horizontal arbitrariness in the criminal justice system. Defendants in similar circumstances receive different treatment (including prison sentences) from court to court and judge to judge (and occasionally even from the same judge). Intuitively, the length of a defendant's sentence should not depend significantly on the name of the court or judge. If an algorithmic approach was followed uniformly in a state, or even nationally in the case of federal courts, sentencing consistency would be presumably reached or certainly improved.[12]

Alas, this utopian state does not currently exist. Instead, algorithms generally reproduce and in certain circumstances aggravate human biases by replicating and augmenting past human practices. Moreover, consistency has not yet been reached. A particular algorithm has not taken hold nationally, and even though certain algorithms have become increasingly employed, whether a court uses algorithms and to what extent judges rely on them varies significantly in law and in practice. We

---

11. *See, e.g.*, Garrett & Monahan, *supra* note 8, at 452 ("Research has shown that quantitative assessments are more reliable in their predictions than those of individual decision-makers."); *cf.* Carmen de Macedo, *Guilt by Statistical Association: Revisiting the Prosecutor's Fallacy and the Interrogator's Fallacy*, 105 J. PHIL. 320, 320 (2008) (discussing "some legitimate probabilistic indicators of guilt" in criminal justice, although not in the algorithmic context).

12. Notwithstanding accompanying drawbacks, some evidence exists that the federal sentencing guidelines made progress on this front, particularly when the guidelines were mandatory:

> Punishment is more certain and predictable, allowing the parties to better anticipate the sentencing consequences of case facts, and allowing the system to better predict the impact of changes in policy on prison populations and correctional resources. Sentence severity has been increased for many types of crime, in some cases substantially. Most important, the guidelines do not admit consideration of factors, such as race or ethnicity, that are irrelevant to the purposes of sentencing. There is less inter-judge disparity for similar offenders committing similar offenses.

U.S. SENT'G COMM'N, FIFTEEN YEARS OF GUIDELINES SENTENCING: AN ASSESSMENT OF HOW WELL THE FEDERAL CRIMINAL JUSTICE SYSTEM IS ACHIEVING THE GOALS OF SENTENCING REFORM, at xvi (2004), https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-projects-and-surveys/miscellaneous/15-year-study/15_year_study_full.pdf [https://perma.cc/X4KT-N5L4].

have some evidence of under-reliance,[13] but dangers of over-reliance also loom.[14] Thus, the future hope of reducing racial and jurisdictional disparities in detention and sentencing decisions will be realized, if at all, long into the future.

This Article focuses mostly on a separate concern with algorithms (at least as currently built)—simply stated, guilt by association. The pretrial detention and sentencing models at issue base their scores on data from other defendants. If, for example, other defendants from broken homes or from crime-stricken neighborhoods have jumped bail or have reoffended, the defendant (who also came from a broken home or a crime-stricken neighborhood) is now saddled with this prediction. Thus, the defendant is guilty by this association.[15] Of course, the defendant is not guilty in the liability sense; instead, the defendant is now labeled as more likely to jump bail or to reoffend. This label has tangible consequences—the defendant is more likely to be locked up in pretrial detention or to be sentenced more harshly if the label influences the judge. It is not clear that judges fully understand, and if so would accept, this form of guilt by association.[16] To be sure, the instruments or

---

13. *See, e.g.*, Garrett & Monahan, *supra* note 8, at 467 ("Indeed, approximately half of the judges stated that they relied equally on the [Nonviolent Risk Assessment] instrument and on their judicial experience in sentencing a drug or property offender, and approximately one-third stated that they relied primarily on their judicial experience.").

14. *See infra* Section IV.A; *see, e.g.*, Andrew Lee Park, *Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing*, UCLA L. REV (Feb. 19, 2019), https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/ [https://perma.cc/95Q8-746R] (discussing a particularly salient type of cognitive bias, known as "automation bias," which "refers to the 'tendency to ascribe greater power and authority to automated aids than to other sources of advice'" (quoting Raja Parasuraman & Dietrich H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381, 391 (2010))).

15. CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 26 (2016) (arguing that details about a "criminal's birth and upbringing" should be irrelevant to their criminal case or sentencing). O'Neil offers an example:

> If a prosecutor attempted to tar a defendant by mentioning his brother's criminal record or the high crime rate in his neighborhood, a decent defense attorney would roar, 'Objection, Your Honor!' And a serious judge would sustain it. This is the basis of our legal system. We are judged by what we do, not by who we are. And although we don't know the exact weights that are attached to these parts of the test, any weight above zero is unreasonable.

Id.

16. Although still controversial, guilt by association does creep into the law, but as compared to algorithmic guilt by association, the more traditional examples tend to be more transparent to the participants (*e.g.*, judges, lawyers, defendants) and involve at least a bit more individual culpability. For example, coconspirators may be held responsible for certain criminal actions of the other coconspirators, and some laws impose various burdens on members of certain gangs.

instructions often come with a disclaimer (*e.g.*, that each case is different) to mitigate the guilt by association, but the fact remains that the methods score the defendant's future behavior on the basis of others' past behavior.[17]

Although the summary above by no means exhausts the many potential benefits and detriments flowing from algorithmic usage, it provides a relevant sample of the promise and dangerousness of algorithmic criminal justice. The stakes are high for both its use and disuse, and thus one might expect to find a robust literature around the implications for judges and judicial ethics. Instead, the judicial ethics of using algorithms in criminal cases occupies under-explored and hazardous territory. For example, if algorithms are color-blind, why do the judges using them still tend to treat defendants of certain races more harshly than others?[18] Apart from a one-off, publicity-generating case here or there, why have judges not faced greater scrutiny for these disparate outcomes, taken greater ethical responsibility for correcting them going forward, or taken greater responsibility for understanding how these algorithms work at all? The American criminal justice system designedly provides a defendant with a barrage of individual rights against the power of the state, but the judiciary has increasingly turned to sophisticated algorithms at critical phases of criminal proceedings. These algorithms currently do not account for individual rights and in a sense are antithetical to individuals. Moreover, the algorithms rely on data from a historically, and currently, racially stratified justice system.

Solutions or mitigators to these concerns are discussed in Part IV, but first I describe current algorithmic usage below for context. I then

---

17. *See infra* Part IV. An important distinction exists between my theory in Section IV.A., which resists comparisons to others' conduct, and comparisons to physical properties of others for identification purposes. The latter phenomena generally are not problematic. *See, e.g.*, Marcello Di Bello, *When Statistical Evidence Is Not Specific Enough*, 199 SYNTHESE 12251, 12264 (2021) ("People are less reluctant to base a verdict of criminal liability when matching shoe prints, matching genetic profiles, etc.—more generally, statistics-based match evidence—show that the suspect is very likely to be the perpetrator.").

18. *See, e.g.*, Tom Simonite, *Algorithms Should've Made Courts More Fair. What Went Wrong?*, WIRED (Sept. 5, 2019, 7:00 AM), https://www.wired.com/story/algorithms-shouldve-made-courts-more-fair-what-went-wrong/; Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. 2007, 2007 (2022) (noting that the biased data being fed into the algorithm is not the sole issue; data is also primarily derived from carceral knowledge sources, *e.g.*, police and court records); Yaniv Yacoby, Ben Green, Christopher L. Griffin & Finale Doshi-Velez, *"If It Didn't Happen, Why Would I Change My Decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment* (Ariz. Legal Studs. Discussion Paper, Paper No. 22-18, 2022), https://ssrn.com/abstract=4203996 (noting that judicial use of pretrial risk assessments may have "increased racial disparities because judges respond to recommendations in more punitive ways when evaluating Black as opposed to white defendants").

discuss how judicial ethics authorities have so far largely neglected to discuss algorithmically informed legal decisions.

II. CURRENT USAGE OF ALGORITHMS IN JUDGING

This Part focuses on the current usage of algorithms, primarily in criminal justice and particularly in pretrial detention, sentencing, and postconviction release determinations.[19] It also preliminarily describes other, more common types of algorithmic usage in judging.

All judges use algorithms, even judges in jurisdictions that do not (yet) employ the more controversial or sophisticated algorithms. Algorithms boast (or perhaps suffer from) many definitions. For these purposes, the following one should suffice: "At its most fundamental level, an algorithm is nothing more than a very precisely specified series of instructions for performing some concrete task."[20]

Thus, in this general sense, judges use algorithms in driving to the courthouse and in putting on robes. Judges also use algorithms in decision-making. For example, judges use calculators (or computers with calculator applications) to calculate damages. Judges use Westlaw or LexisNexis for legal research, and both services use algorithms to find and suggest relevant search results. Judges use law clerks for legal research and opinion drafting, and these clerks (presumably) receive and follow the judges' steps for completing the process. Judges also rule (and if admitted, rely) on evidence resulting in whole or part from algorithms (*e.g.*, statistical reports). All of these uses are common; none is typically controversial.

The focus here is on a new kind of algorithmic use increasingly employed in courts. In particular, judges are using risk assessments,[21]

---

19. *See generally* Garrett & Monahan, *supra* note 8, at 448 ("We are in the midst of a risk assessment revolution in criminal justice. In the pretrial context, New Jersey and Kentucky, for example, along with many local jurisdictions, use the Public Safety Assessment (PSA) developed by Arnold Ventures. State Supreme Courts, such as the Indiana Supreme Court, the Kentucky Supreme Court, the Nebraska Judicial Council, and the New Jersey Supreme Court, have ordered studies or sweeping changes. States are using risk-based instruments to assess conditions of confinement as well." (footnotes omitted)).

20. MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 4 (2020).

21. Garrett & Monahan, *supra* note 8, at 448–49 ("The most widely used definition of risk assessment describes it as the process of using risk factors to estimate the likelihood (i.e., probability) of an outcome occurring in a population. 'Risk factors' are simply variables that (1) statistically correlate with recidivism, and (2) precede recidivism in time. In the case of pretrial decision-making, the relevant population consists of persons facing criminal charges. In the case of sentencing, the relevant population consists of convicted offenders." (footnote omitted)).

based on actuarial methods, in their decisions on pretrial detention, sentencing, and postconviction release.[22] In the typical instance, court (or parole commission) personnel enter the data for the defendant's risk assessment in advance of the bail hearing, sentencing, or parole hearing. Most assessments rely on data from an interview with the defendant or data otherwise collected about the defendant, including the defendant's current charge(s) and the defendant's criminal and court history.[23] For sentencings and postconviction release determinations (*e.g.*, parole and clemency), much of the data would have been collected for, and discussed in, a presentence or other advance report.[24] The predictive turn, however, adds an actuarial risk assessment to the reporting process. The assessment (*e.g.*, COMPAS or PSA) will return, among other features, a risk score for the court personnel and later the judge to consider.[25]

These scores are derived from analyses of a large number of similar cases (sometimes called training data, which is then validated and tested

22. *See generally State v. Loomis*, 2016 WI 68, ¶ 3, 881 N.W.2d 749, 752–53 (noting that initially these risk assessment tools were used only for determining the best supervision and treatment strategies during probation and parole, but now are used for sentencing); Ben Green, *The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness* 594 (Conf. on Fairness, Accountability & Transparency, 2020), https://dl.acm.org/doi/pdf/10.1145/3351095.3372869 [https://perma.cc/35ND-XSAU] ("[R]isk assessments have proliferated in recent years: in 2017, 25% of people in the U.S. lived in a jurisdiction using a pretrial risk assessment, compared to just 10% four years prior. A 2019 scan of 91 U.S. jurisdictions found that more than two-thirds used a pretrial risk assessment." (internal citations omitted)). For certain offenders, even the influential Model Penal Code now recommends in the sentencing context the use of "actuarial instruments or processes, supported by current and ongoing recidivism research, that will estimate the relative risks that individual offenders pose to public safety through their future criminal conduct." MODEL PENAL CODE: SENT'G § 9.08(2) at 632 (Am. L. Inst. 2023).

23. The model's datasets in turn rely on other data (typically from previous defendants and their circumstances) and generally raise a host of other questions. *See, e.g.*, Garrett & Monahan, *supra* note 8, at 456–57 ("[M]any risk assessment instruments used by parole boards, for example, have never been independently validated, have not been validated on in-state populations, are not regularly updated, and are sometimes altered to add additional factors that are not part of the original instrument.").

