University of Cincinnati Law Review

Volume 88 | Issue 4 Article 3

May 2020

Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality: Algorithms are Color Blind and Never Tire

Mirko Bagaric
mbagaric@swin.edu.au

Dan Hunter
Queensland University of Technology, Australia, dan.hunter@qut.edu.au

Nigel Stobbs
Queensland University of Technology, n2.stobbs@qut.edu.au

ERASING THE BIAS AGAINST USING ARTIFICIAL INTELLIGENCE TO PREDICT FUTURE CRIMINALITY: ALGORITHMS ARE COLOR BLIND AND NEVER TIRE

Mirko Bagaric, Dan Hunter,** and Dr. Nigel Stobbs****

ABSTRACT

Many problems in the criminal justice system would be solved if we could accurately determine which offenders would commit offenses in the future. The likelihood that a person will commit a crime in the future is the single most important consideration that influences sentencing outcomes. It is relevant to the objectives of community protection, specific deterrence, and rehabilitation. The risk of future offending is also a cardinal consideration in bail and probation decisions. Empirical evidence establishes that judges are poor predictors of future offending—their decisions are barely more accurate than the toss of a coin. This undermines the efficacy and integrity of the criminal justice system.

Modern artificial intelligence systems are much more accurate in determining if a defendant will commit future crimes. Yet, the move towards using artificial intelligence in the criminal justice system is slowing because of increasing concerns regarding the lack of transparency of algorithms and claims that the algorithms are imbedded with biased and racist sentiments. Criticisms have also been leveled at the reliability of algorithmic determinations. In this Article, we undertake an examination of the desirability of using algorithms to predict future offending and in the process analyze the innate resistance that human have towards deferring decisions of this nature to computers. It emerges that most people have an irrational distrust of computer decision-making. This phenomenon is termed "algorithmic aversion." We provide a number of recommendations regarding the steps that are necessary to surmount algorithmic aversion and lay the groundwork for the development of fairer and more efficient sentencing, bail, and probation systems.

INTRODUCTION

The effectiveness of the sentencing system depends considerably on the accuracy of decisions regarding whether an offender will commit further offenses. This consideration is paramount in ascertaining how the

* Dean of Law, Swinburne University, Australia.

** Executive Dean, Faculty of Law, Queensland University of Technology, Australia.

*** Senior Lecturer, Queensland University of Technology, Australia.

core sentencing objectives of community protection, specific deterrence and rehabilitation should be calibrated in deciding the ultimate sanction.[1] If an offender has a high likelihood of recidivism, this will strongly lean in favor of a harsher penalty, in order to protect the community and to underline to the offender that there are severe consequences for criminal behavior.[2] By contrast, a low risk of reoffending leans in favor of a lower penalty. This is because there is less need for community protection and the offender is likely to have reasonable prospects of rehabilitation.[3]

Despite the importance of risk assessment to sentencing decisions, until recently there has been relatively little research conducted on identifying the characteristics of offenders who are likely to re-offend. Thus, these decisions have been traditionally made by reference to the impressionistic sentiments of judges. The trend of human history shows that decisions made without an underlying scientific methodology tend to be compromised and often wrong. Therefore, it is no surprise that when judges make intuitive and unstructured judgements about the future criminal tendency of defendants, they are very often inaccurate. Research shows that they are breathtakingly wrong: barely more accurate than if they tossed a coin to determine if a defendant was likely to reoffend.[4]

Artificial intelligence has made remarkable advances in the last ten years, and is now making inroads into legal decision-making.[5] Sentencing is one area where there is an obvious opportunity for automated-research based technology to inform decision-making. This is because at the sentencing stage of proceedings, the facts are generally already established and there are an extremely large number of variables that are relevant to sentencing decisions. Hence, it is not surprising that in recent years there have been a number of algorithms that have been developed, trialed and sometimes used to guide sentencing decisions.[6]

The most important computer sentencing tools that have been used relate to predictions of reoffending. These instruments have been demonstrated to be more accurate than judicial assessments.[7] Despite this, they have come under considerable criticism. It has been argued that the algorithms supposedly make decisions which incorporate inappropriate considerations (including the racial profiles of offenders) and the integers

1. *See e.g.*, U.S. SENTENCING COMM'N, U.S. SENTENCING GUIDELINES MANUAL 1-16 (2018).

2. *Id.*

3. *Id.*

4. *See infra* Section III.B. and note 105.

5. *See generally* KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE (Cambridge Univ. Press, 2017); Kevin D. Ashley, *A Brief History of the Changing Roles of Case Prediction in AI and Law*, 36 LAW IN CONTEXT 93 (2019), https://doi.org/10.26826/law-in-context.v36i1.88.

6. *See infra* Part III.

7. *See infra* Part III.

which drive the algorithm are not transparent.[8]

The criticisms of algorithms are, however, misguided. They are based on a misunderstanding of the design process of the algorithms and the nature of the data that the algorithms use. In essence, algorithms are formulas. The results produced by these formulas cannot include types of synthesis which have not been embedded into the formula. Quite simply, as long as the formula for the algorithm does not include racist sentiments and the data does not encode racism, the application of the algorithm cannot have a racist orientation.[9]

The current backlash against the use of artificial intelligence ("AI") within criminal justice should be understood and countered by considering two things. First, humans display a very compelling bias against the use of computers in a range of decision-making areas. The bias is termed "algorithmic aversion."[10] Generally, people have an innate and illogical distrust of decisions being made by computers, coloring their acceptance of automated decision-making in areas such as sentencing, bail, and parole determinations. Further, a key point that is missed by critics of algorithms in the criminal justice system is that the current judge-dominated process for making sentencing decisions has been shown to be heavily biased against disadvantaged groups. For example, it has been established that groups such as African Americans and unattractive people receive disproportionately heavier sentences than other people.[11] Algorithms, by contrast, have no subconscious thinking paths—they do exactly what they are programmed to do, and only that. Together, these biases combine to provide the perfect storm of injustice where flawed human-decision making will continue to be seen to be preferable to better computational decision-making.

This is not to say that algorithmic decision-making is perfect, of course. There is evidence that some algorithms do produce outcomes which have a biased or racist orientation. This does not, however, evidence a generic problem with these formulas. Rather, it demonstrates that there are some bad algorithms and some bad datasets. To be clear, it is not that the algorithms that have been produced in the criminal justice domain seem to intentionally discriminate against certain groups. Instead the problem generally relates to the fact that discrimination can occur indirectly. This commonly occurs when variables that are incorporated into an algorithm impliedly discriminate against groups in the community. If, for example, an offense predictive algorithm determines that people with university qualifications have a low risk of offending, this can operate more harshly

8. *See infra* Part III.

9. *Cf*. Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017).

10. *See infra* Part II.

11. *See infra* Part III.

against African Americans.[12] The key to designing accurate and fair algorithms of this nature is ensuring that all of the integers which are coded into the formula do not discriminate directly or indirectly against any cohort in the community. This is achievable, but it requires an acute understanding of the types of sentencing considerations—such as prior criminal history, marital status, educational level—that can serve as proxies for immutable human traits, such as race and gender.[13]

In this article, we discuss the best methodology for making accurate decisions regarding future criminal offending. This Article proposes a key reform: the sentencing system would be considerably improved if risk assessments were made with algorithms based on large data sets of information relating to the factors that suggest recidivism. Another important recommendation made in this Article is that the integers that inform the algorithm must be transparent and made publicly available. This will ensure that the algorithm does not produce results which are biased against any groups. It will also provide the opportunity for ongoing testing, evaluation, refinement and improvement of the algorithm. The algorithm developed in this context can then also be used or adapted in other areas of the criminal justice system where the risk of recidivism is a cardinal consideration, namely bail and probation decisions. Indeed, as we discuss, risk assessment algorithms are already used relatively extensively in relation to probation decisions, however, significant improvements can be made to the design of such instruments.

In the sentencing context, algorithms regarding the likelihood of recidivism have assumed high level importance with the recent passing of the First Step Act in December, 2018.[14] This has been hailed as the most significant criminal justice legislative reform in decades.[15] The First Step Act introduces prison reforms, as well as sentencing changes, and includes several measures that will reduce the length of prison terms for some offenders and consequently lower the number of inmates in federal prisons. The Act is expected to apply to approximately 30% of federal

12. Andrew Howard Nichols & J. Oliver Schak, *Degree Attainment for Black Adults: National and State Trends*, THE EDUCATION TRUST (2014) https://edtrust.org/wp-content/uploads/2014/09/Black-Degree-Attainment_FINAL.pdf.

13. On the question of proxies, *see generally* Anya Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257 (2020) (demonstrating the dagger of certain neutral-seeming features in data being proxies for other clearly discriminatory features).

14. For details about the Act, *see* Douglas A. Berman, *Prez Trump Signs Historic (Though Modest) First Step Act into Law…and Now Comes the Critical Work of Implementing It Well!!*, SENT'G L. AND POL'Y BLOG (Dec. 21, 2018), https://sentencing.typepad.com/sentencing_law_and_policy/2018/12/prez-trump-signs-historic-though-modest-first-step-act-into-law-and-now-comes-the-critical-work-of-i.html [https://perma.cc/G5BE-58CD].

15. *Id.*

prisoners.[16] The decision of whether to reduce an inmate's sentence is to be made in accordance to a risk assessment algorithm.[17] The Act requires the Attorney General to create a "Risk and Needs Assessment System" to ascertain all inmates' risk of recidivism and the evidence-based recidivism reduction programs that will best suit them, and to provide inmates with access to these programs.[18] The AI system to implement these changes has not yet been developed, and so the recommendations in this Article are highly pertinent and timely.[19]

In the next part of the Article, we explain the nature of algorithms and the advantages of computer-decision making over judgments made by people. In Part II, we discuss the reasons that the uptake of algorithms has been slow. As we discussed in this Part, current research indicates the people have an irrational aversion to use of algorithms in certain contexts. In Part III, we explore the current manner in which risk assessment decisions are made in sentencing. This is followed in Part IV by an analysis of the criticisms of criminal justice algorithms. The manner in which these criticisms can be surmounted is set out in Part V. Reform proposals are made in the Part VI.

I. THE NATURE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence has been in existence for several decades; however, the concept is only now starting to attract a degree of mainstream recognition. Like many emerging developments, it is still not well understood. In crude terms, current data-driven artificial intelligence systems synthesize large amounts of data involving prior action or behavior to make predictions about future behavior. The way in which the data is processed is the key to the efficacy and integrity of AI. The data is processed by a formula, termed an algorithm. As noted by the Pew Research Center, algorithms are not new. They are simply "instructions for solving a problem or completing a task. Recipes are algorithms, as are math equations. Computer code is algorithmic."[20] The increasing use of

16. Gina Martinez, *The Bipartisan Criminal-Justice Bill Will Affect Thousands of Prisoners. Here's How Their Lives Will Change*, TIME (Dec. 20, 2018), http://time.com/5483066/congress-passes-bipartisan-criminal-justice-reform-effort/ [https://perma.cc/8GU4-XAZE].

17. There are concerns about the capacity to develop the instrument, *see* Press Release, Jerrold Nadler & Karen Bass, *Statement on DOJ's Selection of the Hudson Institute to Host First Step Act Independent Review Committee*, HOUSE COMMITTEE ON THE JUDICIARY (Apr. 23, 2019), https://judiciary.house.gov/news/press-releases/nadler-bass-statement-doj-s-selection-hudson-institute-host-first-step-act [https://perma.cc/X75B-EBUF].

18. *Id.*

19. See *NIJ's Role Under the First Step Act*, NAT'L INST. OF JUST. (June 20, 2019), https://www.nij.gov/topics/corrections/reentry/Pages/first-step-act.aspx [https://perma.cc/K7VH-ZS6N].

20. Lee Rainie & Janna Anderson, *Code-Dependent: Pros and Cons of the Algorithm Age*, PEW RES. CTR. (Feb. 8, 2017), http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-

algorithms stems in a large part from the fact that presently "massive amounts of data are being created, captured and analyzed by businesses and governments."[21] Algorithms already play a key role in many aspects of society from risk assessments for insurance premiums to detection of tax fraud,[22] and controlling the timing of lights that facilitate traffic flow.[23]

Artificial intelligence uses algorithms to process and synthesize vast amounts of information and provide answers to problems. Thus, there is an inextricable connection between algorithms and artificial intelligence. All artificial intelligence systems are based on algorithms, however, most algorithms do not operate within the context of an artificial intelligence construct. The main advantages from artificial intelligence systems which incorporate algorithms are that they are capable of providing accurate and efficient answers and solutions to problems that often require the computation or assessment of a large number of variables. The are no limits to the types of subject areas in which AI can operate. One of the most commonly used forms of AI is Siri, which is a virtual assistant which uses voice recognition to provide answers to users of iPhones. Other common examples include ridesharing apps used by entities such as Uber to anticipate driver demand,[24] plagiarism checkers such as "Turnitin,"[25] and Facebook which uses AI to suggest friends.[26]

The recent explosion of interest in AI has been driven by advances in neural network technology, especially what is generally referred to as deep learning systems.[27] At its core, deep learning is a statistical method for classifying patterns based on large amounts of sample data using neural networks that have multiple layers. The networks are constructed with input nodes connected to output nodes via a series of "hidden" nodes which are arranged in a series of layers. The input nodes can represent

algorithm-age/ [https://perma.cc/C9DT-BBTJ].

21. *Id.*

22. Ric Simmons, *Quantifying Criminal Procedure: How To Unlock The Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947, 955 (2016).

23. *Id.* at 1013.

24. Daniel Faggella, *Everyday Examples of Artificial Intelligence and Machine Learning*, EMERO, https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/ [https://perma.cc/CK2K-3XEX] (last updated Apr. 11, 2020).

25. *Id.*

26. *Id.*

27. The field exploded in 2012 when Krizhevsky, Sutskever, and Hinton demonstrated remarkable results in image classification and object recognition using large scale multi-layer, deep networks, *see* Alex Krizhevsky, Ilya Sutskever & Geoffrey E Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, 1 NIPS 1097 (2012). Similar work was being undertaken elsewhere. *See* Dan Cireşan et al., *Multi-Column Deep Neural Network for Traffic Sign Classification*, 32 NEURAL NETWORKS 333 (2012). The seminal review by the leaders in the field is Yann LeCun et al, *Deep Learning*, 521 NATURE 436 (2015).

any data—in the examples of image recognition and speech recognition, they involve pixels or words—and the outputs involve the decision or coding that the researcher is looking for, e.g., the classification of a picture or the meaning of the sentence. All of the nodes (or "neurons") within the network have activation levels, so that a neuron will "fire" if the nodes connected to it add up to a certain activation level or higher. All of the connections initially have a random weight assigned to them, but by using a large training set and a process called back-propagation, eventually the activation levels and weighting are adjusted to the point where any given input will produce the correct output.[28]

A simple example may help to understand how these systems work. Imagine that we have a dataset that provides historical data on every sentencing decision for all criminal defendants in a given jurisdiction. This dataset contains all of the salient factors as inputs to the sentencing decision—the presence of mitigating factors like contrition or juvenile status, the presence of aggravating factors like recidivism or violence, the name of the judge, the nature of the crime, etc.—along with some presumably irrelevant considerations—for example, the time of day of the decision, the color of the defendant's clothes, and so on—along with the eventual sentence given for each case. The sentencing factors are the inputs on the network, and the sentencing determinations are the outputs. The network is initially coded with random activations and weightings, and so it cannot predict accurately the outcome of any case. But if we train it with hundreds of cases—or better, hundreds of thousands of cases—where we know the factors and the sentences, then we will eventually have a fully trained network where the outcome of an undecided case can be predicted accurately based on the presence or absence of various inputs.[29]

Deep neural networks have made good on the promise that, one day, machines could actually learn. This type of AI is now widely applied across a range of legal areas.[30] A number of technology vendors have demonstrated the ability of big-data driven statistical and quantitative techniques to assess the quality of an attorney based on their litigation history,[31] the disposition of legal cases in patent litigation and Supreme

28. *See* IAN GOODFELLOW, YOSHUA BENJIO & AARON COURVILLE, DEEP LEARNING 200 (MIT Press, 2016).

29. For a serious analysis of the limitations of deep learning systems, *see generally* GARY MARCUS, DEEP LEARNING: A CRITICAL APPRAISAL, (at Xiv, Jan 2, 2018), https://arxiv.org/abs/1801.00631.

30. *See e.g.*, Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87 (2014).

31. Daniel Martin Katz, *The 2012 Randolph W. Thrower Symposium Innovation For The Modern Era: Law, Policy, and Legal Practice in a Changing World: Article: Quantitative Legal Prediction – Or – How I Learned to Stop Worrying and Start Preparing For The Data-Driven Future of The Legal Services Industry*, 62 EMORY L.J. 909, 932-34 (2013).

Court determinations,[32] and the likely attorney costs to be awarded in a range of cases.[33] These sorts of systems are also commercially available in technology assisted document review—also known as "predictive coding" in e-discovery—and in large scale contract review. In these situations, deep learning approaches involve training a neural net on a subset of documents that are known to be relevant to the discovery question or due diligence question, and then having the system categorize the remaining, uncategorized documents.[34]

The other important area of big-data analytics/machine learning in law is in the criminal justice field, especially in the area of recidivism assessment. The commercial success of prior data driven recidivism assessment systems like Northpointe's COMPAS, have been balanced with research that questions their accuracy, utility, and fairness.[35] It is to this question that we now turn.

---

32. *See generally* Daniel Martin Katz, Michael J. Bommarito II & Josh Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, *PLOS ONE* 12(4): e0174698 (2017), https://doi.org/10.1371/journal.pone.0174698 [https://perma.cc/GLP7-UB4Q] (demonstrating the use of a random forest classifier algorithm to predict US Supreme Court decisions with greater accuracy than support vector machines or deep layer neural networks); Andrew D. Martin et al., *Competing Approaches to Predicting Supreme Court Decision Making*, 2(4) PERSPECTIVES ON POLITICS 761, 761–68 (2004) (describing a statistical model of Supreme Court outcomes based upon various factors including the political orientation of the lower opinion and the circuit of origin of the appeal that outperformed experts in predicting Supreme Court outcomes and highlighted data relationships not previously understood); Andrew D. Martin & Kevin M. Quinn, *Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999*, 10 POLITICAL ANALYSIS 134 (2002); Theodore W Ruger et al., *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-Making*, 104 COLUM. L. REV. 1150 (2004); Isha Salian, *"Moneyball" Legal Analytics Helps Lawyers Assess Judges*, S.F. CHRONICLE (July 14, 2017) www.sfchronicle.com/business/article/Moneyball-legal-analytics-helps-lawyers-11289892.php.

33. Katz, *supra* note 31 at 929-31.

34. *See* Fed. Housing Fin. Agency v. HSBC North America Holdings Inc., et al., 2014 WL 584300, at *3 (S.D.N.Y. Feb. 14, 2014) ("predictive coding had a better track record in the production of responsive documents than human review"); Monique Da Silva Moore, et. al. v. Publicis Groupe & MSL Group, 287 F.R.D. 182, 193 (S.D.N.Y. 2012) (holding computer-assisted review appropriate in some cases); Nat'l Day Laborer Org. Network v. U.S. Immigration & Customs Enforcement Agency (NDLON), 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012) ("[P]arties can (and frequently should) rely on . . . machine learning tools to find responsive documents."). *But see*, Charles Yablon & Nick Landsman-Roos, *Predictive Coding: Emerging Questions and Concerns*, 64 S.C. L. REV. 633 (2013) (discussing some limitations of these systems).

35. *See e.g.*, Tom Simonite, *How to Upgrade Judges with Machine Learning*, MIT TECH. REV. (Mar. 6, 2017), www.technologyreview.com/s/603763/how-to-upgrade-judges-with-machine-learning [https://perma.cc/3AE9-23UC]; Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008); Sonja B Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 805 (2014); Daniel Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (Cambridge: Harvard University Press, 2015). *But see* Nigel Stobbs, Dan Hunter & Mirko Bagaric, *Can Sentencing be Enhanced by the Use of Artificial Intelligence?*, 41 CRIM. L.J. 261 (2017); Harry Surden, *The Ethics of Artificial Intelligence in Law: Basic Questions*, OXFORD HANDBOOK OF ETHICS OF AI (forthcoming 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3441303.

II. THE TRUST DEFICIT WITH AI DECISIONS

Although algorithms in a wide range of areas have now been around for some time, most studies have concentrated on consumer algorithms, marketing algorithms and social media algorithms which are likely to affect choice, governance and social behaviors. Developments in technology have allowed for further growth of algorithms in all areas of human life. Tensions exist and continue to develop around the ethics, transparency and fairness of algorithmic decision-making, specifically around decisions predominately or at least historically made by humans. One reason for distrust of algorithms is widespread confusion regarding their functionalities and in particular the manner in which computer systems are capable of self-learning. This leads to fears that AI will trump human sovereignty. This fear is misplaced, however. As noted in the discussion below, computers are capable of self-learning, but the autonomous learning relates only to acquiring and collating information regarding the domain in which the computer operates and then applying that to the formula where this knowledge can operate. Importantly, the formula is always coded by human beings who set the parameters of the computer's decision-making capabilities.

Numerous studies have been undertaken which consider these tensions, in particular the perceived lack of trust and lack of control around algorithmic decisions.[36] Indeed, researchers in this field have gone as far as to label this lack of control and bias in favoring human forecasting and outcome predicting as "algorithmic aversion."[37] Essentially, algorithmic aversion, as coined by Dietvorst in his studies in this field, refers to the phenomenon of a positive bias towards human-based decision-making, even when an algorithm has proven more competent than its human counterpart.[38] One theme seen throughout various studies in human-automation trust research is that humans expect algorithmic perfection—

36. *See generally* Berkeley Dietvorst, People reject (superior) algorithms because they compare them to counter-normative reference points (Dec. 6, 2016) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2881503; Berkeley J. Dietvorst, Joseph P. Simmons & Cade Massey, *Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them*, 64 MGMT. SCI. 1155 (2016); Berkeley J. Dietvorst, Joseph P. Simmons & Cade Massey, *Algorithm aversion: People erroneously avoid algorithms after seeing them err*, 144 J. OF EXPERIMENTAL PSYCHOL. GEN. 114 (2015) [hereinafter *Algorithm Aversion Article*]; Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, Paper Presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Aug. 13-17, 2017), https://dl.acm.org/doi/10.1145/3097983.3098095; Min Kyung Lee & Su Baykal, *Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division*, Paper Presented at the CSCW (Feb. 2017), ttps://www.researchgate.net/publication/313738865_Algorithmic_Mediation_in_Group_Decisions_Fairness_Perceptions_of_Algorithmically_Mediated_vs_Discussion-Based_Social_Division.

37. *Algorithm Aversion Article*, *supra* note 36.

38. *Id.* at 114.

meaning zero errors—whilst permitting humans to be imperfect and to make mistakes and still favoring human decision-making.[39] In fact, studies have shown that people prefer flawed human forecasts to flawless algorithmic forecasts.[40] The reason why humans are so averse to trusting algorithms in making correct predictions and decisions is based on several themes that have been deduced by these studies. The main themes around aversion to algorithmic-based decisions and judgments falls into the broad categories of trust/control/transparency which underpin the basics of human nature and social norms. These themes will be discussed in more detail below, considering the current literature surrounding this phenomenon.

As discovered in recent studies,[41] humans are unlikely to use an algorithmic decision when there is a comparable, if somewhat inferior human decision/prediction which they could use instead. The literature affirms that transparent decision-making processes play an important role in justifying any decisions made. Hence, while humans may make mistakes and errors in judgment, they can in turn, be held accountable to rationalize their processes used in arriving at their decision.[42] Indeed, in some cases seen in the literature, intelligent system decisions may be better trusted when they utilize a built-in explanation system[43] which explains to the affected person how the decision was reached. But, for some, the level of detail these explanation systems use may not be sufficient to warrant trust in the system.[44]

A recent study by Binns, et al., looked at the effects that explanations have on people's perceptions of algorithmic decisions.[45] The study had participants review scenarios where an algorithm made decisions for

39. Andrew Prahl & Lyn Van Swol, *Understanding algorithm aversion: When is advice from automation discounted?*, 36 J. OF FORECASTING 691 (2017); Paul Goodwin, M Sinan Gönül & Dilek Önkal, *Antecedents and effects of trust in forecasting advice*, 29 INT'L J. OF FORECASTING 354 (2013).

40. Dalia L. Diab et al., *Lay perceptions of selection decision aids in US and non-US samples*, 19 INT'L J. OF SELECTION AND ASSESSMENT 209 (2011); Dietvorst, Simmons & Massey, *Algorithm aversion: People erroneously avoid algorithms after seeing them err*, 144 J. OF EXPERIMENTAL PSYCHOL. GEN. 114 (2015); Joseph Eastwood, Brent Snook & Kirk Luther, *What people want from their professionals: Attitudes toward decision-making strategies*, 25 J. OF BEHAV. DECISION MAKING 458 (2012).

41. *See Algorithm Aversion Article*, *supra* note 36.

42. Reuben Binns et al., It's Reducing a Human Being to a Percentage: Perceptions of Justice in Algorithmic Decisions, Paper Presented at the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Apr. 21-26, 2018), https://arxiv.org/pdf/1801.10408.pdf.

43. L Richard Ye and Paul E Johnson, *The impact of explanation facilities on user acceptance of expert systems advice*, 19 MIS Q. 157 (1995).