24. In the postconviction release context, this discussion is intended for those instances in which judges make the decisions (*e.g.*, supervised release of convicted defendants). Parole or clemency board members make release decisions as well, and this Article may offer some general, if less direct, insights for them.

25. For examples of COMPAS and PSA reports, see respectively: DEP'T OF CORR., DIV. OF COMM. CORR., PRE-SENTENCE INVESTIGATION (2014), https://web.archive.org/web/20200726145629/http://www.wispd.org/attachments/article/272/COMPAS%20PSI%20Presentation%20by%20DOC.pdf (hypothetical COMPAS report); MILWAUKEE CNTY., PRETRIAL RISK ASSESSMENT REPORT: DARRELL BROOKS (2021) https://wisconsinwatch.org/wp-content/uploads/2021/06/Brooks_pretrial_assessment.pdf (PSA Report for Darrell E. Brooks).

on additional data). To simplify a bit for purposes of illustration, if the training data revealed that defendants accused of robbery with at least two prior and recent convictions were highly likely to jump bail or to commit another offense, the model would award a high risk score to a future defendant with these same attributes.[26] This score might be a "9" out of 10, for example, perhaps with an accompanying color of alarm (red or yellow). The probation officer and later the judge would see this high score and presumably take it into account in the pretrial release or sentencing decision. Whereas judges of yesteryear would not have had this score in the report, judges nowadays increasingly have this score (or something like it) in front of them before making these critical decisions.

Perhaps instantly obvious, there are important differences between the more traditional (and still more common) algorithmic involvement in a general sense and these newer risk assessment models. Although the newfangled tools present opportunities, they also present unique risks. In particular, they can be "opaque, unquestioned, and unaccountable, and they operate at a scale to sort, target, or 'optimize' millions of people."[27] They can also create runaway feedback loops (*e.g.*, sending more and more police to certain neighborhoods).[28] Unlike older tools, moreover, these tools rely more on the conduct of others to make their predictions.[29] Thus, the way in which judges may ethically use these tools is of critical importance, a point to which we now turn.

---

26. As Renée Jorgensen aptly describes:

> Most of these tools employ straightforward statistical analysis on historical arrest databases, seeking to isolate the strongest correlations between a relatively sparse set of recorded variables and a property representing the target outcome (e.g., failure to appear, another arrest, or arrest for violent offense). There is some variation in the variables used: third generation risk assessment measures improve on the original second generation models by using not only static variables—properties that do not change over time, such as age at first arrest, having a prior conviction, sex, etc.—but also dynamic variables (e.g., years since last offense, employment status, present substance abuse) which are responsive to the subject's current behavior, and can reflect reduced (or increased) risk over time. Simplifying a bit, these tools are ultimately algorithms taking the variable values as inputs, assigning them weights, and outputting an estimate of how often someone with those features in the database ends up with the target outcome.

Renée Jorgensen, *Algorithms and the Individual in Criminal Law*, 52 CANADIAN J. PHIL. 61, 62 (2022) (footnote omitted).

27. O'NEIL, *supra* note 15, at 12 (noting also that they can create "pernicious . . . feedback loops").

28. Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger & Suresh Venkatasubramanian, *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACH. LEARNING RSCH. 1, 1–12 (2018); O'NEIL, *supra* note 15, at 86–87, 97 (discussing similar concerns with predictive policing algorithms).

29. *See infra* Section IV.A (discussing a legal and moral objection to this focus if used to harm, rather than help, a defendant).

III. ALGORITHMIC USAGE AND JUDICIAL ETHICS

This Part examines the existing judicial ethics architecture to see whether it anticipates and adequately addresses the increasing algorithmic reliance described above. Although judicial ethics in this context has been largely unexplored,[30] judicial ethics is impactful in the process leading up to, and during, judicial decision-making. Judges face professional jeopardy for failing to follow or exhibit judicial ethics. In addition, judges generally strive for excellence in judging and for the respect of their peers, all of which depends, in part, on their embodiment of judicial ethics. This Part explains the judicial disciplinary system and its ethical canons (overarching principles), followed by its rules (applications of the principles) and pertinent disciplinary decisions. This Part also highlights the need for evolution in this area.

A. The System and Canons of Judicial Ethics

By way of regulatory background, an ethical code binds all judges in the United States.[31] Judges who commit transgressions of this ethical code fall below their peers' standards and are subject to discipline.[32] Discipline generally refers to tangible consequences for professional

30. For example, in a recent, fifty-four page article on risk assessments in judicial practice, the words "judicial ethics" or "judicial conduct" are never mentioned, although the article is quite insightful on judicial practice in this area. Garrett & Monahan, *supra* note 8. One recent article, focusing overall more on lawyers than judges and more on legal practice than criminal courts, is Mark L. Shope, *Lawyer and Judicial Competency in the Era of Artificial Intelligence: Ethical Requirements for Documenting Datasets and Machine Learning Models*, 34 GEO. J. LEGAL ETHICS 191, 211 (2021) (arguing for "the mandatory use of Dataset Disclosure Forms and Model Disclosure Forms" when lawyers or judges use AI models).

31. Cynthia Gray, *The Line Between Legal Error and Judicial Misconduct: Balancing Judicial Independence and Accountability*, 32 HOFSTRA L. REV. 1245, 1246 n.4 (2004) (noting that "[f]orty-nine states, the U.S. Judicial Conference, and the District of Columbia have adopted codes based on (but not identical to) either the 1972 or 1990 model codes" of judicial conduct and that the remaining state also has rules of conduct for judges but which are not based on either model code).

32. Both federal and state judges are subject to discipline for professional misconduct, but as to federal judges, the ethical rules are "instructive" but not dispositive. *See, e.g.*, CHARLES GARDNER GEYH, JAMES J. ALFINI & JAMES SAMPLE, JUDICIAL CONDUCT AND ETHICS § 1.06 (6th ed. 2020) ("While the Judicial Conference has acknowledged that sometimes Code violations can serve as a basis for discipline, the disciplinary process in the federal system operates independently of the Code and subjects judges to limited forms of discipline for conduct 'prejudicial to the effective and expeditious administration of the business of the courts.'" (quoting 28 U.S.C. § 351(c)) (footnotes omitted)); James J. Alfini, *Supreme Court Ethics: The Need for Greater Transparency and Accountability*, 21 PRO. LAW. 10, 10 (2012).

misconduct: reprimand, censure, suspension, or removal from the bench.[33]

Overarchingly, the ethical code seeks to promote three "I's" in judges: integrity, impartiality, and independence.[34] (It also seeks to avoid a negative "I": impropriety and its appearance to the public.[35]) Integrity generally means that the judge must comply with the law in and out of the courtroom and exhibit "probity, fairness, honesty, uprightness, and soundness of character."[36] Impartiality means the "absence of bias or prejudice in favor of, or against, particular parties or classes of parties, as well as maintenance of an open mind in considering issues that may come before a judge."[37] Independence, finally, is defined as "a judge's freedom from influence or controls other than those established by law."[38]

In the spirit of these three general values above,[39] individual ethical rules command or forbid specific action, as listed below.

B. Ethical Code: Rules

The rules of judicial ethics provide some guidance to, and constraints on, judges considering algorithmic tools. These include rules addressing (1) fairness and impartiality, (2) competence, (3) ex parte communications and factual research, and (4) supervisory obligations.

---

33. CYNTHIA GRAY, A STUDY OF STATE JUDICIAL DISCIPLINE SANCTIONS 3, 6 (2002) ("Each of the 50 states and the District of Columbia has established a judicial conduct organization charged with investigating and prosecuting complaints against judicial officers.").

34. MODEL CODE OF JUD. CONDUCT Canon 1 (AM. BAR ASS'N 2007) ("A judge shall uphold and promote the independence, integrity, and impartiality of the judiciary, and shall avoid impropriety and the appearance of impropriety."); *see generally* JUDICIAL ETHICS (Keith Swisher ed., 2016).

35. *See generally* JUDICIAL ETHICS, *supra* note 34. Impropriety "includes conduct that violates the law, court rules, or provisions of this Code, and conduct that undermines a judge's independence, integrity, or impartiality." MODEL CODE OF JUD. CONDUCT Terminology; *id.* r. 1.2 cmt. 5 ("Actual improprieties include violations of law, court rules or provisions of this Code. The test for appearance of impropriety is whether the conduct would create in reasonable minds a perception that the judge violated this Code or engaged in other conduct that reflects adversely on the judge's honesty, impartiality, temperament, or fitness to serve as a judge.").

36. MODEL CODE OF JUD. CONDUCT Terminology; *id.* r. 1.1 ("A judge shall comply with the law, including the Code of Judicial Conduct."). Law "encompasses court rules as well as statutes, constitutional provisions, and decisional law." *Id.* Terminology.

37. *Id.* Terminology.

38. *Id.*

39. In spirit and command: "A judge shall act at all times in a manner that promotes public confidence in the independence, integrity, and impartiality of the judiciary, and shall avoid impropriety and the appearance of impropriety." *Id.* r. 1.2.

1. Fairness and Impartiality

Judges must "accord to every person who has a legal interest in a proceeding, or that person's lawyer, the right to be heard according to law."[40] In providing this legal forum, judges must "perform all duties of judicial office fairly and impartially" and "must be objective and open-minded."[41] Judges thus must maintain objectivity and open-mindedness even when seemingly accurate algorithms are involved and must be willing to consider seriously any countervailing information or arguments.

Of particular note for algorithmic fairness questions, judges must "perform the duties of judicial office, including administrative duties, without bias or prejudice."[42] This includes the following prohibition:

> [Judges must] not, in the performance of judicial duties, by words or conduct manifest bias or prejudice, . . . including but not limited to bias, prejudice, or harassment based upon race, sex, gender, religion, national origin, ethnicity, disability, age, sexual orientation, marital status, socioeconomic status, or political affiliation, and shall not permit court staff, court officials, or others subject to the judge's direction and control to do so.[43]

This prohibition, however, does "not preclude judges or lawyers from making legitimate reference to the listed factors, or similar factors, when they are relevant to an issue in a proceeding."[44] Algorithmic, artificial intelligence, and machine learning models can learn unethical biases from the programmers or the data itself. To rule fairly, impartially, and without bias, a judge should not defer to models whose outputs might be infected with bias.

In sum, judges have a duty to approach and render their decisions open-mindedly, fairly, and impartially, without bias or prejudice.

---

40. *Id.* r. 2.6(A) & cmt. 1 ("The right to be heard is an essential component of a fair and impartial system of justice. Substantive rights of litigants can be protected only if procedures protecting the right to be heard are observed."). Furthermore, a judge may "make reasonable accommodations to ensure [unrepresented] litigants the opportunity to have their matters fairly heard." *Id.* r. 2.2 cmt. 4. Certain algorithmic approaches place pressure on this duty, in light of the opacity of the algorithm and whether (especially pretrial) the defendant is given a realistic opportunity to challenge the algorithm or the data it used.

41. *Id.* r. 2.2 & cmt. 1.

42. *Id.* r. 2.3(A).

43. *Id.* r. 2.3(B).

44. *Id.* r. 2.3(D).

2. Competence and Diligence

In addition to fairness and impartiality, judges must "perform judicial and administrative duties, competently and diligently."[45] To act competently "requires the legal knowledge, skill, thoroughness, and preparation reasonably necessary to perform a judge's responsibilities of judicial office."[46] Judges may understandably be new to the algorithmic approaches discussed in this Article, and to gain competence, "[a] judge should seek the necessary docket time, court staff, expertise, and resources to discharge all adjudicative and administrative responsibilities."[47] Although aimed more toward unnecessary disqualification, the Code notes that judges have a "responsibility to decide," in that they must "hear and decide matters assigned to the judge."[48] Combining these duties suggests that although "court staff, expertise, and resources" may assist judges in competently discharging their duties,[49] judges may not outsource themselves completely.[50] In other words, judges must judge, not simply rubberstamp algorithmic outputs.

Furthermore, the duties of competence and diligence impose, or at least imply, a responsibility to investigate algorithmic models, or at a minimum their outputs, to ensure accuracy and fairness, including the absence of bias.[51] Conversely, these duties imply that judges may need

---

45. MODEL CODE OF JUD. CONDUCT r. 2.5(A).

46. *Id.* r. 2.5 cmt. 1.

47. *Id.* r. 2.5 cmt. 2; *cf.* ABA Comm. on Ethics & Pro. Resp., Formal Op. 512 (2024) ("[A] lawyer's reliance on, or submission of, a [generative AI] tool's output—without an appropriate degree of independent verification or review of its output—could violate the duty to provide competent representation as required by [the ethical rules].").

48. MODEL CODE OF JUD. CONDUCT r. 2.7 & cmt. 1 ("Judges must be available to decide the matters that come before the court."). This at least implies that judges should decide matters themselves, not turn over the work to someone, or something, else.

49. *Id.* r. 2.5 cmt. 2.

50. *See id.* r. 2.9(A)(3) (noting that judges have "the responsibility personally to decide the matter").

51. The ethical code has not had a comprehensive refresh since 2007 (when the most recent version was adopted). Thus, it was drafted when machine learning was in its relative infancy. Although drafted before much of today's algorithmic usages, the official comment provides examples of forbidden conduct:

> Examples of manifestations of bias or prejudice include but are not limited to epithets; slurs; demeaning nicknames; negative stereotyping; attempted humor based upon stereotypes; threatening, intimidating, or hostile acts; *suggestions of connections between race, ethnicity, or nationality and crime*; and irrelevant references to personal characteristics. Even facial expressions and body language can convey to parties and lawyers in the proceeding, jurors, the media, and others an appearance of bias or prejudice. *A judge must avoid conduct that may reasonably be perceived as prejudiced or biased.*

to investigate and potentially adopt algorithmic approaches. As noted elsewhere, the future may bring algorithmic tools that far exceed human judicial capacity.[52] Indeed, it could one day reach a point at which a competent and diligent judge would have a responsibility to employ algorithmic assistance.

In sum, judges cannot incompetently or completely defer to algorithmic assessments in judicial decision-making.[53] Bias may infiltrate the data on which the algorithms rely, and in addition to the general duty to avoid bias, judges have a general duty of competence and diligence toward each case and toward any models they use to help decide the case. Moreover, it almost goes without saying that judges must decide cases on the basis of the actions or inactions of the parties before the judge, not on the actions of parties in other cases. Because current algorithmic models use the actions and outcomes from other defendants' cases—not the defendant before the judge—over-reliance on algorithms might mean not only a dereliction of the duty to decide competently but also a decision involving the wrong case.

Id. r. 2.3 cmt. 2 (emphases added).

Algorithms sometimes suggest connections between race and crime, for example, and may appear to be biased. *See, e.g.*, Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias: There's Software Used Across the Country To Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/E29S-RMTN]; Jeff Larsen, Surya Mattu, Lauren Kirchner & Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perma.cc/BAW9-LFTG]. *But see* WILLIAM DIETERICH, CHRISTINA MENDOZA & TIM BRENNAN, COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY 3 (2016), https://assets.documentcloud.org/documents/2998391/ProPublica-Commentary-Final-070616.pdf [https://perma.cc/4QAB-82D9] (rejecting claims of bias).

52. *See, e.g.*, Angwin, Larson, Mattu & Kirchner, *supra* note 51 (arguing that in the future computer predictions could make the criminal justice system fairer); Shope, *supra* note 30, at 196 (noting that "[m]achine learning is becoming more accurate").

53. West Virginia's Judicial Investigation Commission very recently confirmed this. W. Va. Jud. Investigation Comm'n, Advisory Op. 2023-22, at 4 (2023) (stating that a judge "should never use AI to reach a conclusion on the outcome of a case," in part "because of perceived biases that may be built into the program" and that the judge "alone is responsible for determining the outcome of all proceedings"); *see also* State Bar of Mich. Jud. Ethics Standing Comm., Op. JI-155 (2023) ("[The code of conduct] could be triggered, for example, if a judicial officer uses an AI solution that is considered partial or unfair and may influence the judicial officer's judgment. This could occur if the tool's algorithm or training data creates bias." (citation omitted)); Fla. Bd. Rev. Comm. on Pro. Ethics, Advisory Op. 24-1 (2024) (noting that "a lawyer may not delegate to generative AI any act that could constitute the practice of law such as the negotiation of claims or any other function that requires a lawyer's personal judgment and participation").

3. Ex Parte Communications and Factual Research

Although of less frequent intersection with current algorithmic approaches, it may be worth flagging for foundational purposes and future developments that judges generally may not engage in ex parte communications or conduct factual research.

In particular, judges may "not initiate, permit, or consider ex parte communications, or consider other communications made to the judge outside the presence of the parties or their lawyers, concerning a pending or impending matter."[54] Judges may nevertheless consult with court staff and officials "whose functions are to aid the judge in carrying out the judge's adjudicative responsibilities, or with other judges, provided the judge makes reasonable efforts to avoid receiving factual information that is not part of the record, and does not abrogate the responsibility personally to decide the matter."[55] Furthermore, judges may "not investigate facts in a matter independently, and shall consider only the evidence presented and any facts that may properly be judicially noticed."[56] This "prohibition against a judge investigating the facts in a matter extends to information available in all mediums, including electronic."[57]

Algorithms often rely on factual information from other cases, a problem rarely noted in the literature. If the law authorizes or requires judges to use these algorithms, however, it might not be a violation of the judicial ethics rules to do so provided that the report is disclosed to the parties.[58] The judge, in addition, should ensure that the defendant (and, if interested, the prosecutor) has the chance to learn and scrutinize the information. Furthermore, for judges who wish to use these tools, particularly ones utilizing factual information from other cases in jurisdictions in which the law does not clearly authorize them, the judge should research this question or request an ethics opinion on the issue.

In sum, if judges were to communicate with third-party algorithm vendors about a particular legal matter or if algorithmic products were used to conduct factual research about the case at hand, the conduct might violate the ethical rules. Currently, however, this conduct does not seem prevalent.

---

54. MODEL CODE OF JUD. CONDUCT r. 2.9(A).

55. *Id.* r. 2.9(A)(3).

56. *Id.* r. 2.9(C).

57. *Id.* r. 2.9 cmt. 6.

58. In helpful comments on an earlier draft, Professor Charles Gardner Geyh correctly noted that a legislature's imposition of certain requirements on state judges could infringe on separation of powers principles. In practice, the legislatures thus far seem to be winning (or at least not sufficiently challenged) on the algorithms at issue.

4. Supervisory Rules

Finally, judges may bear responsibility for the actions of others. Even if others (*e.g.*, court employees) are the ones taking the ethically offensive actions, judges are still responsible for those actions.[59] Likewise, judges who order an action that is actually completed by another do not absolve themselves of disciplinary liability.[60] Thus, for example, if a risk assessment discriminated against Black defendants by design, but the court's probation personnel or clerks procured the assessment, the judge could not escape discipline by simply pointing the finger at those personnel. Similarly, if the judge's clerk used the algorithm or an AI tool, the judge is still responsible if the output is erroneous or biased. Both personnel and tools must be supervised.

* * *

In sum, the ethical rules provide some generalized guidance to judges in using, or considering the use of, algorithmic tools. The rules do not, however, specifically address algorithms, even in the official commentary. We therefore need to explore additional ethical sources. Judicial disciplinary decisions, discussed immediately below, provide the most analogous guidance.

---

59. The Model Code states:

> A judge is responsible for his or her own conduct and for the conduct of others, such as staff, when those persons are acting at the judge's direction or control. A judge may not direct court personnel to engage in conduct on the judge's behalf or as the judge's representative when such conduct would violate the Code if undertaken by the judge.

MODEL CODE OF JUD. CONDUCT r. 2.12 cmt. 1. *Cf.* Fla. Bd. Rev. Comm. on Pro. Ethics, Advisory Op. 24-1 (2024) ("[J]ust as a lawyer must make reasonable efforts to ensure that a law firm has policies to reasonably assure that the conduct of a nonlawyer assistant is compatible with the lawyer's own professional obligations, a lawyer must do the same for generative AI."); ABA Comm. on Ethics & Pro. Resp., Formal Op. 512 (2024) ("Managerial lawyers must establish clear policies regarding the law firm's permissible use of [generative AI], and supervisory lawyers must make reasonable efforts to ensure that the firm's lawyers and nonlawyers comply with their professional obligations when using [generative AI] tools.").

60. *See supra* note 59. *Cf.* MODEL RULES OF PRO. CONDUCT r. 8.4(a) (AM. BAR ASS'N 1983) (noting that attorneys cannot break the ethical rules by using an agent, *e.g.*, a legal assistant).

C. Judicial Disciplinary Decisions

Judges can be disciplined (*e.g.*, removed from the bench) for violating the ethical rules, including those described above.[61] Although virtually no judicial disciplinary decisions are exactly on point in this new frontier, a body of disciplinary caselaw provides some analogous constraints.

Perhaps most directly, judges cannot abrogate their duty to decide according to the law and facts of the case. Thus, judges who have used, or appeared to use, tools or gimmicks for deciding cases, such as flipping a coin or asking courtroom spectators to decide the case or the legal question, have been disciplined.[62] In addition, judges who have delegated their decision-making responsibility to one of the parties in a case (*e.g.*, the prosecution) have also been disciplined.[63] Thus, a judge cannot completely defer to an algorithm, no matter how seemingly accurate.

Judges also have been disciplined for failing to permit a party to be heard according to law or to allow the parties to make the record.[64] This is relevant to the algorithmic context. Many algorithms are technical,

---

61. However, the federal disciplinary regime generally gives less force to the ethical rules and has available fewer forms of punishment for offending judges. *See supra* note 32 and accompanying text.

62. *See, e.g.*, *In re Best*, 719 So. 2d 432, 434 (La. 1998) (censuring a judge in part for asking the courtroom audience: "If you think I ought to find him not guilty, will you stand up?"); *see also* JEFFREY M. SHAMAN, STEVEN LUBET & JAMES J. ALFINI, JUDICIAL CONDUCT AND ETHICS 39–40 (3d ed. 2000) (discussing cases in which "judges make a decision by flipping a coin in open court, or by throwing a dart at a dart board, or by taking a vote of the spectators in the courtroom" and denouncing this type of behavior as "a complete abdication of the duty to exercise judgment").

63. *In re Sanchez*, 512 P.2d 302, 303 (Cal. 1973) (censuring a judge because he effectively allowed the bondsman to determine the amount of bail); Stephen B. Bright & Patrick J. Keenan, *Judges and the Politics of Death: Deciding Between the Bill of Rights and the Next Election in Capital Cases*, 75 B.U. L. REV. 759, 803–11 (1995) (documenting instances of improper delegation to the prosecution in capital cases); Gerald Stern, *Is Judicial Discipline in New York State a Threat to Judicial Independence?*, 7 PACE L. REV. 291, 333 (1987) (noting the discipline of a judge who allowed the district attorney to draft the judge's order). Somewhat analogously, the ABA Ethics Committee concluded that prosecutors act unethically when they permit private companies to use the letterhead of the prosecutorial agency to collect debts, concluding that the practice creates a false impression that the prosecutors reviewed the facts and determined that prosecution is warranted in the specific matter. ABA Comm. on Ethics & Pro. Resp., Formal Op. 469, 2–3 (2014).