44. Adrian Bussone, Simone Stumpf & Dympna O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, Paper presented at the International Conference on Healthcare Informatics (ICHI) (Oct. 2015), https://www.researchgate.net/publication/283079634_The_Role_of_Explanations_on_Trust_and_Reliance_in_Clinical_Decision_Support_Systems.

45. Binns et al., *supra* note 42.

loans, insurance and so forth based on the data/information input of the hypothetical user. The major themes that emerged from this study included concerns around the lack of human touch, lack of understanding around interpretations of the system's reasoning, lack of knowledge about the use of statistical inference, and uncertainty over the degree of actionability in an explanation and important aspects which were unaccounted for by the system.[46] Participants stated that algorithmic decision-making could be "impersonal" or "dehumanizing" for recipients of the decision. They also considered the lack of negotiation and opportunity for human interaction as a negative. Participants also reflected on the generalization and statistical inference of the decision-making process as unacceptable, stating that "this is just simply reducing a human being to a percentage."[47]

Other studies have seen similar results, with Prahl and Van Swol investigating trust factors associated with automated advice versus human advice.[48] The authors considered advice response theory (ART), a concept generally used in interpersonal advice where characteristics such as politeness of message, expertise of the advisor and the emotional state of the receiver all play at part in usage of the advice. However, due to algorithmic advice sometimes being presented numerically, rather than through words, and with no interpersonal attributes, the authors chose to focus on advisor characteristics only (such as expertise or credibility) to determine how people react to human or computer advice.

The authors argued that perceived competence, credibility and expressed confidence are linked to advice usage.[49] Prahl and Van Swol argue that algorithms are generally evaluated only on their competence and nothing else. Due to the lack of interpersonal connection between the algorithm and the human, human-like characteristics such as emotions, social cues and intentions were not considered in the study.[50] Therefore the study focused specifically on competency of advice in advice utilization.

The study considered five hypotheses related to advice utilization in favor of human advice. Differing from other similar studies, the output of the advice was numerical only, presented in identical formats. The first hypotheses posited—that human advice would be favored over the algorithm's advice—was not supported. Notably, hypothesis two—which posited that after receiving bad advice from the algorithm, participants would defer to advice from human advisors—was supported. This finding

46. *Id.* at 6.

47. *Id.* at 7.

48. *See generally* Prahl & Van Swol, *supra* note 39.

49. *Id.* at 692.

50. *Id.*

supports the "perfection schema" theory which suggests that when devices make an error, it feels "especially negative to the advice recipient and they lose trust rapidly."[51] The authors state that participants viewed an error as indicative of a fundamental flaw of the algorithm which would reoccur, while humans who err have the ability to correct and improve on their performance over time.

Similarly, aspects of fairness and transparency also play a role in determining trust and confidence in algorithmic decisions. In a study examining algorithmic decision-making and perceptions of fairness, researchers found that the algorithms they used did not allow for "multiple concepts of fairness, altruistic behaviors and norms, or the social psychology of users."[52] They also discovered that "fair division algorithms" make several assumptions, namely that users will be rational, users will apply the same intensity to their preferences and that user inputs will reflect their true preferences.

Interestingly, the researchers touch on a key issue associated with fairness in relation to algorithmic decisions where the algorithm determines the final outcome: "is fairness based on equal distribution of resources, regardless of the people those resources are distributed to, or [is] distribution only fair if it takes individual differences into account?" The authors empirically studied individuals' experiences and perceptions of algorithmically mediated group division where the "fair division" algorithm determines the final outcome. Using the website Spliddit, they investigated the division of rent, house chores, snacks and credit for a game outcome to determine fair solutions for the participants. Following the division tasks, the participants were asked to rate, on a 7-point Likert scale, whether they agreed or disagreed that the divisions were fair to themselves, to others and to the group as a whole. They were then interviewed to discuss the outcomes of the Spliddit results.

What the authors found was that there are multiple concepts of fairness. Some participants agreed that the algorithmic divisions were somewhat fair on average, while others rated the divisions high when the outcomes reflected their preferences. Conversely others preferred equal distribution even when it was not "fair" to their own interests and some considered preferences and even distribution of equal fairness. The study also found that the input interfaces of Spliddit were akin to potential biases as they did not always embody accurate assumptions about users. For example, it assumed that each participant cared in equal amounts about the baseline task, while, in reality, some had strong preferences for or against the

51. Poornima Madhavan & Douglas A Wiegmann, *Similarities and differences between human–human and human–automation trust: an integrative review*, 8 THEORETICAL ISSUES IN ERGONOMICS SCI. 277, 297 (2007).

52. Lee & Baykal, *supra* note 36.

specific tasks and were willing to increase or reduce their overall input.

Human behavior, emotions and social norms play a decisive part in determining people's perceptions of fairness. As noted by one of the participants of the study, "we do our best to make people happy … [b]ut with the computer there's no emotions in it."[53] Participants also compared the algorithmic decision-making through Spliddit to discussion-based decision-making. They noted advantages in discussion since they were made aware of other participant's preferences and could hence reach what they considered fairer results.

This led the authors to create another study to compare algorithmically-mediated versus discussion-based divisions. Considering social justice and fairness literature which suggests "greater perceived control over and trust in the decision-making process increase[s] people's fairness of outcomes,"[54] the authors set a study up where participants would have perceived control over the process, which they determined would increase their fairness perceptions of the algorithm.

Overall, the researchers found similar results to study one—that participants thought decisions made via discussion were fairer than those of the algorithm. Interestingly, this study also considered other variables, including interpersonal power and fairness and influence of choice and social transparency. The effects of these variables found that participants were more likely to blame their own choices, as they had volunteered to do certain tasks during the study with high interpersonal power. These participants judged discussions as more fair, while those with low interpersonal power felt similar judgements to the algorithm and discussion-based decision. The influence of choice meant that some participants blamed their own choices for the outcomes they got, which they perceived as "fair." They also assumed other participants perceived their own results as fair due to voluntary choice. Again, the variable of social transparency through discussion led to compromise and understanding of others' preferences, increasing the overall perception of fairness.

In sum, the literature suggests that people prefer advice from human advisors rather than from automation and tend to discount automation advice in favor of less than perfect human advice.[55] As noted by Prahl and Van Swol, this discounting of automation advice is known in clinical psychology research as the "clinical versus actuarial" debate[56] or what

53. *Id.*

54. *Id.*

55. Goodwin, Gönül & Önkal, *supra* note 39.

56. Robyn M. Dawes, *The robust beauty of improper linear models in decision making*, 34 AM. PSYCHOLOGIST 571 (1979); PAUL E MEEHL, CLINICAL VERSUS STATISTICAL PREDICTION: A THEORETICAL ANALYSIS AND A REVIEW OF THE EVIDENCE (1954).

Dietvorst refers to as "algorithmic aversion."[57]

As we shall see below, algorithmic aversion is manifest in literature regarding the desirability of the use of algorithms in the criminal justice system. Perhaps the most strident and common criticism of these algorithms is that they discriminate against certain groups in the community, especially African-Americans. These criticisms miss the fundamental point that human-decision making in the criminal justice system also has a profound bias against these groups. Before examining this in detail, in order to contextualize the remainder of the discussion, we provide an overview of the sentencing, bail and probation systems.

III. PROBLEMS WITH THE CURRENT APPROACH TO DECISION-MAKING

A. Overview of Sentencing, Bail and Parole

Algorithms are currently used in some areas of criminal justice which involve decisions regarding the risk of future offending. As noted above, there are three main points in the criminal justice system where evaluation of future offending is relevant. They are sentencing, bail, and parole. Although algorithm usage has made some inroads into these areas, the nature and extent of the reliance on algorithms varies markedly across the United States and generally, there is no doctrinal basis underpinning the use of these instruments and no clear direction regarding their future use. Prior to discussing the current use of algorithms in these contexts, we provide an overview of the criminal justice stages where the likelihood that a defendant will commit future criminal offences is an important consideration. These stages are: sentencing, bail and parole. We consider them in that order.

1. Sentencing Law and Practice

Sentencing is the process whereby courts impose sanctions on offenders. The sentencing systems in the Federal jurisdiction and the 50 States are different;[58] however, they have similar overarching frameworks in that they share similar objectives in the form of retribution, specific deterrence, general deterrence, rehabilitation and community protection (known as incapacitation in some jurisdictions).[59] Though each case and

57. *Algorithm Aversion Article*, *supra* note 36.

58. Sentencing (and more generally, criminal law) in the United States is mainly the province of the states. *See* United States v. Morrison, 529 U.S. 598, 613 (2000) (citing U.S. v. Lopez, 514 U.S. 549, 564 (1995)).

59. *See* UNITED STATES SENTENCING COMMISSION GUIDELINES MANUAL (U.S. SENT'G COMM'N 2016), http://www.ussc.gov/guidelines/2016-guidelines-manual [hereinafter U.S. SENT'G COMM'N 2016].

jurisdiction places different emphasis on these goals, community protection is widely regarded as being the paramount consideration.[60] The advent of severe prescriptive sentencing laws[61] that are in place in all State and Federal jurisdictions have been driven largely by the perceived need to protect the community.[62]

Extensive guideline sentencing is now used in twenty jurisdictions across the United States[63] This type of proscription means that sentencing grids are used to outline prescribed penalties, and penalties are calculated principally by reference to two considerations: criminal history and offense severity.[64] Criminal history is effectively used as the key proxy for the likelihood of future offending.

The US Sentencing Commission Guidelines—often referred to as the "Federal Sentencing Guidelines" or the "Guidelines"—are key to understanding how prescribed penalty laws and guideline sentencing works in the US. The Guidelines have affected the development of state sentencing systems and determined the sentence for offenders, more they any other system.65 As Hamilton has noted, by one measure the federal government has the largest criminal justice system in the U.S., and the federal prison system—leaving aside state counterparts—is larger than the prison systems of most countries.[66] Additionally it has been accepted that:

> … history proves that decisions made in Washington affect the whole criminal justice system, for better or worse. Federal funding drives state policy, and helped create our current crisis of mass incarceration. And the federal government sets the national tone, which is critical to increasing public support and national momentum for change. Without a strong national movement, the bold reforms needed at the state and local level

60. NAT'L RESEARCH COUNCIL, THE GROWTH OF INCARCERATION IN THE UNITED STATES, EXPLORING CAUSES AND CONSEQUENCES 9 (Jeremy Travis et al., eds., 2014) [hereinafter NAT'L RESEARCH COUNCIL].

61. For the purposes of clarity, these both come under the terminology of fixed or standard penalties in this Article.

62. NAT'L RESEARCH COUNCIL, *supra* note 60, at 325.

63. Alabama, Kansas, Oregon, Alaska, Maryland, Pennsylvania, Arkansas, Massachusetts, Tennessee, Delaware, Michigan, Utah, District of Columbia, Minnesota, Virginia, Federal (U.S. courts), North Carolina, Washington, Florida, Ohio. *See* Richard S. Frase & Kelly Lyn Mitchell, *What Are Sentencing Guidelines?*, U. OF MINN. ROBINA INST. CRIM. L. AND CRIM. JUST. (Mar. 21, 2018), http://sentencing.umn.edu/content/what-are-sentencing-guidelines [https://perma.cc/37QZ-J9CW].

64. This is based mainly on the number, seriousness, and age of the prior convictions.

65. *See* Douglas A. Berman & Stephanos Bibas, *Making Sentencing Sensible*, 37 OHIO ST. J. CRIM. L. 37, 40 (2006). There are more than 200,000 federal prisoners. *See* E. Ann Carson, *Prisoners in 2013*, BUREAU OF JUST. STAT. (Sept. 16, 2014), http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5109 [https://perma.cc/EVK7-F4DF]. Also, as noted below, the broad structure of the Federal Guidelines is similar to many other guideline systems in that the penalty range is not mandatory and permit departures in certain circumstances.

66. Melissa Hamilton, *Sentencing Disparities*, 6 BRIT. J. AM. LEG. STUD 178, 182 (2017).

cannot emerge.[67]

The Supreme Court, in *United States v. Booker*,[68] concluded that the Guidelines are only advisory, however, they have had an outsized influence on sentencing decisions[69] Recent data establishes that courts are still considerably influenced by the guideline range in sentencing a penalty. In 2015, 47% of sentences were in line with the Guidelines in 2015, 49% in 2016 and 2017[70] and 51% in 2018.[71]

In keeping with other grid sentencing systems, the Guidelines uses a formula where the offenders' previous convictions and seriousness of the offence impact dramatically the penalties imposed.[72] That is not to say that offence history and severity are the only factors involved, however. The Guidelines list all the factors which can affect the sanction, including "adjustments" and "departures," which allow to deviation from the Guidelines due to mitigating or aggravating circumstances.[73] Adjustments are alterations to the sentence by a fixed amount.[74] By way of example, an offender may get a reduction of three levels if there was an early guilty

67. AMES C. GRAWERT, NATASHA CAMHI & INIMAI CHETTIAR, Brennan Ctr. for Just., A FEDERAL AGENDA TO REDUCE MASS INCARCERATION 1 (2017), https://www.brennancenter.org/sites/default/files/publications/a%20federal%20agenda%20to%20reduce%20mass%20incarceration.pdf [https://perma.cc/KG4T-NFBR].

68. U.S. v. Booker, 543 U.S. 220 (2005). In *Booker*, the Supreme Court held that aspects of the Guidelines that were mandatory were contrary to the Sixth Amendment right to a jury trial.

69. Sarah French Russell, *Rethinking Recidivist Enhancements: The Role of Prior Drug Convictions*, 43 U.C. DAVIS L. REV. 1135, 1160 (2010); *see also* AMY BARON EVANS & JENNIFER NILES COFFIN, NO MORE MATH WITHOUT SUBTRACTION: DECONSTRUCTING THE GUIDELINES' PROHIBITIONS AND RESTRICTIONS ON MITIGATING FACTORS (2011), https://www.fd.org/sites/default/files/criminal_defense_topics/essential_topics/sentencing_resources/deconstructing_the_guidelines/no-more-math-without-subtraction.pdf [https://perma.cc/66LS-H7TJ]. For a discussion regarding the potential of mitigating factors to have a greater role in federal sentencing *see* William W. Berry III, *Mitigation in Federal Sentencing in the United States*, *in* MITIGATION AND AGGRAVATION AT SENTENCING 247 (Julian V. Roberts ed., 2011). U.S. SENTENCING COMM'N, FINAL QUARTERLY DATA REPORT, FISCAL YEAR 2014 (2014), http://www.ussc.gov/sites/default/files/pdf/research-and-publications/federal-sentencing-statistics/quarterly-sentencing-updates/USSC-2014_Quarterly_Report_Final.pdf [https://perma.cc/W9Q5-HYRH].

70. U.S. SENTENCING COMM'N, ANNUAL REPORT, FISCAL YEAR 2016 (2016), http://www.ussc.gov/about/annual-report-2016 [https://perma.cc/L9W4-GCGF]; U.S. SENTENCING COMM'N, ANNUAL REPORT, FISCAL YEAR 2017 (2017), http://www.ussc.gov/about/annual-report-2017 [https://perma.cc/YF85-AX2Z].

71. U. S. SENTENCING COMM'N, 2018 ANNUAL REPORT (2018), https://www.ussc.gov/about/annual-report-2018 [https://perma.cc/HHJ5-W8GQ].

72. *See* Carissa Byrne Hessick, *Why Are Only Bad Acts Good Sentencing Factors?*, 88 B.U. L. REV. 1109, 1135-36 (2008).

73. AMY BARON EVANS & PAUL HOFER, NAT'L SENTENCING RESOURCE COUNSEL, LITIGATING MITIGATING FACTORS: DEPARTURES, VARIANCES, AND ALTERNATIVES TO INCARCERATION, i (2011), https://static1.squarespace.com/static/551cb031e4b00eb221747329/t/5883e40717bffc09e3a59ea1/1485038601489/Litigating_Mitigating_Factors.pdf.

74. These are set out in Chapter 3 of the U.S. Sentencing Guidelines., U.S. SENTENCING GUIDELINES MANUAL 357-91 (U.S. SENTENCING COMM'N 2016).

plea or if there is a showing of remorse.[75] Departures[76] are the most common way for sentences to be handed down outside the Guidelines prescribed range.[77] Additionally, under 18 U.S.C. § 3553, the courts are permitted to use considerations not specified in the Guidelines to justify departures from the suggested guideline range.[78] However, judges must outline explicit reasons for not following the Guidelines stated range in sentencing.[79]

Importantly, for the purposes of this Article, key sentencing objectives that inform the structure of the Guidelines and the application of some departures and adjustments are community protection, rehabilitation and specific deterrence. The risk that an offender will reoffend is a cardinal consideration to these factors. The goal of community protection is best advanced by placing offenders who are at risk of reoffending in prison (or in rare cases, executing them). Specific deterrence is the theory that offenders can be discouraged from reoffending by imposing harsh sanctions on them, typically prison terms, in an attempt to teach them that crime does not pay off. It is especially relevant to offenders who are felt to be at risk of reoffending.[80] When this consideration is relevant, it serves to increase sentence severity. Rehabilitation aims to invoke internal attitudinal reform in offenders by educating them that criminal behavior is inappropriate. It operates to reduce sentence severity. An important consideration regarding whether rehabilitation is tenable is an assessment of whether an offender is likely to reoffend.[81] Offenders who are regarded as being incorrigibly bad are poor candidates for rehabilitation and hence will not receive a penalty reduction on this account. Thus, an assessment of an offender's recidivism rating is a crucial decision in the sentencing calculus.

75. *Id.* § 8C2.5 cmt. n.14. However, section 5K2.0(d)(4) provides that the court cannot depart from a guideline range as a result of "[t]he defendant's decision, in and of itself, to plead guilty to the offense or to enter a plea agreement with respect to the offense (i.e., a departure may not be based merely on the fact that the defendant decided to plead guilty or to enter into a plea agreement, but a departure may be based on justifiable, non-prohibited reasons as part of a sentence that is recommended, or agreed to, in the plea agreement and accepted by the court." *See also Id.* § 6B1.2.

76. *Id.* § 5K.

77. *Id.* § 1A.

78. *Id.* § 5K2.0(a)(2)(B). *See also* Gall v. United States, 552 U.S. 38 (2007); Pepper v. United States, 131 U.S. 476 (2011).

79. U.S. SENTENCING GUIDELINES MANUAL § 5K2.0(e) (U.S. SENTENCING COMM'N 2016).

80. Mirko Bagaric & Theo Alexander, *Specific Deterrence Doesn't Work, Rehabilitation Might and What it Means for Sentencing*, 35 CRIM. L.J. 159 (2012).

81. Mirko Bagaric, et al, *Mitigating America's Mass Incarceration Crisis Without Compromising Community Protection: Expanding The Role of Rehabilitation in Sentencing*, 22 LEWIS & CLARK L. REV. 1 (2018).

2. Bail Law and Practice

When defendants are charged with criminal offenses, they are either released back into the community or placed in custody, pending the finalization of the charges. If they are released into the community, this mechanism is termed bail. Traditionally, bail was thought of as "the posting of security to ensure the presence of an accused at subsequent judicial proceedings."[82] While posting security is a common aspect of bail, in it is not essential in all jurisdictions in the United States. However,

> today, an individual's release pending subsequent criminal proceedings is often predicated on conditions other than, or in addition to, the posting of an appearance bond, secured or unsecured. As a consequence, rather than speaking of bail, existing federal law refers to release or detention pending trial, to release or detention pending sentencing or appeal, and to release or detention of a material witness.[83]

The principal purpose of bail is to ensure that a defendant will appear in court for his or her trial.[84] The other important objective of bail is to ensure that an accused does not commit an offense pending trial.[85]

In many cases, the person charged will be required to pay the court a sum of money, set either by a schedule for minor offenses or by the judge at the first appearance for more serious crimes, which they forfeit if they fail to appear back in court when required. The Eighth Amendment to the Constitution of the United States, which applies only to federal pre-trial detention, provides that "Excessive bail shall not be required," and most states have similar statutory provisions prohibiting the imposition of excessive bail.[86] Yet there are approximately 450,000 people currently in custody awaiting trial as a result of being unable to afford bail.[87]

Unsurprisingly, there is an over-representation among this pre-trial detainee population of the most vulnerable and disadvantaged demographics within the community.[88] Those least likely to be able to afford cash bail are the poor, indigent, unemployed, undereducated, or mentally ill.[89] These detainees are sometimes arrested for vice or street offences such as public intoxication, or minor crimes such as failure to pay fines or driving with a suspended license. The effect then is that what

82. CHARLES DOYLE, CONG. RES. SERV., R40221, BAIL: AN OVERVIEW OF FEDERAL CRIMINAL LAW 1 (2017), https://fas.org/sgp/crs/misc/R40221.pdf [https://perma.cc/ATX2-ZPJ5].

83. *Id.*

84. *Id.* at 2.

85. *Id.*

86. James A. Allen, Note, *Making Bail: Limiting the Use of Bail Schedules and Defining the Elusive Meaning of Excessive Bail*, 25 J.L. & POL'Y 637, 640 n.11 (2016).

87. *Id.* at 640.

88. *Id.*

89. *Id.*

constitutes "excessive" for the purposes of bail varies according to demographics, and the effect that structural inequity eradicates any chance of fairness. It has been argued that it is cruel, arbitrary and ironic to insist, as a matter of law, that a defendant be required to post a cash bail amount that is higher than what is reasonably likely to ensure the defendant's presence at the trial[90] without a similar requirement that the court satisfy itself that the defendant is able to reasonably afford or obtain that amount. Thus, it is manifest that the current bail system (which is driven almost entirely by unstructured human judgements) is to a large degree dysfunctional and operates in a suboptimal manner.

For the purposes of this Article, the most important aspect of bail are the two key considerations that inform eligibility. The first is the risk that the defendant will offend during the period of bail. The second is whether the defendant is likely to abscond. In reaching these decisions, courts most commonly use unstructured judgments, uninformed by empirical evidence. As discussed below, there is considerable scope to enhance the integrity and rectitude of this approach.

3. Probation Law and Practice

Parole is the third aspect of the criminal justice system where the likelihood of offending is a consideration. Parole is the process through which offenders who are in prison are released prior to the expiration of their complete prison term. Thus, parole is a post-incarceration order which involves a statutory body, typically known as a Parole Board, releasing an offender into the community. Parole is a common sanction. Currently, there are approximately 875,000 offenders on parole.[91]

Parole orders involve the imposition of certain conditions. The conditions generally come in two main forms: standard conditions and special conditions. These conditions are designed to achieve the principal aims of parole, which include community protection and rehabilitation.[92]

The United States Sentencing Commission recommends parole after any prison sentence of longer than a year.[93] 18 U.S.C. § 4209 sets the mandatory conditions for probation and supervised release. All offenders on parole must observe three standard rules to stay in compliance: they must: (1) refrain from committing a new offense; (2) refrain from illegal

90. Stack v. Boyle, 342 U.S. 1, 5 (1951).

91. *Probation and Parole Systems Marked by High Stakes, Missed Opportunities*, PEW (Sep. 25, 2018), https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2018/09/probation-and-parole-systems-marked-by-high-stakes-missed-opportunities [https://perma.cc/Z85C-8YT6].

92. *Probation and Pretrial Services - Mission*, U.S. CTS, http://www.uscourts.gov/services-forms/probation-and-pretrial-services/probation-and-pretrial-services-mission [https://perma.cc/2ZCR-UCHN] (last visited Feb. 7, 2018).

93. U.S. SENTENCING GUIDELINES MANUAL § 5D1.1 (U.S. SENTENCING COMM'N 2016).

drug possession; and (3) submit to one drug test within fifteen days of release and two subsequent drug tests. Further, offenders who have committed certain sexual offenses must maintain current registration as a sex offender. For other crimes, submission to DNA testing may be required.[94] If a parolee is a first-time domestic violence offender as defined by 18 U.S.C. §3561(b), a court-approved rehabilitation program must be completed.

Federal statutes 18 U.S.C. §§ 3563(b) and 3583(d) allow courts to set additional requirements for a defendant's probation or supervised release.95 Courts are given wide discretion to:

> [M]odify, reduce, or enlarge the conditions of supervised release, at any time prior to the expiration or termination of the term of supervised release, pursuant to the provisions of the Federal Rules of Criminal Procedure relating to modification of probation and the provisions applicable to the initial setting of the terms and conditions of post-release supervision.[96]

Any discretionary conditions imposed must be directly connected to the five statutory factors defined in 18 U.S.C. §3553(a)(2). The factors are: the nature and circumstances of the offense; the history and characteristics of the defendant; deterrence; protection of the public; and providing needed correctional treatment to the defendant.[97] The effect of this is that when a board is assessing parole for an offender, the offender's likelihood of reoffending is a major consideration.

When parole is violated, the available sanctions are set out in 28 U.S.C. §994(a)(3). Sanctions may include a custodial sentence being re-imposed upon the offender. In 2015, a study examined the breach rates over ten years of 454,223 offenders that were serving probation or on some form of supervised release. Within the first year, the number of offenders who had their supervision revoked as a result of committing a second crime was approximately 16.2%.[98] The percentage of offenders increased to 33.7% in a three year time frame, and increased again to 41.1% over the next two years.[99] The chance that offenders might have their order revoked because of a technical violation was 5.9%, 10.8% and 11.2% respectively over the same timeframes.[100] In the three years after supervision, the study found a 15% chance that an offender would be

94. *Id.* § 5D1.3(a)(8).