64. *See, e.g.*, *In re Complaint Against Staley*, 486 N.W.2d 886, 902 (Neb. 1992) (removing judge in part for preventing a record of the proceedings); *In re Dash*, 564 S.E.2d 672, 673 (S.C. 2002) (censuring judge for finding defendant guilty without allowing defendant to present evidence); *Miss. Comm'n on Judicial Performance v. Wells*, 2001-JP-00229-SCT (¶ 1) 1794 So. 2d 1030, 1031–32 (Miss. 2001) (publicly reprimanding judge for entering guilty verdict without a hearing); *In re Milhouse*, 605 N.W.2d 15, 15–16 (Mich. 2000) (similar).

opaque, or proprietary, and thus the defendant may face significant difficulty learning about and challenging a particular algorithm's result. These issues suggest that judges should pay respect to these challenges, permit defendants to discover the algorithm's workings, engage qualified experts, file motions or lodge objections, and make an adequate record for possible appeal.[65]

As a final, pertinent example, judges have been disciplined for exhibiting racial bias in (and out of) the courtroom. Thus, although proof of judges' intent is often hard to obtain, judges who rule based on racial animus are disciplined.[66] Similarly, judges who use racist language are disciplined.[67] As algorithms have the potential to discriminate on the basis of race and may regurgitate and even amplify racial biases in the data to which the algorithm was applied, judges should scrutinize the algorithmic inputs and outputs to ensure that the algorithm is not manifesting inappropriate bias or prejudice.[68]

Thus, caselaw provides some (wide) guiderails to algorithmic usage. As noted above, judges must remain engaged, impartial, and open-minded and must refrain from manifesting bias on the basis of race (or other protected classes). Moreover, as discussed in Part IV, current algorithmic tools may not directly violate these general rules, but they present plenty of cause for scrutiny. The next section urges judicial ethics tailoring to this new context.

---

65. *See State v. Guise*, 921 N.W.2d 26, 34 (Iowa 2018) (Appel, J., concurring) ("[I]f the state intends to offer risk assessments for the court to rely upon in sentencing, the defendant has a right to an adequate opportunity to attack it. If the court does not give the defendant [this] opportunity . . . , it should not be utilized in sentencing."); Hillman, *supra* note 1, at 37 ("An algorithm-generated risk assessment score presents itself to the court as a presumptive factual determination. In essence, predictive technology becomes another witness against the defendant without a concomitant opportunity to test the data, assumptions, and even prejudices that underlie the conclusion. A predictive recidivism score may emerge oracle-like from an often-proprietary black box.").

66. *See, e.g.*, *In re Quirk*, 705 So. 2d 172, 178 n.14 (La. 1997) (disciplining a judge in part because his "primary consideration in his setting of bail [was] the fact that the racial group to which the defendants belonged 'voted against' [the judge's] brother in an election").

67. *See, e.g.*, *In re Baldwin*, CJC Nos. 19-1291 & 19-1160, at 2 (Tex. Comm'n on Jud. Conduct Apr. 9, 2021), https://www.scjc.texas.gov/media/46829/baldwin19-1297-19-1160pubrep-oae4921.pdf [https://perma.cc/HY98-R2G3] (reprimanding and ordering education for a judge who stated that a Black defendant "needs to be hung," among other misconduct); *cf. In re Poe*, Complaint No. 17-2021, at 2 (W. Va. Jud. Investigation Comm'n Mar. 12, 2021), https://www.courtswv.gov/sites/default/pubfilesmnt/2023-07/17-2021MagCharlesPoe.pdf (admonishing and accepting resignation of a judge who made "inappropriate sexual, homophobic and racist comments" verbally and by text message to court personnel).

68. As noted above, manifesting bias or prejudice violates the ethical rules. *See supra* Section III.B.1.

D. Judicial Ethics: Taking Stock and Making Preliminary Wishes

As shown above, judicial ethics provides a general framework for guiding judges in their algorithmically aided decisions. It does not, however, provide particular guidance or even necessarily contemplate algorithmic uses and advances. Furthermore, judicial ethics authorities have not addressed the issues of guilt by association and the inclusion of systemic biases in algorithmic deployment. Therefore, updates and applications are needed, as discussed below.

One particularly startling absence appears to be in judicial ethics advisory opinions. Formal bodies regularly provide advisory opinions to judges to help navigate the ethical and professionalism issues arising in judicial practice.[69] Surprisingly, virtually no advisory opinions have been issued in this area. Perhaps the committees have not to date recognized that the use or misuse of these algorithmic models might violate the judicial ethics rules in the ways noted above (*e.g.*, discrimination, improper deference). Three templates for these needed opinions follow.

First, the opinions could analyze a particular algorithm or several types of algorithms used in the criminal justice system to explain the pertinent ethical rules (many of which are highlighted above) and apply them (*i.e.*, what type of tool or usage would run afoul of rules requiring judicial competence or the absence of bias or prejudice). Second, the opinions could holistically provide guidance on how a judge should assess and apply an algorithm or category of algorithms when making decisions. Third, the opinions could analyze what the judiciary should look for and be skeptical of when deciding whether to procure and use an algorithm in judicial decision-making. Although all three of these interrelated topics are briefly addressed in this Article, a formal judicial ethics opinion issued by an official body is more generally impactful than an article and would provide a new voice in this important area.

Indeed, as the dearth of opinions illustrates, judicial ethics sources (codes, cases, opinions) have paid insufficient attention to algorithmic uses in criminal (and civil) court cases. This is surprising given the American Bar Association's recent resolution essentially calling for this

---

69. *See* Cynthia Gray, *Helping Judges Look Before They Leap: Judicial Ethics Advisory Committees*, 54 CT. REV. 152, 152 (2019) (noting that almost all states have a judicial ethics committee and that, in 2017, these committees "posted more than 325 advisory opinions online," most frequently advising about "acting as a reference; court staff issues; financial activities; the conduct of senior or part-time judges; teaching, writing, and speaking; and political and campaign conduct"). Although AI has garnered some recent attention in ethics opinions, algorithmic models in criminal courts have not. *See generally* ABA Comm. on Ethics & Pro. Resp., Formal Op. 512 (2024) ("It is anticipated that this Committee and state and local bar association ethics committees will likely offer updated guidance on professional conduct issues relevant to specific [generative AI] tools as they develop.").

attention. Although it was couched in "AI" terminology, its warnings apply to this highly related context.[70] One area in the judicial ethics code is notably ripe for an update: incorporating a technological competence component into the general competence rule.[71] This is particularly important in the algorithmic arena, and the resulting update should consider its unique challenges.[72]

Better late than never, judicial regulatory authorities should begin considering these issues, issuing guidance, and refining that guidance as new insights and technology emerge.[73] Deep into the future, a system of artificial intelligence, in addition to or in lieu of humans, may be bound

---

70. ABA HOUSE OF DELEGATES, RESOLUTION 112 (2019), https://www.americanbar.org/content/dam/aba/directories/policy/annual-2019/112-annual-2019.pdf (urging "courts and lawyers to address the emerging ethical and legal issues related to the usage of artificial intelligence ("AI") in the practice of law including: (1) bias, explainability, and transparency of automated decisions made by AI; (2) ethical and beneficial usage of AI; and (3) controls and oversight of AI and the vendors that provide AI"); Nicolas Economou, *Artificial Intelligence and the Law: The ABA's Important and Timely Contribution*, ABA: LEGAL ANALYTICS COMM. NEWSLETTER (Aug. 26, 2019), https://web.archive.org/web/20200920025642/https://www.americanbar.org/groups/business_law/publications/committee_newsletters/legal_analytics/2019/201908/ai_law/. Of note, the ABA generally supports pretrial risk assessment. STANDARDS FOR CRIM. JUST.: PRETRIAL RELEASE, § 10-1.10 (AM BAR ASS'N 2002). Moreover, the Conference of Chief Justices has likewise issued support. CONF. OF CHIEF JUSTICES & CONF. OF STATE CT. ADM'RS, RESOLUTION 12: IN SUPPORT OF SENTENCING PRACTICES THAT PROMOTE PUBLIC SAFETY AND REDUCE RECIDIVISM (2007), https://ccj.ncsc.org/__data/assets/pdf_file/0019/23536/08012007-support-sentencing-public-safety-reduce-recidivism.pdf [https://perma.cc/Z73A-Y7EY]).

71. As in Indiana, the competence rule should at least read: "the legal knowledge, skill, thoroughness, and preparation reasonably necessary to perform a judge's responsibilities of judicial office, including the benefits and risks associated with the *technology relevant to service as a judicial officer*." Shope, *supra* note 30, at 203 (citing IND. CODE OF JUD. CONDUCT r. 2.5 cmt. 1). Surprisingly, while citing Indiana, the previous article does not mention that the italicized language is missing from the Model Code of Judicial Conduct, which is the Code that influences most jurisdictions. *Compare* MODEL CODE OF JUD. CONDUCT r. 2.5 cmt. 1, *with* IND. CODE OF JUD. CONDUCT r. 2.5 cmt. 1 (2018). The corresponding ethical code for lawyers (the Model Rules of Professional Conduct) already includes a technological competence component. MODEL RULES OF PRO. CONDUCT r. 1.1 cmt. 8.

72. *Cf. Loomis*, 2016 WI 68, ¶ 4, 881 N.W.2d at 753 ("As data changes, our use of evidence-based tools will have to change as well. The justice system must keep up with the research and continuously assess the use of these tools."). The Michigan State Bar recently acknowledged this point, at least generally. State Bar of Mich. Jud. Ethics Standing Comm., Op. JI-155 (2023) ("[A]s AI rapidly advances, judicial officers have an ethical duty to maintain technological competence and understand AI's ethical implications to ensure efficiency and quality of justice."); *see also* State Bar of W. Va. Jud. Investigation Comm'n, Op. 2023-22 (2023).

73. As noted in Part I, algorithms can offer ethical potential, not just danger. *See, e.g.*, Raymond J. McKoski, *Rewriting Judicial Recusal Rules with Big Data*, 2020 UTAH L. REV. 383, 388, 410–17 (discussing "how big data can cure shortcomings in appearance-based disqualification").

by these rules and standards of judicial ethics.[74] The judicial ethics framework explained in this Part has endured for decades, and it needs an update if it is to offer applicable and wise guidance in the future. Considering generally the judicial ethics principles above and incorporating additional legal and moral principles below, the next Part offers three approaches to judges to assist them in ethically addressing algorithms in their cases.

IV. ALGORITHMIC CRIMINAL JUSTICE: THREE APPROACHES TO BOOST JUDICIAL ETHICS

This Part offers three distinct approaches to increase judicial responsibility in the use of criminal justice algorithms. Two of the approaches are directed to the individual judge: (1) the Principle of Personally Identifiable Punishment and (2) the Principle of "Caution About Outputs."[75] The third approach is aimed at the judiciary: the Principle of Most Favorable Demographic Parity. If judges and the judiciary follow these approaches, they will pay greater respect to judicial ethics, due process, and equal protection in criminal cases.

A. The Principle of Personally Identifiable Punishment

From an ethical standpoint, this Article focuses primarily on the following concern with algorithms (at least as currently built): guilt by association. This Section briefly builds on this concern, updates it, and proposes a rule to avoid the ills of guilt by association.

The pretrial detention, sentencing, and postconviction release models at issue largely base their scores on data from other defendants. If, for example, other defendants from broken homes or crime-stricken neighborhoods have jumped bail or have reoffended, the defendant (who also came from a broken home or a crime-stricken neighborhood) is now

74. If the judiciary were to turn on an algorithmic autopilot (contrary to what is currently permissible) to decide cases, that autopilot would presumably need to follow the same ethical rules. This would include exhibiting the three key "I's" of modern judicial ethics (integrity, impartiality, and independence) and avoiding the negative "I" (impropriety and even the appearance of impropriety). I would like to explore this and related questions in a future work, including: Do we have an obligation (and from where?) to craft the perfect algorithmic judge? This judge, if crafted well, would not suffer from human limitations and could, again if crafted well, mimic or perhaps even exceed human capacities.

75. Both the label "Caution About Outputs" and the approach, adapted below, stem from Annette Zimmermann & Chad Lee-Stronach, *Proceed with Caution*, 52 CANADIAN J. PHIL. 6 (2022).

saddled with this prediction.[76] Thus, the defendant is guilty by this association.[77] It is as if a judge were to detain or sentence Defendant A based on Defendant B's behavior in another matter. Although the defendant is not deemed guilty in a liability or verdict sense, the defendant is labeled as likely to skip bail or to reoffend; this is the defendant's model-inferred guilt. Serious consequences to the defendant may follow from it: the defendant is more likely to be locked up in pretrial detention, to be sentenced more harshly, or not to be released from prison.