95. 18 U.S.C. § 3563(c).

96. *Id.*

97. 18 U.S.C. §§ 3553(a)(1), (a)(2)(B)-(D), 3583(d)(1).

98. Laura M. Baber, *Inroads to Reducing Federal Recidivism*, 79 FED. PROBATION3, 6 tbl. 5 (2015), http://www.uscourts.gov/sites/default/files/federal_probation_journal_dec_2015_0.pdf [https://perma.cc/VZB2-E444].

99. *Id*.

100. *Id*. at 6 tbl. 4.

arrested for another offence.[101]

The takeaway from this study is obvious: human-made parole decisions are generally quite poor. The data suggests it is very common for an offender to reoffend subsequent to release.

B. The Current Process for Determining Likelihood of Offending is Highly Inaccurate

As discussed, a defendant's likelihood to commit a future crime strongly influences the three key stages of the criminal justice system. These stages are important because if errors are made in these decisions, adverse consequences flow to either the defendant or the community. If a decision-maker wrongly decides that a defendant will commit an offense, then the defendant is likely to experience unnecessary suffering by being sentenced to prison or a longer term of detention (either pre or post trial). On the other hand, if the decision-maker errs by falsely determining that a defendant will not commit another offense, then the community will suffer as a result of the commission of a crime. Depending on the nature of the offense, this can have catastrophic consequences on members of the community.

Currently, three methodologies are used to forecast offenders' likelihood of recidivism: (1) clinical assessments; (2) actuarial-based assessments; and (3) risk and needs assessments.[102] Clinical assessments are unstructured and involve an evaluator ascertaining an offender's risk of recidivism. Evaluators generally do so by referring to subjective criteria and experience, rather than empirically-validated information.[103] In effect, this is the conventional approach employed by judges in sentencing offenders, and is especially relevant in assessing an offender's rehabilitative prospects and the extent to which offenders threaten community safety. As Reitz observes, "prison sentence lengths in most U.S. jurisdictions are already based on predictions or guesses about offenders' future behavior, and this has been true—in multiple settings—for at least a century".[104]

Unstructured assessments relating to the future risk of offending are not confined to sentencing. This approach has been typically used by

101. *Id*. tbl. 3.

102. As discussed further in this section, the main three methodologies are unstructured clinical assessments; actuarial methodologies; and structured professional judgment assessments. *See* Michael Davis & James R. P. Ogloff, *Key Considerations and Problems in Assessing Risk for Violence*, *in* PSYCHOLOGY AND LAW: BRIDGING THE GAP 191, 195-96 (David Canter & Rita Zukauskiene eds., 2008); Christopher Slobogin, *Risk Assessment*, *in* THE OXFORD HANDBOOK OF SENTENCING AND CORRECTIONS 196, 198-99 (Joan Petersilia & Kevin R. Reitz eds., 2012).

103. Slobogin, *supra* note 102, at 208.

104. Kevin R. Reitz, *"Risk Discretion" at Sentencing*, 30 FED. SENT'G REP. 68, 70 (2017)

parole boards and in bail determinations.[105] Irrespective of the forum in which such assessments are made, they are notoriously inaccurate—in fact, they are barely more accurate than tossing a coin.[106]

We ought not to quarantine the deliberations of judges and other decision-makers within the criminal justice system from our skeptical eye simply on the grounds that they seem to be subject matter experts. We can indeed expect those who make decisions about the application of the law to have knowledge of the law commensurate with their professional status. Yet, when fair and valid applications of the law are contingent upon predictions of recidivism, that contingency tends to negate the value of expertise. This is because the amount of data needed to make accurate predictions is beyond the processing speed of our conscious mind to meaningfully correlate; experts are no more immune than the lay person to the confounding influence of extraneous and irrelevant data,[107] and humans invariably rely on heuristic reasoning to make complex decisions.

Judges are, of course, not the only professionals to attract criticism for their inability to make accurate predictions about the future behavior of those they are tasked with managing. A 2006 metastudy found that up until the end of the twentieth century, the accuracy of psychiatric diagnoses and predictions that a particular patient would go on to develop certain mental illnesses averaged around 50-54%.[108] More recently, machine learning prediction models trained on functional, neuroimaging, and combined baseline data, have been able to outperform psychiatrists in predicting one year outcomes for patients in in clinical high-risk states for psychosis and for patients with recent-onset depression.[109]

What then of the calculation of reoffending? The predominant method uses "risk assessment tools,"[110] which involve actuarial assessments.[111]

---

105. Thomas Mathiesen, *Selective Incapacitation Revisited*, 22 L. & HUM. BEHAV. 455, 458–59 (1998),

106. Mirko Bagaric & Theo Alexander, *The Fallacy that is Incapacitation: an argument for limiting imprisonment only to sex and violent offenders*, 2 COMMONWEALTH CRIM. L. REV. 95 (2012).

107. Brite Englich, Thomas Mussweiler & Fritz Strack, *Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making*, 32 PERSONALITY & SOC. PSYCHOL. BULL. 188–200 (2006).

108. Ahmen Aboraya et al., *The Reliability of Psychiatric Diagnosis Revisited: The Clinician's Guide to Improve the Reliability of Psychiatric Diagnosis*, 3 PSYCHIATRY 41, 43 (2006), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2990547/ [https://perma.cc/5C8A-EEF2].

109. Nikolaos Koutsouleris et al., *Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or with Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis*, 75 JAMA PSYCHIATRY 1156 (2018).

110. *See* Pari McGarraugh, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1091 (2013).

111. *See* Melissa Hamilton, *Back to the Future: The Influence of Criminal History on Risk Assessments*, 20 BERKELEY J. CRIM. L. 75, 91-92 (2015); Michael Tonry, *Legal and Ethical Issues in the Prediction of Recidivism*, 26 FED. SENT'G REP. 167, 171 (2014). Such tools are in fact now used in the majority of states in the United States. *See* Shawn Bushway & Jeffrey Smith, *Sentencing Using Statistical*

This approach evaluates "an individual's chances of endangering public safety by reoffending."[112] Richard Berk and Jordan Hyatt observe:

> Forecasting has been an integral part of the criminal justice system in the United States since its inception. Judges, as well as law enforcement and correctional personnel, have long used projections of relative and absolute risk to help inform their decisions. Assessing the likelihood of future crime is not a new idea, although it has enjoyed a recent resurgence: an increasing number of jurisdictions mandate the explicit consideration of risk at sentencing.[113]

Risk assessment tools specify the event variables that have caused offenders to reoffend in the past,[114] and then develop "rules" about how likely these events are to happen in an offender's future. So, these types of tools are "actuarial instruments [that] manipulate existing data in an empirical way to create rules. These rules combine the more significant factors, assign applicable weights, and create final mechanistic rankings."[115] There are numerous risk assessment tools which are distinguished by the integers they incorporate in their formulas and the weightings accorded to each integer. The most basic of these tools consider the person's criminal history,[116] known associates, personality, and attitudes towards crime.[117] Other tools use more advanced and/or more fluid factors. An example of this is the Post-Conviction Risk Assessment ("PCRA") system, which is used in the federal jurisdiction in probation matters.[118] This tool considers factors relating to the offender's employment history, previous familial circumstances, and level of education.[119]

When compared with an unstructured judgement of the court, risk

---

Treatment Rules: What We Don't Know Can Hurt Us, 23 J. QUANTITATIVE CRIMINOLOGY 377, 378 (2007).

112. Bushway & Smith, *supra* note 111. In addition, actuarial methodologies and other risk assessment approaches include unstructured clinical assessments and structured professional judgment assessments. *See* Davis, *supra* note 102; Slobogin, *supra* note 102, at 198.

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

114. *See* McGarraugh, *supra* note 110, at 1091-92.

115. Hamilton, *supra* note 111, at 92.

116. *Id.* at 90.

117. *Id.*

118. *Id.* at 94.

119. *Id.* Another common similar tool is the Level of Service instrument, which incorporates 54 considerations. *See* Slobogin, *supra* note 102, at 199. In terms of predicting future violence, it has been noted that dynamic measures are slightly more accurate than static measures for short- to medium-term predictions of violence. *See* Chi Meng Chu et al., *The Short- to Medium-term Predictive Accuracy of Static and Dynamic Risk Assessment Measures in a Secure Forensic Hospital*, 20 ASSESSMENT 230, 231 (2011). Given that these tools go beyond the use of static factors and incorporate dynamic factors, they are sometimes referred to as "structured professional judgment tools." Davis & Ogloff, *supra* note 102, at 200.

assessment tools provide a much more accurate outlook on offenders' chance of recommitting crimes. It has been shown that "the best models are usually able to predict recidivism with about seventy percent accuracy—provided it is completed by trained staff."[120] When compared to an unstructured assessment of reoffending, these tools produce significantly higher true positives, between the 50% to 85% ranges.[121] Although risk assessment tools have a high accuracy rate, they have not generally been used effectively within the United States in sentencing matters.[122] The Brennan Center made the following observations about the contribution of risk assessment tools to sentencing determinations in different jurisdictions:

> Driven by advances in social science, states are increasingly turning toward risk assessment tools to help decide how much time people should spend behind bars. These tools use data to predict whether an individual has a sufficiently low likelihood of committing an additional crime to justify a shorter sentence or an alternative to incarceration . . . Some courts have implemented risk assessments to determine whether defendants should be held in jail or released while waiting for trial; similarly, some parole boards use them to decide which prisoners to release. States such as Kentucky and Virginia have implemented the former, while Arkansas and Nevada have implemented the latter. More recently, states are applying risk assessments to guide sentencing decisions. The first state to incorporate such an instrument in sentencing was Virginia in 1994. By 2004, the state implemented risk assessments statewide, requesting judges to consider the results in individual sentencing decisions. Courts in at least 20 states have begun to experiment with using risk assessments in some way during sentencing decisions . . . Because these instruments do not change existing sentencing laws, which the authors believe are a root cause of overly long sentences, this report does not delve further into the use of risk assessment in sentencing.[123]

---

120. Edward J. Latessa & Brian Lovins, *The Role of Offender Risk Assessment: A Policy Maker Guide*, 5 VICTIMS & OFFENDERS 203, 212 (2010). Moreover, generally risk assessment tools are more accurate than predictions based solely on clinical judgment. *See* D.A. Andrews et al., *The Recent Past and Near Future of Risk and/or Need Assessment,* 52 CRIME & DELINQ. 7, 12 (2006); William M. Grove et al., *Clinical Versus Mechanical Prediction: A Meta Analysis*, 12 PSYCHOL. ASSESSMENT 19, 25 (2000).

121. Slobogin, *supra* note 102, at 201.

122. They are most commonly used in Virginia, Missouri and Oregon. *Id.* at 202-03.

123. James Austin, Lauren-Brooke Eisen, James Cullen & Jonathan Frank, *How Many Americans are Unnecessarily Incarcerated?*, BRENNAN CTR. FOR JUSTICE AT NYU SCH. OF LAW 18-19 (2016), https://www.brennancenter.org/sites/default/files/publications/Unnecessarily_Incarcerated_0.pdf [https://perma.cc/G2PH-USE2]. Judges often pay little regard to the results of risk assessment tools. *See also* Slobogin, *supra* note 102, at 202, 207. In Virginia, fifty-nine percent of defendants who were considered to be at low risk of reoffending by a risk assessment tool were still sentenced to prison. Simmons, *supra* note 22, at 966 n.76. *See also* Steven Chanenson & Jordan Hyatt, *The Use of Risk Assessment at Sentencing: Implications for Research and Policy* (Villanova Univ. Charles Widgar School of Law Working Paper Series, 2016),

Risk assessment tools are used specifically to predict whether a person will reoffend or endanger the public in the future.[124] On the other hand, risk-and-needs assessment tools aim to address the offender's needs, so that interventions can be applied and reduce the chance of reoffending.[125] This is different from non-traditional risk assessment tools that use the actuarial base to predict the likelihood of the offender proceeding down the path to reoffending.[126]

Risk-and-needs instruments differ in that they "focus on treatment or rehabilitation of the offender to prevent reoffending, rather than simply predict recidivism. This approach to risk differs importantly from the correctional use of static risk for preventive or selective incapacitation, diversion, or deterrence of recidivism through the administration of harsh penalties."[127]

The Ohio Risk Assessment System ("ORAS') is a commonly used risk-and-needs assessment tool.[128] This tool has a range of sophisticated variables, including the offender's family relationships, academic performance, employment history, community involvement, and history of substance abuse.[129] Risk-and-needs assessment tools are used broadly when making decisions about parole[130] and probation;[131] however, tools of this type have also been used increasingly for sentencing decisions.[132]

Risk assessment has been used more frequently in bail decisions as well. In 2017, the National Council of State Legislatures recorded that "nine states enacted laws allowing or requiring courts to use risk assessments to assist in establishing bail and release conditions [and] another five passed bills directing studies or development of risk

---

https://digitalcommons.law.villanova.edu/cgi/viewcontent.cgi?article=1201&context=wps.

124. McGarraugh, *supra* note 110, at 1091.

125. NATHAN JAMES, CONG. RES. SERV., R44087, RISK AND NEEDS ASSESSMENT IN THE CRIMINAL JUSTICE SYSTEM, FED'N OF AM. SCIENTISTS 4-5 (2018), https://fas.org/sgp/crs/misc/R44087.pdf [https://perma.cc/2H29-N5TR].

126. Slobogin, *supra* note 102, at 199.

127. Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 276 (2013).

128. For an explanation of the manner in which it is used, *see* SUPERIOR COURT WORKING GRP. ON SENTENCING BEST PRACTICES, COMMONWEALTH OF MASS., CRIMINAL SENTENCING IN THE SUPERIOR COURT: BEST PRACTICES FOR INDIVIDUALIZED EVIDENCE-BASED SENTENCING (2019) https://www.mass.gov/doc/criminal-sentencing-in-the-superior-court-best-practices-for-individualized-evidence-based/download [https://perma.cc/B84K-QB7Z].

129. JAMES, *supra* note 125, at 7-8.

130. *See Id.* at 1, 10.

131. PAMELA M. CASEY, ROGER K. WARREN & JENNIFER K. ELEK, NAT'L CTR. FOR STATE COURTS, USING OFFENDER RISK AND NEEDS ASSESSMENT INFORMATION AT SENTENCING: GUIDANCE FOR COURTS FROM A NATIONAL WORKING GROUP, 7, 16-17 (2011), http://www.ncsc.org/~/media/microsites/files/csi/rna%20guide%20final.ashx.

132. *Id.* at 9, 13-15.

assessment tools."[133]

Current risk assessment tools provide a way for those overseeing a post-release offender to plan and execute remedial interventions. These interventions can be focused on individual needs and provide the offender with the best chance of success. Thus, the offender's ability to communicate factors such as living situation, reading and cognitive understanding, and transport options can help create an individualized management plan. This has led to improved chances of success.[134] This approach to risk assessment, one that considers each person's way of learning and thinking ability, comes from a model called the Risk-Needs-Responsivity ("RNR") which uses both managerial and actuarial approach base. This approach also utilizes clinical evidence-based methods and rehabilitation methods. In considering the multiple factors considered in risk assessment, this model is ideally used in tandem with algorithmic coding.[135]

The pre-trial services program, used in New Jersey, compiles a Public Safety Assessment ("PSA") by a utilizing algorithmic risk assessment tools, which combine submissions made by the parties deciding applications for release.[136] Defendants who are released following conditions imposed are monitored by Pre-Trial Services staff. The New Jersey Courts' Report to the Legislature on the implementation of the program, notes that:

> This redefined pre-trial process represents a significant improvement in the criminal justice system. The Judiciary has automated many tasks, including production of the PSA, to facilitate faster and more efficient processing of cases. Utilizing a risk measurement and risk management model, judges have the benefit of specific objective information about a defendant in order to make an informed release or detention decision.[137]

However, compliance rates were not as high as expected and numerous, unanticipated problems were encountered. The evaluation

133. NAT'L CONFERENCE OF STATE LEGISLATURES, TRENDS IN PRETRIAL RELEASE: STATE LEGISLATION UPDATE CIVIL AND CRIMINAL JUSTICE 1 (Apr. 2018), http://www.ncsl.org/portals/1/ImageLibrary/WebImages/Criminal%20Justice/pretrialEnactments_2017_v03.pdf [https://perma.cc/3Y89-PSVK].

134. *See generally* THOMAS H. COHEN, ET AL., THE FEDERAL POST-CONVICTION RISK ASSESSMENT INSTRUMENT: A TOOL FOR PREDICTING RECIDIVISM FOR OFFENDERS ON FEDERAL SUPERVISION IN HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 77 (Jay P. Singh, et al. eds., 1st ed. 2017).

135. Jill Viglione, *The Risk-Need-Responsivity Model: How Do Probation Officers Implement the Principles of Effective Intervention?*, 48 CRIM. JUST. & BEHAV. 655 (May 2019), https://journals.sagepub.com/doi/pdf/10.1177/0093854818807505 [https://perma.cc/K937-4KYX].

136. GLENN A. GRANT, CRIMINAL JUSTICE REFORM REPORT TO THE GOVERNOR AND THE LEGISLATURE FOR CALENDAR YEAR 2017, 11 (2018), https://www.njcourts.gov/courts/assets/criminal/2017cjrannual.pdf [https://perma.cc/E9WY-2TWR].

137. *Id.* at 9.

reported that the compliance monitoring staff faced significant challenges due to the lack of affordable community-based substance abuse treatment, mental health treatment, and housing assistance programs. If the risk-needs assessment recommends release, the relevant statute provides a range of conditions which the court can attach to the release order. These include requiring the defendant to undergo medical, psychological or psychiatric treatment, drug or alcohol treatment, obtain or maintaining employment, and obtain or maintain attendance in an educational program. Compliance staff reported an inadequate supply of available programs to meet the demand. The programs that were available were often unaffordable. Where pro bono services were offered for a particular service, there would typically be waiting lists months long, which meant the pre-trial period would expire before they could be utilized.

C. Other Shortcomings of the Current Approach

As we have seen, the main approach to risk assessment in the criminal justice area continues to be unstructured judgments that rely on human-based decision-making strategies. Thus, this is the principal reference point against which alternative approaches—such as AI—should be measured. It is desirable, therefore, to examine more carefully some of the other drawbacks of this approach; especially because, as we shall see, they are of the same nature as the criticisms levelled against AI risk assessment decision making. In addition to the fact that unstructured decision making is poor at determining reoffending rates, there are numerous other problems associated with this approach. Another serious flaw with human decision-making in the criminal justice context is that the outcome of many decisions is influenced by subconscious bias. Many studies have demonstrated that certain groups in society receive harsher criminal justice outcomes than the wider community. The groups that are disproportionately burdened are the already disadvantaged groups in the community.

Empirical studies have uncovered that offenders from minority groups, especially African-Americans, often receive more severe sentences than white offenders who have committed comparable crimes.[138] Researchers have found that racial bias has contributed to this disparity, thereby undermining the rule of law. As Walker notes, a critical component of the rule of law is "the rules of natural justice," which include "the requirement of an unbiased tribunal."[139]

In a study involving more than 77,000 offenders, researchers found that

138. Rose Matsui Ochi, *Racial Discrimination in Criminal Sentencing*, 24 JUDGES J. 6, 8 (1985).

139. GEOFFREY DE Q. WALKER, THE RULE OF LAW: FOUNDATION OF CONSTITUTIONAL DEMOCRACY 1 (1988).

black offenders were sentenced to terms of incarceration more than 12% longer than white offenders, once other variables were controlled.[140] When applied in the federal jurisdiction, the Federal Sentencing Guidelines were found to have the same level of disparity.[141] This research showed that between 2005 and 2012, offenders who were black received sentences that imposed imprisonment that were 5 to 10% longer than those of white offenders who had committed similar or identical crimes,[142] even when accounting for the variables set out in the guidelines.[143] It was also posited that the Supreme Court's decision in *Booker* has lead to inconsistent determinations when sentencing black and white offenders because of the increased discretion given to judges.[144] The report states:

> We are concerned that racial disparity has increased over time since *Booker*. Perhaps judges, who feel increasingly emancipated from their guidelines restrictions, are improving justice administration by incorporating relevant but previously ignored factors into their sentencing calculus, even if this improvement disadvantages black males as a class. But in a society that sees intentional and unintentional racial bias in many areas of social and economic activity, these trends are a warning sign. It is further distressing that judges disagree about the relative sentences for white and black males because those disagreements cannot be so easily explained by sentencing-relevant factors that vary systematically between black and white males . . . We take the random effect as strong evidence of disparity in the imposition of sentences for white and black males.[145]

Unstructured sentencing by judges has led to a system where decisions are opaque and inconsistent because they are based on a judge's personal predisposition. This has caused certain groups in the community to be sentenced harsher than others. Judges, like most of the community, view themselves as having high standards of fairness and objectivity. But as all

140. Ronald S. Everett & Roger A. Wojtkiewicz, *Difference, Disparity, and Race/Ethnic Bias in Federal Sentencing*, 18 J. QUANTITATIVE CRIMINOLOGY 189, 198 (2002); David S. Abrams, et al., *Do Judges Vary in Their Treatment of Race?*, 41 J. LEGAL STUD. 347, 350 (2012).

141. William Rhodes et al., *Federal Sentencing Disparity: 2005–2012* (Bureau of Justice Statistics Working Paper Series, Paper No. 2015:01, 2015), https://www.bjs.gov/content/pub/pdf/fsd0512.pdf [https://perma.cc/HL89-SQRE] (documenting previous studies in the United States, which support the conclusion that subconscious bias causes racial disparity in sentencing).

142. *Id.* at 41.

143. *Id.* at 23.

144. *Id.* at 66.

145. *Id.* at 68. A more recent study focusing on sentencing patterns in Florida noted that African Americans often received markedly longer prison terms than white offenders for the same offense. *See* Elizabeth Johnson et al., *Black defendants get longer sentences in Treasure coast system*, DAYTONA BEACH NEWS-JOURNAL (Dec. 19, 2016, 1:09 PM), http://www.news-journalonline.com/news/20161218/black-defendants-get-longer-sentences-in-treasure-coast-system [https://perma.cc/J9PM-2P76].

people do, judges have their own biases and ideals, which knowingly or unknowingly affect their decision-making. It has been found that in making decisions, judges have severe problems recognizing inherent bias that exist in the way they make these decisions.[146] The most prevalent of these and hardest for them to recognize are ones they do not even know they have. Judge Richard Posner states in his book How Judges Think that "we use introspection to acquit ourselves of accusations of bias, while using realistic notions of human behavior to identify bias in others."[147] While most people would like to think "their judgments are uncontaminated"[148] with bias, known or unknown, this is clearly not true. The different path through life that every judge takes impacts on how they think and they "are more favorably disposed to the familiar, and fear or become frustrated with the unfamiliar."[149]

Numerous studies have been made of implicit bias within judicial settings. The most explosive of these findings involve race and socio-economic status. For example, offenders who are considered attractive by society receive lenient penalties, except in cases where their attractiveness was used to advance the crime.[150] We also know that race plays a huge part in sentencing: while black judges show no favoritism or preference in their courts, white judges have given less severe sentences to white offenders,[151] and black offenders targeting white victims are given a more severe sentence than in cases where the victim was black.[152] The economic status of those before the court also can affect the outcome of the case. When dealing with child custody, it has been shown that judges will give more preference and show favor to those who are wealthy then those who come from poor backgrounds.[153]

There are also a range of other more subtle factors that have been found

146. Jennifer K. Robbennolt & Matthew Taksin, *Can Judges Determine Their Own Impartiality?*, 41 MONITOR ON PSYCHOL. 24, 24 (2010).

147. RICHARD POSNER, HOW JUDGES THINK 121 (2008).

148. Timothy Wilson et al., *Mental Contamination and the Debiasing Problem*, *in* HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT 185, 190 (Thomas Gilovich et al. eds., 2002).

149. Ochi, *supra* note 138, at 53.

150. Birte Englich, *Heuristic Strategies and Persistent Biases in Sentencing Decisions*, *in* SOCIAL PSYCHOLOGY OF PUNISHMENT OF CRIME 295, 304 (Margit E. Oswald et al. eds., 2009). In one study, seventy-seven percent of unattractive defendants received a prison term, while only forty-six percent of attractive defendants were subjected to the same penalty. *See* John E. Stewart II, *Defendant's Attractiveness as a Factor in the Outcome of Criminal Trials: An Observational Study*, 10 J. APPLIED SOC. PSYCHOL. 348, 354 (1980).

151. Jeffrey J. Rachlinski & Sheri L. Johnson, *Does Unconscious Racial Bias Affect Trial Judges?*, 84 NOTRE DAME L. REV. 1195, 1210 (2009).

152. Mirko Bagaric, *Sentencing: From Vagueness to Arbitrariness: The Need to Abolish the Stain that is the Instinctive Synthesis*, 38 UNIV. OF NEW SOUTH WALES L.J. 76, 107 (2015) [hereinafter Bagaric, *From Vagueness*].