Of course, guilt by association is not unique to these models. For example, if a defendant associates with gang or conspiracy members, and if solely by virtue of that association the jury convicts the defendant for the criminal activities of the codefendants, this is guilt by association.[78] As another more egregious example, judges have sentenced defendants more harshly if they were Black.[79] The defendant thus was associated, in a sense, with other Black defendants. Both examples are legally (and morally) problematic, which is why the law has long banned guilt by association.[80] Guilt by association generally violates defendants' rights; these defendants—their actus reus and mens rea—are on trial, not those with whom they appropriately or inappropriately associated. Morally, moreover, guilt by mere association may not respect the individual's autonomy, among other ills.

The predictive turn in criminal justice, however, has snuck guilt by association into the law. In several states, this form of guilt by association

76. *See, e.g.*, *Loomis*, 2016 WI 68, ¶ 15, 881 N.W.2d at 754 ("[R]isk scores are intended to predict the general likelihood that those with a similar history of offending are either less likely or more likely to commit another crime following release from custody. However, the COMPAS risk assessment does not predict the specific likelihood that an individual offender will reoffend. Instead, it provides a prediction based on a comparison of information about the individual to a similar data group.").

77. For a discussion of guilt by association, see *supra* note 16.

78. *See, e.g.,* Mitchell Eisen, Brenna Dotson & Gregory Dohi, *Probative or Prejudicial: Can Gang Evidence Trump Reasonable Doubt?*, 62 UCLA L. REV. DISCOURSE 2, 12–13 (2014).

79. *See, e.g.,* Rachlinski & Wistrich, *supra* note 10, at 221.

80. *See United States v. Polasek*, 162 F.3d 878, 886–87 (5th Cir. 1998) (reversing conviction because the improper "guilt by association evidence likely had substantial impact on the jury's verdict"); *United States v. Diffoot*, 54 M.J. 149, 153 (C.A.A.F. 2000) ("[P]ermitting convictions on the basis of a theory of guilt by association would establish 'a principle alien to American standards of justice.'" (quoting *United States v. Jacobs*, 1 C.M.A. 209, 211 (1952))). Guilt by association also can have negative First Amendment (freedom of association) implications, but those are beyond the scope of this Article.

is now legally permissive or even required.[81] These newly legal statistical methods necessarily draw from other defendants. Using the (mis)behavior of other defendants, the methods lodge a prediction about the defendant's likely behavior (*e.g.*, recidivism or flight risk).[82] In other words, through the (mis)behavior or circumstances of *other defendants*, the methods lodge a prediction about *this defendant's* likely behavior. To be sure, the instruments or instructions often come with a disclaimer (*e.g.*, that each case is different). These disclaimers attempt to mitigate guilt by association, but they also serve as an admission: The methods categorize the defendant's future behavior on the basis of others' past behavior.

Admittedly, this Article's use of "guilt by association" is in a sense broader than commonly understood. This acknowledgment has both minor and major implications. On the minor side, the term "guilt" is too narrow, because "guilt" typically refers to the determination (often of the jury) that the defendant committed the charged offenses. Here, it is used to refer to earlier determinations—whether to detain the defendant pending trial—and later determinations—whether, after guilt has been found, the defendant should receive a harsher sentence or lose supervised release. The major implications relate to updating the guilt-by-association concept to capture what, I believe, is the primary issue with statistical and similar approaches that extrapolate from other defendants.

This Article argues that a defendant has the following right, flowing from due process, judicial ethics, and our shared notion of moral responsibility:

> *The Principle of Personally Identifiable Punishment:* An individual has a right not to be subjected to state punishment[83] (and the state has a corresponding duty not to punish) based on

81. Although not referencing guilt by association, many authors have noted the increasing use, or even requirement, of algorithmic tools in pretrial detention, sentencing, and postconviction release decisions. *See, e.g.*, Garrett & Monahan, *supra* note 8, at 453 ("Some states, such as Kentucky, Ohio, Oklahoma, Pennsylvania, and Washington, have required judges to consider risk assessment during sentencing.").

82. *See State v. Loomis*, 2016 WI 68, ¶ 69, 881 N.W.2d 749, 764 ("Risk scales are able to identify groups of high-risk offenders—not a particular high-risk individual. A pointed example of potential misunderstanding arising from the use of group data is that an individual who has never committed a violent offense may nevertheless be labeled as a high risk for recidivism on the violent risk scale." (footnote omitted)).

83. "Punishment" and "punish" here primarily refer to decisions to detain or incarcerate (and to prolong detention or incarceration). Although detrimental state action stretches far beyond detention and incarceration, and although this right might have broader implications, those other actions and implications are not explored here.

> predictions extrapolated from others' data; rather, the punishment must be based solely on the individual's data.[84]

Individuals thus have a right to be punished, in essence, only on their personally identifiable information (now commonly referred to as PII in data governance spheres).[85] Correlatively, the state has a duty not to use others' data (often from other criminal defendants or arrestees) to aggravate the defendant's punishment. Because the duty is limited to algorithmic predictions, it does not limit a judge's ability to consider the sentences of other defendants (including codefendants) when sentencing a defendant who has been found guilty. Criminal sentences should not be arbitrary or widely disparate, and a judge's consultation of other sentences can limit these problems, among others. Judges thus retain the ability to increase a sentence (subject to statutory maxima and due process), but judges should not increase punishment based on a predictive algorithm.[86]

Under the Principle, the permissible data can be from the defendant (*e.g.*, fingerprints, allocution) or about the defendant (*e.g.*, eyewitness testimony, conviction record), but if the data instead comes from others (*e.g.*, those in a statistical sample), it should not be used as a basis to punish the defendant. Because this right precludes only state-imposed detriments, statistical models may still be employed to benefit

---

84. I would like to develop further the foundation and meaning of this right in a later article, but the brief articulation above should work for present purposes.

85. In *Loomis*, the Wisconsin Supreme Court cautioned lower courts regarding the use of COMPAS risk assessment scores because they are based on group data and not the particular individual. *Loomis*, 2016 WI 68, ¶ 74, 881 N.W.2d at 765.

86. *See, e.g.*, *State v. Samsa*, 2014 WI App 6, ¶ 13, 859 N.W.2d 149, 154 (rejecting the argument that courts are "prohibited from considering criminogenic needs when assessing a defendant's risk of reoffense"). If any defendants were to try to suggest that additional treatment or monitoring also amounted to increased "punishment," judges could simply and credibly respond that without those conditions, judges would have imposed additional jail or prison time. In any event, the constraint defines the prohibited punishment only as follows—increased or prolonged jail or prison time on the basis of the algorithm. *See Loomis*, 2016 WI 68, ¶ 98, 881 N.W.2d at 769 (noting that risk scores may not be used to determine sentence severity, incarceration, or as the determinative factor "in deciding whether an offender can be supervised safely and effectively in the community"); *cf.* JENNIFER K. ELEK, ROGER K. WARREN & PAMELA M. CASEY, NAT'L CTR. FOR STATE CTS., USING RISK AND NEEDS ASSESSMENT INFORMATION AT SENTENCING: OBSERVATIONS FROM TEN JURISDICTIONS 5 (2015), https://ncsc.contentdm.oclc.org/digital/collection/criminal/id/267 (concluding that risk assessment "should not be used as an aggravating or mitigating factor in determining the severity of an offender's sanction").

defendants.[87] Using data from others to help the defendant is neither morally offensive nor imposes a cognizable harm on the defendant.

This beneficial reading, moreover, is consistent with the purpose of the relatively recent adoption of algorithmic tools to thin the high numbers of detained and incarcerated criminal defendants in the United States. When these tools might be used to sustain or increase detention or incarceration, however, they are inconsistent with this Principle and with their initial purpose.[88] Moreover, by focusing on data, this Article's argument updates the concept to current technology and terminology (on, *e.g.*, big data and data privacy). It also avoids sometimes unconvincing causal theories. Finally, it guards against allowing the (mis)behavior of others to creep into the judge's decision, while it does not unrealistically restrict the judge from considering the defendant's own behavior.[89]

One counterargument to the Principle is past practice. In the absence of algorithms, judges nevertheless relied, to some unclear extent, on others' behavior to inform the judges' decisions. For example, if four out of the last five sex offenders failed to appear in the judge's courtroom, that judge might refuse to release the sixth sex offender pending trial. Likewise, if a judge's experience indicated that violent offenders tend to reoffend, the judge might have been more likely to deny pretrial release or impose a longer sentence on the defendant. Why, then,

---

87. In addition, in the private sectors, such models may impose detriments (*e.g.*, denial of a loan) when consistent with other law. Morally, however, the right perhaps should be adapted and applied to the private sector as well. I leave this argument for another day.

88. In other words, the commitment in law (and elsewhere) to individualism and agency, along with a skepticism toward these new statistically based methods, arguably combine to mean that we should use these methods only to the advantage of the individual. Thus, as noted above, algorithmic usage in this context is permissible only to the extent that it advantages—or at least does not disadvantage—the individuals subject to the decision. If this is followed, it is a Pareto improvement: The new technology lifts all tides, while the right guarantees that no individual is swept aside for the supposed greater good.

89. Thus the right is narrower than Wasserman's articulation, which covers others' conduct *and* the defendant's own past conduct. *See* David T. Wasserman, *The Morality of Statistical Proof and the Risk of Mistaken Liability*, 13 CARDOZO L. REV. 935, 942–43 (1991) (arguing "that what is objectionable is the reliance on others' conduct, *or the defendant's past conduct*, to infer his commission of a wrongful act" (emphasis added)). According to Wasserman, "[w]e object to this inference because it ignores the defendant's capacity to diverge from his associates or from his past, thereby demeaning his individuality and autonomy." *Id.* at 943. Likewise, it is "inconsistent with the law's commitment to treat the defendant as an autonomous individual, free to determine and alter his conduct at each moment." *Id.* Although not covered in this Article, epistemic and moral concerns can still exist with using the individuals' own past acts against them, including amplifying the risks of wrongful conviction. *See* Marcello Di Bello & Collin O'Neil, *Profile Evidence, Fairness, and the Risks of Mistaken Convictions*, 130 ETHICS 147, 151 (2020).

should it matter if an algorithm provides similar data to the judge? Two replies follow.

First, it was never necessarily fair for a judge to treat a defendant more harshly based simply on the misconduct of others. Just because the reasoning may have been opaque or less common did not make it fair. For example, judges rarely put it explicitly: "Mr. Doe, I am going to sentence you to thirty years in prison because I have seen or heard about other violent offenders committing additional offenses on release. You therefore would do so as well, and I am not going to give you that opportunity." Tying the defendant to the deeds or trajectory of past defendants does not treat the defendant as an autonomous individual with unique circumstances. We presumably all use causal reasoning, but judges should not take the causal chains from other defendants and simply apply them to the defendant at hand.

Second, through their datasets, these algorithmic systems aggregate and analyze the (mis)conduct of others on an unprecedented scale. The systems' proponents then sell their reliability to judges and judicial administrators in highly influential terms (*e.g.*, the model is ninety-four percent accurate in its predictions), while often being less than fully transparent in releasing the dataset, listing the variables, explaining the calculations, and so on.[90] Whereas before a judge might have permitted the behavior of a handful of past defendants in the judge's courtroom to influence how the judge would treat the current defendant, the judge now must consult a purportedly highly reliable score based on hundreds or thousands of other defendants with whom the judge has never met and about whom the judge may know nothing else. Thus, while these algorithms may not raise entirely novel problems, they raise problems and do so on a scale never before seen.

In sum, the Principle would remedy this new form of guilt by association in courts. But other approaches also exist and might profitably be applied. In addition to honoring the Principle above, judges should approach current algorithms with caution about their inputs and outputs, as illustrated below.