153. Bagaric, *From Vagueness*, *supra* note 152, at 106-07; Michele Benedetto Neitz, *Socioeconomic Bias in the Judiciary,* 61 CLEV. ST. L. REV. 137, 158–60 (2013).

to influence the mindset of judges and their decisions. Thus, it has been noted that judges who think about negative matters, such as their own death, set bail at higher levels than other judges.[154] Another study observed that judges were far more likely to grant parole if the decision was made shortly after they had taken a meal break than prior to doing so.[155] The researchers speculated on the reason for this:

> All repetitive decision-making tasks drain our mental resources. We start suffering from "choice overload" and we start opting for the easiest choice . . . And when it comes to parole hearings, the default choice is to deny the prisoner's request. The more decisions a judge has made, the more drained they are, and the more likely they are to make the default choice. Taking a break replenishes them.[156]

Thus, a judge's preferences play a part when making decisions, and this will not be reduced just through the judges' own free will. It was suggested correctly by Posner that, as with all members of the community, judges are utility maximizers, and they will gain a sense of completion and prestige from fulfilling their role in society.[157] The decisions made by judges are influenced and shaped by their biases and preferences. These, in turn, will be guided by "background, temperament, training, experience, and ideology, which shape [their] preconceptions and thus [their] response to arguments and evidence."[158]

It is completely human of judges to want their decisions to be familiar with their own understanding of what they believed to be just and correct, in line with their own life experience. But the sentencing system and any other judicial system that uses an unstructured assessment has great impact upon the well-being of offenders, their victims, and the community as a whole. In this domain, there is no place for subjective judgments.[159]

Thus, a significant problem with human decision-making in the areas of sentencing, bail and parole is that the subconscious sentiments of the decision-makers lead to inconsistent outcomes. These differences often operate disproportionately against already disadvantaged sectors of the

154. Bagaric, *From Vagueness*, *supra* note 152, at 107; Abram Rosenblatt et al., *Evidence for Terror Management Theory: I. The Effects of Mortality Salience on Reactions to Those Who Violate or Uphold Cultural Values*, 57 J. PERSONALITY & SOC. PSYCHOL. 681, 682 (1980).

155. Shai Danzinger et al., *Extraneous Factors in Judicial Decisions*, 108 PROC. NAT'L ACAD. SCI. 6889, 6889–90 (2011).

156. Bagaric, *From Vagueness*, *supra* note 152 at 107-08; Ed Yong, *Justice is Served, but More so After Lunch: How Food-breaks Sway the Decisions of Judges*, DISCOVERMAG.COM (Apr. 11, 2011, 3:00 PM), https://www.discovermagazine.com/the-sciences/justice-is-served-but-more-so-after-lunch-how-food-breaks-sway-the-decisions-of-judges [https://perma.cc/AJT6-Y48F].

157. POSNER, *supra* note 147, at 35–36.

158. *Id.* at 249.

159. Bagaric, *From Vagueness*, *supra* note 152, at 110–11.

community.

IV. CURRENT USE OF ALGORITHMS AND AI IN THE CRIMINAL JUSTICE SYSTEM

As discussed in Part II of this Article, there is considerable hesitance to using algorithms in areas involving numerous variables and where traditionally judgments have been made the human. This has contributed to the slow adoption of artificial intelligence in the criminal justice system. The ongoing criticism of algorithmic decision-making in the areas of sentencing and bail determination are particularly acute. These criticisms derive their sting from the fundamental inequity of the biases (both perceived and actual) in design and application, leveraged against the gravity of the consequences of error—that is, an unjustified deprivation of liberty. If not resolved, these concerns have the capacity to erode the value and degree of adoption of algorithms which promise great results. Careful, conceptually robust and evidence-based responses to these criticisms are therefore essential if the trust and confidence of users and stakeholders is to be earned. It is also necessary to acknowledge and incorporate elements of critique so that as AI becomes more ubiquitous on the criminal justice domain, and is not seen as a panacea or simplistic cost saving measure.

One common criticism of AI is that algorithms used to predict recidivism may discriminate against offenders with immutable traits and entrench racism in decision-making about sentences.[160] An offender's race is not an explicit consideration in risk assessment tools or sentencing law generally.[161] Nevertheless, due to the fact that more African-Americans have prior convictions than white Americans, the inclusion of prior criminality as a consideration in risk assessment tools and as an aggravating factor in sentencing determinations can have the effect of discriminating against African-American offenders.[162] Racial bias leads to over policing of African-American neighborhoods, and criminal activity then becomes more visible. This generates data which is likely to

160. *See infra* Part III (discussing the use of risk assessment tools). *See, e.g.*, *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/X8F2-SMZD]; *See also* Laurel Eckhouse, Opinion, *Big Data May be Reinforcing Racial Bias In the Criminal Justice System*, WASH. POST (Feb. 10, 2017), https://www.washingtonpost.com/opinions/big-data-may-be-reinforcing-racial-bias-in-the-criminal-justice-system/2017/02/10/d63de518-ee3a-11e6-9973-c5efb7ccfb0d_story.html?utm_term=.6a19034da71a.

161. United States v. Taveras, 585 F. Supp. 2d 327, 336 (E.D.N.Y. 2008).

162. Mirko Bagaric, *Three Things that a Baseline Study Shows Don't Cause Indigenous Over-Imprisonment: Three Things that Might but Shouldn't and Three Reforms that Will Reduce Indigenous Over-Imprisonment*, 32 HARV. J. ON RACIAL AND ETHNIC JUST. 103, 107 (2016).

influence future concentrations in patrol deployment policing algorithms,

In fact, research suggests that one risk and assessment tool can incorporate race lacks accuracy. Dressel and Farid undertook a study investigating the tensions around bias for the COMPAS system used by courts in various jurisdictions within the United States for bail risk assessment.[163] The study found that the popular risk-assessment tool COMPAS was as accurate as an online poll of random people with no criminal or legal training. This finding was alarming to the researchers, especially when considering the weight that the courts may place on decisions made by the algorithm. The study analyzed COMPAS's predictions on recidivism for approximately 7,000 defendants in a number of U.S. states and found that inherent bias had crept into the algorithm. The algorithm incorrectly categorized a number of black defendants as high-risk.[164] To determine whether the algorithm improved on human predictions of recidivism, the researchers designed an experiment to test their theories. They used Amazon Mechanical Turk and recruited about 400 participants to predict recidivism using a sample of 1,000 real defendants. They used seven data points for the experiment (whereas COMPAS uses 137 data points via its defendant questionnaire). Interestingly, the researchers found that the untrained participants were roughly as accurate in their predictions as the COMPAS algorithm with a 67% accuracy as opposed to a 65% COMPAS accuracy.

The bias indicates that certain data can act as proxies for racial data even when race is not specifically considered as a data point. The researchers undertook the same experiment with another 400 participants with results of similar accuracy.[165] This is a structural problem with the design of some algorithms, but it is not a necessary aspect of all AI systems.[166]

Algorithms are designed to discriminate or discern information. They are not privy to what is socially acceptable.[167] Things that are considered protected characteristics, such as gender, race, pregnancy status, religion, sexuality and disability all play a part in human decision-making

163. Julia Dressel and Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism* 4 SCI. ADVANCES 1 (2008), https://advances.sciencemag.org/content/4/1/eaao5580.full [https://perma.cc/3UMC-LHQA].

164. Issie Lapowsky, *Crime-Predicting Algorithms May Not Fare Much Better Than Untrained Humans*, WIRED (Jan. 17, 2018, 2:16 PM), https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans/ [https://perma.cc/9MAT-CHXN].

165. *Id.*

166. *See generally* Sandra Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Lauren Eckhouse, Kristian Lum, Cynthia Conti-Cook & Julie Ciccolini, *Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment*, 46 CRIM. JUS. AND BEHAV. 185 (2018).

167. Lilian Edwards and Michael Veale, *Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18 (2017).

processes. This suggests that algorithms which use "past biased data" are likely to recreate the same biases in decision-making processes which exacerbate discrimination and unfairness.[168]

It is, however, possible to engineer AI systems that do not consider the aforementioned traits, thus minimizing the chance for indirect discrimination. Standard statistical and big data methods allow us to see which features are proxies for race or other protected characteristics. These can then be controlled, removing bias from the system as it is developed.

The PCRA demonstrates the capacity to develop algorithms that do not incorporate biases.[169] Jennifer Skeem and Christopher Lowenkamp conducted a study of the PCRA risk assessment process concerning the probation of 34,794 offenders.[170] In the federal system, risk assessment is not used when dealing with sentencing matters, so this was not examined by Skeem and Lowenkamp.[171] In addition to finding that the PCRA was accurate in more than 70% of cases,172 the authors discovered the following:

> First, there is little evidence of test bias for the PCRA. The instrument strongly predicts re-arrest for both Black and White offenders. Regardless of group membership, a PCRA score has essentially the same meaning, i.e., same probability of recidivism. So the PCRA is informative, with respect to utilitarian and crime control goals of sentencing. Second, Black offenders tend to obtain higher scores on the PCRA than White offenders (d= .34; 13.5% nonoverlap). So some applications of the PCRA might create disparate impact—which is defined by moral rather than empirical criteria. Third, most (66%) of the racial difference in PCRA scores is attributable to criminal history—which strongly predicts recidivism for both groups, is embedded in current sentencing guidelines, and has been shown to contribute to disparities in incarceration (Frase et al., 2015). Finally, criminal history is not a proxy for race. Instead, criminal history partially mediates the weak relationship between race and a future violent arrest.[173]

These data methods can deal with other problematic features. Slobogin suggests, for example, that increasing punishments based on immutable,

168. *Id.* at 28.

169. Jennifer Skeem & Christopher T. Lowenkamp, *Risk, Race, & Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 700 (2016).

170. *Id.* at 680.

171. *Id.* at 686.

172. *Id.* at 691.

173. *Id.* at 700. *See also* Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not that Clear*., WASH. POST (Oct. 17, 2016) https://www.washingtonpost.com/news/monkeycage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

non-behavioral features of an offender is profoundly unfair. Features like gender or age should not affect the outcome of a case.[174] There must be an understanding of how these immutable traits work in a sentencing algorithm and why they work in this manner. It is necessary to understand this so that these traits do not unfairly affect the outcome. This, too, has been noted by Slobogin:

> The Supreme Court, however, does not believe that risk assessment is antithetical to criminal justice. It has even approved death sentences based on dangerousness determinations (*Jurek v. Texas* 1976, 275–276). If sentences can be enhanced in response to risk, then neither society's nor the offender's interests are advanced by prohibiting consideration of factors that might aggravate or mitigate that risk simply because they consist of immutable characteristics. In any event, risk-based sentences are ultimately based on a prediction of what a person will do, not what he is; immutable risk factors are merely evidence of future conduct, in the same way that various pieces of circumstantial evidence are not blameworthy in themselves.[175]

The first case at a state appellate level to consider the appropriateness of risk- and needs-assessment in sentencing matters stated that it was not discriminatory for a judge to use a risk assessment tool that considered one of these immutable traits.[176] The court reasoned that all sentencing law:

> mandates that pre-sentence investigation reports include "the convicted person's history of delinquency or criminality, social history, employment history, family situation, economic status, education, and personal habits." Furthermore, supporting research convincingly shows that offender risk assessment instruments, which are substantially based on such personal and sociological data, are effective in predicting the risk of recidivism and the amenability to rehabilitative treatment.[177]

Nonetheless, when risk-assessment and other systems are used to calculate the chance of reoffending, it is important that all factors used in this consideration are expressly identified. This can prevent unwanted factors, such as social and economic background, from being inappropriately used and becoming intertwined with the immutable traits. This means that unless evidence is provided that these factors are relevant to the sentencing process, they should not be used in the calculation of sentencing decisions. When using a computer sentencing system, it would be possible to ensure that all the irrelevant factors are discarded, and the process followed without deviation. This would prevent socio-economic

174. Slobogin, *supra* note 102, at 204-05.

175. *Id.* at 205.

176. Malenchik v. Indiana, 928 N.E.2d 564, 573 (Ind. 2010).

177. *Id.* at 574.

disadvantaged offenders or offenders of different races from suffering harsher penalties. This process can be achieved with computers far more easily and effectively than when relying on human judgment.