B. Caution About Outputs (CAO)

This Section offers judges an approach to help make just decisions when faced with a criminal justice algorithm. As Zimmermann and Lee-Stronach put it, "[n]avigating the moral risks involved in deploying predictive algorithmic systems . . . requires moral deliberation and

90. *See* Recent Case, State v. Loomis, *881 N.W.2d 749 (Wis. 2016)*, 130 HARV. L. REV. 1530, 1535–36 (2017); *see also* Garrett & Monahan, *supra* note 8, at 458.

doxastic restraint uniquely available to human agents."[91] In light of the algorithmic risks, including the risk of perpetuating or even amplifying racial prejudice and structural injustice, they suggest using the following approach, which in turn this Article applies to judges:[92]

> *Caution About Outputs (CAO)*: Instead of restricting inputs or relying on insufficiently informative inputs, human agents relying on algorithmic procedures should (i) aim to work with inputs that are maximally informative with respect to the impact of structural injustice, and (ii) exercise caution when basing decisions on algorithmic outputs. This may, depending on the context-sensitive features, entail suspending belief and remaining agnostic about a particular issue (e.g. a question like "does individual A merit X?"), or it may entail not taking any action for the moment.[93]

Judges of course are essential to judicial decision-making, even in our increasingly algorithmic criminal justice system.[94] Zimmermann and Lee-Stronach provide illustrative questions for a decision-maker, *A*, applying the Principle of Caution About Outputs in the context of a financing decision for applicants *B* and *C*, who received the same algorithmic score.[95] These questions are adaptable to the judge, as discussed below. Although the questions are certainly not exhaustive, they may well assist the judge in approaching the algorithmically informed information with appropriate caution and sensitivity.

> *1. Has [the judge] considered the full range of available options (e.g., denial, approval, different loan rates or conditions, social services)?* [96]

One current advantage of a human decision-maker, such as the judge, is the ability to "address counterfactual scenarios that humans are

---

91. Zimmermann & Lee-Stronach, *supra* note 75, at 21–23 (noting that there may be "strong moral reasons to ascribe similarity to a broader range of cases than we currently do . . . if doing so would get us closer to a plausible ideal of substantive justice").

92. In most algorithm-involved decisions (at least at the time of this writing), humans are still a part of the process. *See id.* at 21 ("We view human deliberation as an appropriate object of procedural justice constraints.").

93. *Id.* at 20–21.

94. *See, e.g.*, *Malenchik v. State*, 928 N.E.2d 564, 566 (Ind. 2010) (concluding that assessment instruments do not replace a trial court's sentencing determination).

95. Zimmerman & Lee-Stronach, *supra* note 75, at 21–22.

96. *Id.* at 22.

capable of recognising, but which are not reflected in the system's data."[97] The judge can thus search for an intervention appropriate to a defendant beyond the algorithmic score or instrument.[98] The fairest outcome might be unknown to, or otherwise beyond, the instrument. Judges may fill in the gaps.

Indeed, the gaps may be wide. Algorithms often are not designed to make specific recommendations or decisions; those instead are left to the judge (or probation officer or parole board). The algorithm might simply make a prediction about the relevant behavior (*e.g.*, likelihood to reoffend).

Although perhaps some judges have assumed that an algorithm's prediction means or implies more than this, the ultimate decision is still the judge's. Instead of jail or prison time, for example, the judge could impose a treatment or monitoring program. Algorithms (to date) are not designed to insist on incarceration; a judge who assumes that the bare prediction should be dispositive of detention or incarceration seems both to misunderstand the extent of the algorithmic tool (and therefore lacks competence) and to underestimate the judge's own responsibility for the ultimate decision. Indeed, the judge may even overstep current law if the algorithm is supposed to remain only one factor in the judge's determination.

> *2. How much is at stake for the decision subject (i.e., loan applicants B and C)? Will an adverse decision outcome impact B and C in the same way, or will structural injustice lead to vastly different effects for B and C?* [99]

The judge can recognize that the costs of detention (whether jail or prison) might impact defendant *B* more harshly than another defendant, here *C*, and the judge should interrogate the fairness of the algorithm, given its known limitations.[100] But let us address preliminarily a

---

97. *Id.*

98. This in turn highlights the need for states to provide the judge with an array of practical sentencing options, which of course involves funding and other complications.

99. Zimmermann & Lee-Stronach, *supra* note 75, at 22.

100. For example, at least in Wisconsin under *Loomis*, every sentencing judge should consider the following (among other considerations):

- The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined.

complication and a proposed solution: Judges do not generally compare the sentencing fates of defendants *B* and *C* unless they are codefendants or otherwise similarly joined.[101] But *C* in particular seems unnecessary to the analysis. Furthermore, it might be morally and logically infirm to base a more lenient sentence for *B* on the fact that *C* has enjoyed better circumstances or to base a harsher sentence for *C* on the fact that *B* has faced worse circumstances (unless, again, *B* and *C* are connected in some meaningful way).[102] In other words, sentencing for *B* and *C* is not interdependent. Finally, hundreds or thousands of *C's* were analyzed in the algorithmic result, but the judge typically receives and lacks access to any other information about the *C's*. Thus, relying on comparisons with some particular *C* or a group of unspecified *C's* generally seems inappropriate and contrary to showing caution.

The way out, however, seems fairly easy: The judge seemingly could arrive at a substantively just disposition by looking at *B's* circumstances, without reference to *C's* circumstances and sentence (if any).[103] The judge could consider, for example, *B's* low education level

---

- Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders—not a particular high-risk individual.
- Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism.
- A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed. Risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations.
- COMPAS was not developed for use at sentencing, but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole.

State v. Loomis, 2016 WI 68, ¶ 100, 881 N.W.2d 749, 769–70. *But see* Recent Case, *supra* note 90, at 1537 (arguing that this advisory does not sufficiently explain to judges how they would go about scrutinizing algorithmic tools for bias and that instead courts should exclude secretively built or insufficiently validated tools).

101. To be sure, occasionally judges will compare the sentencing outcomes for a class of defendants, *e.g.*, those convicted of bank robbery in the jurisdiction. This exercise is not dispositive, as every case is different (*e.g.*, each defendant presents a unique and abundant arrangement of aggravating and mitigating factors for the sentence). Furthermore, a general class comparison is not necessarily the same enterprise as comparing unrelated individual defendants.

102. The judge presumably would consider some comparisons (*e.g.*, to sentence lengths for more, less, or equally serious crimes), but nothing otherwise seems to connect *C* in this question.

103. Again, though, if we are considering codefendants or the functional equivalent, there is a doctrine and an urge toward intra-case proportionality. In other words, a defendant who acted less egregiously should not receive a harsher sentence than

and *B's* likelihood to be arrested due to his neighborhood or biased policing, and these factors could favor a reduced or completely non-carceral sentence. Defendant *C* or some general class of *C's* should not negatively impact Defendant *B*. The judge may, however, reference common sentences for the offense at issue to gather a sense of a baseline and to avoid arbitrary sentences.

In sum, the judge can recognize that the defendant has faced structural injustice and that the impacts for the defendant if incarcerated are particularly grave. Of course, to recognize more clearly structural injustice and disproportionate harms potentially resulting from a carceral sentence, the judge will at least have to consider comparatively the circumstances of others. As a brief reminder, however, this does not violate the Principle of Personally Identifiable Punishment because the judge would not be *increasing* the defendant's punishment on the basis of others' conduct.

> *3. How much is at stake for [the judge]: might [the judge] have morally relevant biases which impair their ability to know what justice would require in this context? How much is [the judge]'s judgment about B being actually similar to C affected by higher-order uncertainty? Can [the judge] reasonably adopt any belief about B, or is [the judge] morally required to suspend judgment for the time being?* [104]

Modern judicial education has made judges increasingly aware of their biases (at least generally).[105] Although the evidence is mixed as to whether this education has actually dampened these biases,[106] judges of course still get to decide cases. They should proceed cautiously, knowing that they (like other humans) have implicit and cognitive biases and that they often do not share the same circumstances as the defendants before them (*e.g.*, judges tend to be richer and more homogenous than defendants).[107] The judge should interrogate these asymmetries in every case.

---

a defendant who acted more egregiously during the same crime or scheme. Inter-case, however, the criminal justice system tolerates large differences in sentences even for seemingly similar offenses and circumstances. This is often explained away, sometimes dismissively, as "every case is different."

104. Zimmermann & Lee-Stronach, *supra* note 75, at 22.

105. *See, e.g.*, Michele Benedetto Neitz, *Socioeconomic Bias in the Judiciary*, 61 CLEV. ST. L. REV. 137, 163–64 (2013).

106. *See id.* at 164–65.

107. *See, e.g.*, *id.* at 141 (noting "the existence of implicit socioeconomic bias on the part of judges through examination of recent Fourth Amendment and child custody cases").

Fortunately, modern sentencing judges typically have a range of information available about the defendant (but typically with much more available postconviction than pretrial).[108] Nevertheless, if the judge "is unsure about the extent to which structural injustice has disadvantaged *B* in comparison to *C*, [the judge c]ould suspend—at least temporarily—belief about *B*, and adopt a *blanket rule* that gives *B* the benefit of the doubt."[109] This rule "counteracts historically unjust negative assumptions about 'people like *B* defaulting on loans'"—or here, assumptions about people like *B* having a propensity to commit crime in general or certain types of crime.[110] In the United States, for example, the judge might follow a default rule that gives Black (or another category of) defendants the benefit of the doubt in cases of uncertainty. A default rule based on race, however, might run afoul of the Equal Protection Clause (particularly given the corollary: presumably defendants of other races would be sentenced comparatively more harshly than Black defendants under this default rule). This conceivably could also violate the judge's oath not to discriminate on the basis of race.

The equal protection complication is not fatal, however. Aside from adopting a "blanket rule" based on race, equal protection generally would not block a judge from following CAO. Nothing stops a judge from examining critically the judge's own potential biases or suspending judgment temporarily to gather more information, both of which are of course fully consistent with the CAO approach. The judge could order an updated presentencing investigation or report as needed to gather more information about the defendant or the circumstances. The circumstances, which might be correlated with race, could warrant sentence mitigation, and it likely would not violate equal protection to rely on those circumstances (instead of race).

* * *

The CAO approach is an example of how a judge might approach algorithms skeptically, particularly in light of historic and lingering racial injustices in the criminal justice system. These injustices impacted the data on which current algorithms rely. Although there is nothing magical about the CAO approach in particular, following CAO or something like it would help the judge to avoid repeating these injustices and to mitigate the algorithmic formula of reliance on other defendants' conduct. Beyond individual judges, the judiciary collectively can address racially harmful algorithms, as discussed below.

---

108. *See, e.g.*, ELEK, WARREN & CASEY, *supra* note 86, at A-12, A-107.

109. Zimmermann & Lee-Stronach, *supra* note 75, at 22.

110. *Id.* (emphasis omitted).

C. Most Favorable Demographic Parity in Criminal Justice

Although algorithms typically do not use race explicitly, many use proxy variables for race in the United States (*e.g.*, neighborhoods, associates, arrest records).[111] If the deployed algorithm relies on race, it might have to meet strict scrutiny.[112] Given the imperfect fit with existing constitutional law, however, most algorithms have not received strict scrutiny (or in some instances, any significant judicial scrutiny).[113] In this void, this Article argues for the following approach, which rests on the concept of demographic or statistical parity.[114]

Of the dozens of technical constraints that have been proposed to address bias in algorithms,[115] demographic parity is often dismissed in

---

111. Note, *Beyond Intent: Establishing Discriminatory Purpose in Algorithmic Risk Assessment*, 134 HARV. L. REV. 1760, 1766–67 (2021).

112. In the criminal justice area, Supreme Court precedent generally suggests that a showing of discriminatory purpose, not merely discriminatory impact, is required, and as many authors have noted, this is an imperfect fit in the algorithmic context. *See, e.g.*, *id.* at 1764–76 ("The focus of the discriminatory purpose requirement on intent is inapposite to algorithmic systems. Algorithms do not possess intent. Disparate impact rather stems from inevitable data biases or the intentional choices of their creators or the judges applying them." (footnote omitted)); *Henderson v. Stensberg*, No. 18-CV-555, 2021 WL 1221249, at *5–6, *8 (W.D. Wis. Mar. 26, 2021) (denying defendant's equal protection claim against COMPAS primarily because the defendant was unharmed—he received a 1, the lowest possible COMPAS risk score—but also noting that the defendant had essentially failed to "produce evidence of discriminatory intent"). Although the area is unsettled and arguments could be (and have been) made under current Equal Protection doctrine and under reasonable extensions of it that criminal justice algorithms violate the law, a normative argument is sufficient for purposes of this Article.

113. *See* Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 851–52 (2020); Huq, *supra* note 8, at 1087–88.

114. Kearns and Roth define and illustrate statistical parity as follows:

> Suppose for some reason we are concerned about discrimination against Squares in the granting of loans by a lender, so we ask that race be a protected attribute. Statistical parity simply asks that the fraction of Square applicants that are granted loans be approximately the same as the fraction of Circle applicants that are granted loans. That's all. The definition doesn't specify how many loans we have to give, or which particular Circle and Square citizens should receive them—it's just a crude constraint saying that the rate of granted loans has to be roughly the same for both races. Note that while our concern might have been discrimination against Squares, the definition is two-sided and thus also demands we not discriminate against Circles (though we could define a one-sided variant if we wanted).

KEARNS & ROTH, *supra* note 20, at 69.

115. *See, e.g.*, Arvind Narayanan, *21 Fairness Definitions and Their Politics*, YOUTUBE (March 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk.

the literature.[116] As discussed below, however, a certain type of racial parity—what I will call "Most Favorable Demographic Parity"—appears to be the best choice for an algorithmic constraint in criminal justice. Therefore, the judiciary, as a state actor and as the contracting or otherwise responsible party for using an algorithm with defendants,[117] should insist on Most Favorable Demographic Parity before the algorithm is deployed in its courts.

1. Most Favorable Demographic Parity

To reach most favorable demographic parity, the algorithm must meet two elements. First, it must achieve demographic parity (or come very close), *i.e.*, an equal proportional percentage by race in its outputs. Second, and equally importantly, the parity is set at the most positive or least negative percentage. For example, if twenty percent of the best-scoring racial class would receive a benefit or avoid a detriment under the algorithmic tool, then twenty percent of the other classes would receive the same.

As currently implemented, criminal justice algorithms tend, in effect, to disadvantage some races and to privilege other races.[118] For example, compared to other races, an algorithm might classify more Black defendants as highly likely to skip trial or reoffend.[119] Presumably the algorithm or its developers did not explicitly use race to reach this

116. For example, Kearns and Roth call it "generally . . . weak and flawed," in part because "while statistical parity is not at odds with exploration, it is at odds with *exploitation*—that is, with making optimal decisions—any time the optimal thing to do from an accuracy perspective differs between the two populations." KEARNS & ROTH, *supra* note 20, at 71; *see also* Sofie Goethals, David Martens & Toon Calders, *PreCoF: Counterfactual Explanations for Fairness*, MACH. LEARNING (Mar 28, 2023), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10047477/pdf/10994_2023_Article_6319.pdf [https://perma.cc/C9LQ-UXU5] (discussing weaknesses in demographic parity and other competing fairness concepts).

117. When employed by state actors, racial classification disparities generally receive additional scrutiny. If classification discrimination exists on the basis of race in particular, it might be a violation of equal protection. Private actors, such as Northpointe or Equivant, are not directly bound by the Constitution—which in the main protects the public against government actors—but once the government uses the products or services on the public, those products or services must typically comply with constitutional requirements. *See, e.g.*, *Henderson v. Stensberg*, No. 18-CV-555, 2020 WL 1320820, at *1–2 (W.D. Wis. Mar. 20, 2020) (refusing to grant Northpointe's motion to dismiss in part because "private actors act under color of law when they work jointly with state actors to violate a person's rights"). To be sure, most algorithms do not use race directly, and constitutional law so far has not been applied. Judges, however, who take an oath not to discriminate, will generally be concerned with racially discriminatory outputs, even if the algorithm's purpose is not racial discrimination.

118. *See* Huq, *supra* note 8, at 1076–82.

119. *See id.* at 1080.

disparate result, but the result is discriminatory nonetheless and warrants attention. Although other algorithmic constraints have been proposed, demographic parity is the only constraint that actually guarantees members of different races "equal protection of the laws."[120] The other approaches, in contrast, tend to mitigate algorithmically created racial disparities in the outcome or score, and some even potentially aggravate those disparities, but none equalize the disparities under the law.[121] The proposal of demographic parity, adjusted up or down to the most beneficial (or least detrimental) outcome by race, necessarily ensures equal treatment and addresses the preexisting injustices in the system. Demographic parity also avoids resorting to scrutiny of the decision-maker's mental states (whether of a programmer, administrator, or judge), which do not map on well to the algorithmic context.[122]

To be sure, this approach likely will mean that more individuals receive the benefit or avoid the detriment than without this constraint. This can be costly in criminal justice,[123] but this is the corresponding

120. U.S. CONST. amend. XIV, § 1. Race is one of, and perhaps the, most strongly protected class in equal protection law, and of course the one on which this argument is based. Even if the law has not yet fully reached algorithms' connection to race, the law informs this section. Other classes, perhaps most notably sex, have been recognized under constitutional law, although this section and constraint deal only with race in part because the other classes have generally received less protection in this context. *See generally State v. Loomis*, 2016 WI 68, ¶ 83, 881 N.W.2d 749, 766 ("Thus, if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose." (citing Melissa Hamilton, *Risk–Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 255 (2015))).

121. *See* Huq, *supra* note 8, at 1113–32.

122. Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, 81 PROC. MACH. LEARNING RSCH. 149, 152 (2018) (noting that "mental state accounts of discrimination do not naturally transfer to the context of algorithmic decision-making"). We might also demand something akin to the demographic parity proposal for judicial decisions generally, even ones not involving algorithmic tools, but this is beyond the scope of the present Article.

123. As Sandra G. Mayson recently summarized:

> In a recent study, . . . Richard Berk adjusted juvenile justice data to produce statistical parity in predictions of rearrest for violent crime. This resulted in extremely high rates of false negatives; the algorithm missed 92% of violent-crime rearrests of white juveniles and 98% of violent-crime rearrests of black juveniles. Such high false-negative rates are a hefty cost in accuracy in and of themselves. But the increase in error also had a disparate racial impact. Because the base rate of arrest for violence in this data was significantly higher among black juveniles than white juveniles, and because most violent crime is intraracial, the astronomical false-negative rates would mean a much greater *absolute* number of violent-crime arrests missed in the black community than in the white community. To the extent that violent-crime arrests correspond to violent crimes, the effort to achieve statistical parity for

equal protection and fairness costs of using the racially harmful algorithmic tool. Choosing any lower cutoff, even if applied proportionately across the classes, would negatively impact at least one of the groups on the basis of race. If that eventuates, then a member of the aggrieved class could rightfully object.[124] Under this proposal, however, no one is harmed, while many are helped.

Another key concern with this approach is intersectionality. The algorithm audit would presumably need to account for the intersections of the protected racial classes, and the designer might choose or ultimately be required to factor in additional classes (*e.g.*, those with disabilities or those over a certain age). As the classes considered and their intersectional possibilities grow, however, the more the algorithm will be attending to demographic parity rather than accuracy or other goal. This of course is a valid concern, but the proposal reduces its force through two mechanisms: (1) limiting its recommendation to race (one of the few areas informed by strict scrutiny and in a criminal justice system with long-standing racial disparities in inputs and outputs), which reduces significantly the puzzle of fully accounting for intersectionality, and (2) in part to account for racial intersectionality, a Four-Fifths Rule is proposed in the next Section. In brief, the Rule provides some leeway to adjust for intersectionality and other concerns during the implementation of demographic parity.

Of the two responses above, the first is significantly more responsive to the intersectionality objection: Policing percentages across only the racial factor is realistic, whereas policing percentages involving multiple factors presents a more difficult and perhaps unworkable question. The next Section explains the Four-Fifths Rule, adapted from the employment discrimination context.

2. Most Favorable Demographic Parity Compromised:

> black juveniles subject to risk assessment comes at the cost of disparate harm to black victims.

Sandra G. Mayson, *Bias in, Bias out*, 128 YALE L.J. 2218, 2250 (2019) (footnotes omitted) (discussing Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEGAL STUDS. 175 (2019)). Putting aside the point that no algorithm to my knowledge has yet to be highly accurate at predicting violent crime, the argument, while generally weighty, omits the most obvious legal interest at stake—namely, the individual defendant who (absent controlling for parity) would be subject to state punishment based at least in part on his race.

124. *Cf. Henderson v. Stensberg*, No. 18-CV-555, 2021 WL 1221249, at *1 (W.D. Wis. Mar. 26, 2021) ("Henderson's recidivism score in the COMPAS assessment was favorable; he has failed to adduce admissible evidence that he was harmed by his COMPAS assessment or that he was denied parole for a discriminatory reason.").

The Four-Fifths Rule

Most Favorable Demographic Parity may be paired with the Four-Fifths Rule, as explained below. Although not without its drawbacks, this addition effectively addresses a number of potential objections to demographic parity.

Although the Four-Fifths Rule has been passingly mentioned in the algorithmic context,[125] it has been put to little or no use to date. In the employment context, the Four-Fifths Rule flags presumptively discriminatory employment selection practices with disparate impact on the basis of a protected characteristic.[126] As a simple example, a large employer uses a screening or selection mechanism that selects fifty percent of applicants listed as Black and ninety percent of applicants listed as Asian. Because the former's selection rate is under four-fifths of the highest rate (ninety percent), this will generally be regarded as adverse impact. Pairing Most Favorable Demographic Parity with the Four-Fifths Rule will provide two benefits: one data-related and one audit-related, as illustrated below.

A brief example reveals the benefits: Drawing from data in a local five-thousand-person sample, a vendor crafts a two-point risk assessment tool in a region with only two races. It would be shocking if, without adjustment, the proportion of each race receiving each score was exactly the same (*e.g.*, score 1: 25%, 25%; and score 2: 75%, 75%).[127] Real data are rarely that perfect; some amount of variation should be expected even without injustice having infiltrated the data or model. The four-fifths threshold leaves plenty of room for natural variation (perhaps too

125. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 701–02 (2016); Mayson, *supra* note 123, at 2242 ("Statistical parity requires that the same percentage of each group be forecast for arrest. That is, it requires parity in the total-population impact of the prediction at issue. This is the simplest measure of intergroup equality. It is also the one that dominates disparate impact law. EEOC guidance, for example, provides that too great a divergence from statistical parity is prima facie evidence of 'adverse impact.'"); Huq, *supra* note 8, at 1119–20.

126. The Uniform Guidelines on Employment Selection Procedures states:

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

29 C.F.R. § 1607.4(D) (2015).

127. This is true even assuming some theorem (*e.g.*, the law of large numbers) will bring the proportions close(r) together.

much room, as soon discussed below).[128] In addition, a four-fifths or other threshold serves an auditing purpose, as it analogously does in the employment context.[129] In other words, if members of a certain race are faring below the threshold, it raises a presumption against the practice (here, the deployment of or reliance on the algorithm). To be sure, this is the weaker reason of the two, because any variation from perfect demographic parity could in theory raise a similar presumption, however unrealistic.

A final objection might be that implementation of this regime could face a constitutional challenge. For example, a member of the best-treated class, from which the threshold is set, might complain that the member is being treated less favorably than the other classes. Technically, of course, this is false—the member is being treated the same and (given the four-fifths leeway) potentially better than the other classes. Moreover, the Rule, as framed, runs both ways: When the member belongs to a racial class that falls lower on the model, the member will benefit from the Rule. In other words, this is not a rule that favors only members of a certain race (although perhaps such a rule could survive strict scrutiny given in part the historical inequities perpetuated in the data and the discrimination is designed to help, not hurt, the

128. *Cf.* § 1607.4(D) ("[A] greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. . . . Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant, or where special recruiting or other programs cause the pool of minority or female candidates to be atypical of the normal pool of applicants from that group.").

129. *See* Barocas & Selbst, *supra* note 125, at 701–02.

disadvantaged racial group).[130] Thus, the member will fail to point to a harm,[131] and this challenge will presumably fail.

In sum, Most Favorable Demographic Parity, coupled with the Fourth-Fifths Rule, seems both workable and compelling. The judiciary should demonstrate its commitment to equal protection and judicial ethics by ensuring that algorithmic tools align, at least roughly, with these goals. If the tools cannot be made to roughly equalize their outputs according to race—either technically or constitutionally—then the judiciary in response should either not use the tools or treat those impacted by them leniently.

3. Most Favorable Demographic Parity in Action

Even with the four-fifths rule, operationalizing Most Favorable Demographic Parity will not be easy, although several different approaches might suffice and thereby grant flexibility to the designers. By treating the demographic parity principle as a post-hoc auditing tool, moreover, discretion and creativity as to the model and its input will be left to the designers. The complexity and the potential feasibility can be illustrated by the below example.

To predict recidivism, a vendor creates a model with a 3-point scoring output: 1 (low risk: 0–25% likelihood of recidivism); 2 (medium risk: 25.1–50% likelihood of recidivism); or 3 (high risk: >50% likelihood of recidivism). The jurisdiction has four races: Blues, Reds, Greens, and Purples. The vendor presumably optimized the model for

130. *Cf., e.g.*, *Grutter v. Bollinger*, 539 U.S. 306, 328 (2003) (referring to diversity as a compelling interest); Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 809 (2020) ("[C]onclud[ing] that at least some forms of algorithmic affirmative action, to the extent they are racial classifications at all, ought to survive strict scrutiny as narrowly tailored solutions designed to mitigate the effects of past discrimination."); Cheryl I. Harris, *Whiteness as Property*, 106 HARV. L. REV. 1707, 1783 (1993) ("If affirmative action is viewed through the prism of distributive justice, the claim of white innocence no longer seems so compelling, because a distributive justice framework does not focus primarily on guilt and innocence, but rather on entitlement and fairness. Thus, distributive justice as a matter of equal protection requires that individuals receive that share of the benefits they would have secured in the absence of racism."). *But cf. Students for Fair Admissions, Inc. v. President of Harvard Coll.*, 143 S. Ct. 2141, 2175 (2023) (concluding that two race-conscious college admissions programs violate equal protection because they "lack sufficiently focused and measurable objectives warranting the use of race, unavoidably employ race in a negative manner, involve racial stereotyping, and lack meaningful end points").

131. *See, e.g.*, Hellman, *supra* note 113, at 863–64 ("[W]here an algorithm designer is aware that an approach will have a racially disparate impact in the abstract and so makes changes to avoid that impact, we have no specific, known people who are harmed, nor any reliance.").

accuracy, and it produced the following results (which for purposes of this example were later cross-checked with the race of the defendants):

Simulation: Racial Outputs of 3-Point Risk Model [132]

Defendant's Race	1 (low risk)	2	3 (high risk)
Blue	**80**	10	**10**
Red	60	10	30
Green	75	**5**	20
Purple	70	15	15

In this model, the most beneficial (least harmful) category is score 1. That score tells the judge or other decision-maker that the defendant is a low risk (and thus suggests or requires that the decision-maker refrain from detaining or sentencing the defendant more harshly). The Blue Race fared the best (80% of whom received a 1). Where possible, the judiciary should insist that the vendor attempt to produce a model in which all racial groups land near 80% in category 1. At a minimum, applying the four-fifths rule, no race should fall below 64% (four-fifths of 80%). As currently constructed, the treatment of the Red Race (only 60% of whom received a low-risk score) should raise concern. The judiciary might reject the model on this basis.

If, however, category 1 complies with the threshold, the two additional scores (2 and 3) should then be scrutinized for equality (or at least for four-fifths of it) to the best-scoring group.[133] Once the process has been completed, the model or adjusted model could be a candidate for the judiciary's purchase or deployment. When initially deploying the model, however, the judiciary should effectively roll it out as a pilot.[134] In particular, the model should have no negative effect on a defendant until the model's scores on actual defendants in the jurisdiction have been audited to ensure compliance with the principles above. Finally, at this point, the judiciary can safely rely on the model in practice (subject to further auditing and the other concerns in and beyond this Article).

One immediate objection to the proposal above rests on accuracy: The proposal will likely reduce accuracy to some extent whenever it

132. Numerical values represent the racial percentage in each of the three score columns; the most favorable percentage in each column is boxed.

133. The vendor presumably could have worked the other direction (*e.g.*, starting with the proportion of each race with a 3). This is just an example, not meant to be a set-in-stone algorithm on how to certify the model for most favorable demographic parity.

134. Both courts and data scientists often roll out new models (construed broadly) as temporary pilots to analyze the performance and impacts, before wide or permanent implementation.

prompts readjustments in the algorithmic model. But less short-term accuracy proportionately results in an increase in fairness, for at least two reasons. First, it seems conspicuously dubious that one race—a broad, social construct—would be more prone to criminal activity than another, absent some widespread, negative differences in circumstances.[135] Yet the model in fact seems to be suggesting this. Instead of following the model to the group member's detriment, the judge should dig in to reach a determination based on the member's specific circumstances. In any event, it is individually unfair to impose a detriment simply by some racial association with others. Second, and relatedly, the justice system historically imposed, and still imposes, different outcomes depending on the defendant's race.[136] Beyond the constitutional law point—in other words, even if constitutional law would permit the racial classification disparities—courts and the legal community should want to improve the historically underperforming criminal justice system. We thus encounter a value conflict between fairness and predictive accuracy. It may well mean that an algorithmic system free (or freer) from racial disparities (however predictive) is less accurate—until we remedy the underlying structural unfairness accounting for the predictive value of race. It is at least reasonable to insist in the meantime that the state accord greater weight to fairness over accuracy. Furthermore, this would be anticipating a brighter future, that is, grooming the model to predict in a world without widespread racial injustice.

Another objection is technical. At least at present, the model likely cannot take account of race, unless race is inputted. This injects race explicitly into the model, whereas currently the models do not rely on race but may produce racially disparate outputs.[137] This complication might not be fatal given this Article's "most favorable" constraint; race might be used to equalize most favorable treatment without hurting any

---

135. Huq maintains:

> Statistical parity is a clear and simple idea. . . . On the other hand, *there is no a priori reason why state coercion should be equally distributed among racial groups*. To be sure, there is some evidence that at least for certain sorts of offenses, such as narcotics crimes, there are 'no statistically significant differences' in offending rates for different racial and ethnic groups. But on the assumption that the algorithm's training data are not flawed, the hypothetical would simply not capture such cases.

Huq, *supra* note 8, at 1119–20 (emphasis added) (footnotes omitted). A just system, however, should start with the converse proposition: "there is no a priori reason why state coercion should be [*un*]equally distributed among racial groups." *Id.* In addition, given the broad, social construct of race, an algorithm heavily impacting race in its outputs might be too course-grained in its use of proxies.

136. *Id.* at 1104–07.

137. Binns, *supra* note 122, at 155.

race. Perhaps the more difficult technical objection is that this model might be unworkable; in other words, the model could not be configured to comply perfectly with the Principle. This might well be true, but the four-fifths leeway gives designers discretion to design and adjust a model that results in compliant outcomes (however implemented in the model). Moreover, the Principle functions as a post-hoc audit tool for the judiciary. Whatever model is proposed to the judiciary, it could be scrutinized against the Principle, and if noncompliant, the judiciary could reject the model altogether, request adjustments, or discount the outputs.

Finally, the Four-Fifths Rule permits significant discretion to account for accuracy (and data issues, such as initially small sample sizes) in the model and any readjustments. Indeed, some might object that the Four-Fifths Rule is too solicitous of accuracy or too gameable to guard against racial discrimination. This is a valid concern, and furthermore, the particular four-fifths (eighty percent) threshold is of course arbitrary. A state or its judiciary could instead choose, for example, ninety-five percent at the outset or could increase the eighty percent threshold over time as the model is audited and improved.

* * *

Until the law evolves favorably, the above constraints rest primarily on moral arguments, not legal ones.[138] These arguments are primarily lodged with the judiciary, not necessarily individual judges. It is the judiciary (including its administrative offices) in the applicable jurisdiction that procures, or at a minimum deploys, the criminal justice algorithm in its courts. The judiciary should insist that the tool on which its judges are relying satisfies demographic parity; it should insist this of the tool at the time of its implementation and in later audits. The Four-Fifths Rule will provide some leeway in the design and in future audits, but if the judiciary fails to enforce this proposal (or something like it), its judges will necessarily be contributing to the historic and continuing racial injustices in our courts. Most judges want, or at least profess to want, the opposite—to remedy these injustices.[139] Insisting on Most

138. *Cf., e.g.*, Garrett & Monahan, *supra* note 8, at 492 ("The use of risk assessment in criminal justice should be regulated with rules to inform and structure decision-making, and a process for developing those rules should be articulated. The Due Process Clause and other constitutional criminal procedure sources do not provide sufficiently definitive or informative guidance in this new world where risk is judged at each stage of the criminal process.").

139. Indeed, judges might be cross-checking or following algorithmic scores precisely to avoid unfairness, in the sense of avoiding or mitigating their implicit or cognitive biases. Although these are important intentions, they will not lead to substantive

Favorable Demographic Parity would be one tangible and presumably obtainable step in that direction.

D. Summation: Intersection or Confliction

The three proposed innovations helpfully buttress one another, for the most part. For the individual judge, using the principle of Personally Identifiable Punishment would ensure that defendants are not punished, only aided, by the guilt-by-association design of the statistical models. But this principle is not law or necessarily even practice at present. Moreover, even when a conscientious judge employs the principle, the statistical outputs might be too narrow or skewed. Therefore, another principle, Caution About Outputs (CAO), would ensure that judges scrutinize these statistical outputs, seek more fine-grained information, and consider individualized or novel resolutions outside the statistical model's scope. Lastly, Most Favorable Demographic Parity is aimed at the judiciary (and in turn the algorithmic developers and vendors). It seeks to limit the impact of racial bias soaked into the statistical data and reiterated in the models. It is both broader and narrower than the principles of personally identifiable punishment and CAO: broader in that it covers all judicial models and would make no provision for case-by-case determination; narrower in that it covers only racial bias (although other biases, such as those based on gender, might be mitigated with a similar approach).

To be sure, none of the three approaches completely eliminates the guilt-by-association inherent in the current statistical models. Indeed, it is not clear whether that would be possible without eliminating the usage altogether. Thus, until models exist that do not rely on a guilt-by-association technique, policymakers have only the following choices: refrain from using the models; fully rely on the models; or partially rely on the models. The criminal justice system already operates in the world of partial reliance, and the three approaches developed above would likely improve the fairness of that reliance.

CONCLUSION

Turning to statistics and technology, policymakers and judges strive for judgments less beholden to injustices and biases. In the courts, however, the efforts to date have not clearly produced less harm in terms of injustices and biases. As shown above, judges have ethical duties to

---

or procedural fairness if the relied-on tool rests on or replicates inappropriate bias in its data or in its model (*e.g.*, if the tool follows or even aggravates racially prejudiced law enforcement practices, as some do).

judge fairly, impartially, and competently, yet judges currently base their decisions, in part, on guilt-by-association and racially harmful algorithmic tools. Perhaps contrary to the opinions of certain commentators, however, the algorithmic project does hold promise. It could maximize benefits, reduce human biases, and tailor treatments.[140]

To seize on the positives and mitigate the negatives, constraints along the lines proposed above should be considered and implemented. On the regulatory level, judicial authorities should clarify and amplify the judicial ethics rules and opinions in this novel, question-provoking domain. But improvements need not wait until judicial ethics authorities catch up. Judges right now have the power to implement beneficial constraints at the decisional level, while the judiciary as a whole has the power to require fairness improvements in the algorithmic models. In conclusion, by basing punishment on the defendant's data (not those in the sample), by exercising caution and diligence toward the algorithmic inputs and outputs (not deferring to the easy numerical scores or red flags), and by requiring Most Favorable Demographic Parity in the algorithmic models (not allowing structural demographic disparities to persist), the judges and the judiciary will boost algorithmic judicial ethics.

140. *See, e.g.*, *State v. Loomis*, 2016 WI 68, ¶ 88, 881 N.W.2d 749, 767 ("Although it cannot be determinative, a sentencing court may use a COMPAS risk assessment as a relevant factor for such matters as: (1) diverting low-risk prison-bound offenders to a non-prison alternative; (2) assessing whether an offender can be supervised safely and effectively in the community; and (3) imposing terms and conditions of probation, supervision, and responses to violations." (footnote omitted)).