In 2019, the Centre for Court Innovation studied a sample of arrests made in New York City of white, black and Hispanic peoples in 2015.[178] This study was made up of 86,227 black (49%), 64,109 Hispanic (36%), and 25,117 white (14%) defendants, totaling 175,000 defendants. The researchers used this sample in a custom-made risk assessment tool that did not consider any express mention of ethnicity; it only considered criminal history and demographic factors that had strong correlation to future arrest. Using this sample, the researchers created nine risk factors, broken into three categories, that were used to predict the chance of a new arrest of current defender. These categories were criminal history, current case characteristics and demographic characteristics. The criminal history category considered an offender's prior convictions, failure to appear before the court, probation status and prior sentences. The case characteristics that were considered were the nature and number of charges that the defendant had pending. The key demographic considerations were the defendant's age and gender. When the risk assessment tool was structured in this way, the study found that, irrespective of the defendant's race or ethnicity, the tool could accurately make predictions on who would be arrested, concluding that "re-arrest rates increased progressively, in near-lockstep, as risk categories move from minimal to high."[179]

This study also noted that most existing risk-assessment tools weighed criminal history too heavily, especially the factors of previous arrests or current warrants. Black defendants, in this case, have a severe disadvantage concerning sentencing. There are more black defendants in the criminal justice system that can be affected by these outcomes. Additionally, for the aforementioned reasons, the systems are also highly likely to consider black defendants at a disproportionately high risk of reoffending. This study also considered the pre-trial detention rate and the number of false positives (those not re-arrested) based on race/ethnicity and then compared the effects of different decision-making systems on these two considerations. The "business as usual" system, where detention is determined subjectively by the judges, as is in New York City, was the first tested. By following that system of all defendants that appeared, 26% of them were detained. Out of these defendants, 22% of

178. SARAH PICARD, MATT WATKINS, MICHAEL REMPEL & ASHMINI KERODAL, CTR. FOR COURT INNOVATION, BEYOND THE ALGORITHM PRETRIAL REFORM, RISK ASSESSMENT, AND RACIAL FAIRNESS (2019), https://www.courtinnovation.org/publications/beyond-algorithm [https://perma.cc/EGJ6-ZHAV].

179. *Id.* at 6.

them were white, 25% Hispanic and 31% black.[180] The second system used the risk-assessment algorithm discussed above. When this was used, the amount of false positives that occurred decreased by 10% (of these, 22% were black, 16% Hispanic and 10% white). Additionally, the total percentage of people that were detained decreased by 9%.[181] The final system that was used contained both a risk-assessment and a restriction on detention, where only severe cases would be detained, to create a "hybrid." This restriction meant that only those that were a violent felony, or a domestic violence case would fall into the moderate to high risk areas. Because of the use of this hybrid system, there was a 51% decrease into pre-trial detention, and when dealing with false positives there was basically no racial disparity, with 13% black, 14% Hispanic and 13% white.[182]

In 2015, the Office for Civil Rights within the U.S. Department of Justice investigated a complaint from Equal Justice Under Law (a non-profit civil-rights advocacy organization). The complaint alleged that the pre-trial bail decision making process used by judges in Davidson County Tennessee (Twentieth Judicial District) impermissibly discriminated against African Americans.[183] The complaint argued that requiring defendants to post money bail as a pre-trial condition of release unfairly discriminated against African-Americans, as they were disproportionately detained in jail prior to trial.[184] This complaint led to an investigation on why decision-makers were significantly more likely to deny bail opportunities to African Americans in the jurisdiction. The investigation found that at the time the complaint arose, judicial officers assigned weight to the relevant statutory criteria for bail decisions. They did not make use of a fixed schedule to determine the amount of secured cash bail that would be payable, but virtually all defendants were required to post some secured bail. Then, in April of 2018, the jurisdiction decided to adopt an algorithmic risk assessment tool[185] as a mechanism of consistency. The jurisdiction used case data from the County to identify risk factors statistically correlated with likelihood of re-arrest or failure to appear. These factors were then tested to see if they accurately predicted pre-trial outcomes. A retrospective analysis was also carried out to test

180. *Id.* at 11.

181. *Id.*

182. *Id.* at 12.

183. Chrysse Haynes, *Press Release: Nashville and Davidson County Tennessee Take First Steps to Reform Money Bail*, EQUAL JUSTICE UNDER LAW (Aug. 17, 2018), https://equaljusticeunderlaw.org/thejusticereport/2018/8/24/press-release-nashville-and-davidson-county-tennessee-take-first-steps-to-reform-money-bail [https://perma.cc/F57K-CNV5].

184. At that time the average amount bail levied for misdemeanors in Davidson County was in excess of $5,000.

185. Developed in conjunction with the Crime and Justice Institute (CJI).

the risk factors against a cohort of African-American defendants. The results showed that each factor accurately predicted offending while on bail and failing to appear, both individually and collectively, for that cohort. Compared to the overall offender group, these identified risk factors "also did not have a statistically significant disparate impact on African Americans."[186]

The real benefit of predictive algorithms in decisions about bail is in predicting risk of flight or of offending while on bail. Risk-assessment is a critical factor for a bail court to consider, but it is not the only consideration. A risk assessment algorithm, therefore, is not a panacea for over-incarceration; its mere availability does not guarantee that users will include it within an overall approach to bail which guarantees equitable treatment of defendants. Users can still take a "set and forget" or "plug and play" solution to risk assessment. Koepke, Logan and Robinson observe that:

> Pre-trial risk assessment instruments, as they are currently built and used, cannot safely be assumed to support reformist goals of reducing incarceration and addressing racial and poverty-based inequities. [S]takeholders who share those goals are best off focusing their reformist energies on other steps that can more directly promote decarceral changes and greater equity in pre-trial justice.[187]

Algorithmic tools for bail can be expected to more accurately predict risk than judicial intuition alone. These tools need to be continually validated against local data, scaffolded on properly resourced data infrastructure, and be to be used to augment rather than replace human decision-making. In its closure letter relating to the Davidson County investigation, the Office for Civil Rights ("OCR") made the following four recommendations regarding their pre-trial release program, which illustrate those requirements:

- Collect and analyze data on race, national origin, and sex for all individuals eligible for pretrial release, including those detained and those released;
- Monitor concurrence rates between judicial decisions and the terms of release recommended by the risk-assessment tool and any associated decision-making framework;
- Document the reasons for overriding the risk-assessment tool's

186. Courts in the jurisdiction are also provided with matrices for each misdemeanor and felony, with one axis plotting the particular offender's risk level for reoffending, and the other axis their risk for failure to appear at future court dates. So, for example, a matrix plotting an offender with low risk for rearrest and high likelihood of turning up to future court appearances, the tool recommends that the decision maker grant bail without a cash surety and text-messaging reminders about court appearances.

187. John Logan Koepka & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725 (2018).

recommendations and analyze any trends that could contribute to systemic bias;
- Measure concurrence rates between the outcomes predicted by the risk-assessment tool and actual outcomes for the pretrial population.[188]

Even if the risk assessment tool and pre-trial release process is purportedly fair in terms of demographics, this does not guarantee that those who are offered bail with a secured cash surety will be able to afford it. In Davidson County, the OCR found existing research, with control for pretrial assessment levels, had established that a defendant released on bail without secured cash bail was no less likely to reappear or to be a greater safety risk than who was required to post a secured bond.[189]

V. INJECTING CONFIDENCE IN ALGORITHMIC DECISION-MAKING: REBUILDING TRUST/CONTROL AND REFORM RECOMMENDATIONS

Algorithms are better at assessing whether people will commit criminal offenses than judges, so long as they are properly coded. This reality is not sufficient to cause a meaningful shift from unstructured assessments of risk to those driven by AI. To achieve this outcome, it is necessary to understand the reluctance towards algorithms and then suggest a pathway for overcoming the difficulties.

Researchers today are investigating ways to "increase trust in automation advice,"[190] including "providing confidence intervals"[191] or allowing people to "slightly modify automation forecasts."[192] As noted above and similarly seen throughout the recent literature, people are much more inclined to trust and maintain confidence in an algorithm when they have, or believe they have, some level of control over the outcome.[193] They are also more likely to trust and use an algorithm when they have seen how it works and how well it determines correct outcomes.[194]

188. Equal Just. Under L. v. Metro. Gov't of Nashville & Davidson Cty. & Twentieth Jud. Dist. of Tenn. (15-OCR-970) Closure Letter from Office for Civil Rights to Mayor Briley and Judge Binkley (July 30, 2018), https://static1.squarespace.com/static/5aabd27d96e76f3205f18a55/t/5b80552770a6ad58d02d7c43/1535137065465/15-OCR-970+Davidson+County+Closure+Final.pdf

189. Michael R. Jones, *Unsecured Bonds: The As Effective and Most Efficient Pretrial Release Option*, PRETRIAL JUSTICE INSTITUTE (2013), https://pdfs.semanticscholar.org/5444/7711f036e000af0f177e176584b7aa7532f7.pdf.

190. Dietvorst, Simmons, & Massey, *Algorithm Aversion*, *supra* note 36.

191. *Id.*

192. *Id.*

193. Binns et al, *supra* note 42; Dietvorst, *supra* note 36; Dietvorst, Simmons, & Massey, *Algorithm Aversion*, *supra* note 36; Dietvorst, Simmons, & Massey, *Overcoming Alogirthm Aversion*, *supra* note 36; Prahl and Van Swol,*supra* note 39.

194. Dietvorst, Simmons, & Massey, *Algorithm Aversion*, *supra* note 36; Dietvorst, Simmons, &

Another avenue for building control and trust is the opacity and transparency of algorithmic decision-making processes. Legislation such as Europe's General Data Protection Regulation ("GDPR") regulates the right to an explanation for a decision made by an algorithm. Although new, and as some scholars note "restrictive, unclear and paradoxical,"[195] this is a step in the right direction for increasing trust in algorithmic decisions. In terms of public decision-making, this is similar to an explanation of rights made under the Freedom of Information Act, where transparency is seen "as one of the bastions of democracy, liberal government, accountability and restrain on arbitrary or self-interested exercises of power."[196] However as Edwards and Veale note, "the apparatus of accountability of private decision-making"[197] is less than transparent due to commercial and trade secrets and protection of IP rights.[198] Transparency and accountability are important in the use of algorithmic decision-making especially where it may have an adverse effect on an individual. Edwards and Veale state that transparency rights "remain intimately linked to the ideal of effective control of algorithmic decision-making."[199] Furthermore, social values such as "human dignity," "information accountability," and "autonomy and respect" all play a part in how society views decision-making processes.

So how does one achieve transparency and accountability without breaching privacy or IP rights? Kroll et al. argue that disclosing the source code is not the solution and may "create harms of its own."[200] It has been suggested that disclosing code may even lead to "gaming" the system, where people attempt to subvert the algorithms efficiency and fairness. The authors argue that accountability can be achieved by auditing and looking to the external inputs and outputs of the process of the decision instead.[201] The algorithm is not the important aspect here, it is the data. Access to the data provides the necessary explanatory information to ensure an absence of bias.

Institutional biases such as racism, sanism, ableism and sexism are often (and perhaps even in the majority of manifestations) intersectional. So, another dimension to the debate about effectiveness and fairness in the use of algorithmic prediction needs to take place at a higher level of abstraction. It must also be a debate which attempts to shine a spotlight

Massey, *Overcoming Algorithm Aversion*, *supra* note 36; Prahl and Van Swol, *supra* note 39.

195. Edwards & Veale, *supra* note 167 at 18.

196. *Id.* at 39.

197. *Id.*

198. *Id.*

199. *Id.* at 41.

200. *Joshua A Kroll* et al, Accountable Algorithms 165 U. Pa. L. Rev. 633 (2016).

201. *Id.*at 641.

on potential ideological differences in how we view the nature of crime detection and mitigation in a polity which is increasingly data driven. At the most fundamental level, we can use the analogy of algorithmic content filtering on the internet. DNS based or search engine filters may be utilized to block access to content which either the accessing party, the content supplier or a third party (such as government) finds objectionable. These content filters are invariably inexact, as the not all of them have input data and code which is sufficiently granular to prevent some over or under blocking. For example, an algorithm which blocks sites containing the word "breast' notoriously prevented access in some libraries to education to material about detection and treatments for breast cancer.[202] How a filtering algorithm classifies a web page may also depend on the ideology (either explicit or implicit) of the designer or sponsor. For example, "one person's women's health page is another person's pro-abortion" page.[203]

There are those who suggest, for example, that we need to reconceptualize the process in which police intercept citizens to investigate actual, perceived or suspected offending as "programs" rather an as discrete, isolated events. Goel and Perelman et al. propose that this needs to be part of an evolving approach in which "the judiciary will need to grow more comfortable with statistical proof of discriminatory policing, and the police will need to be more receptive to the assistance that algorithms can provide in reducing bias."[204]

Thus, there is considerable debate regarding the desirability and utility of using algorithms to predict future offending in the criminal justice system. The solution to this complex issue depends in a large part on recognizing several incontestable aspects relating to the workings of algorithms. The most important threshold reality is that in contrast to humans, computers have no instinctive or unconscious bias. Additionally, they are incapable of inadvertent discrimination and are uninfluenced by extraneous considerations, assumptions and generalizations that are not embedded in their programs. They operate simply by applying variables that have been pre-programmed.

Bias can infiltrate computerized predictive programs only if an algorithm incorporates existing variables or encodes features that result in disproportionately harsh outcomes on offenders from minority groups. Consequently, for computerized sentencing to eliminate bias from criminal justice decisions, the algorithm and the data must be free of the

202. Lizabeth Elaine Stem, Censorship: Filtering Content on the Web, 64 SE. LIBR. 17, 17 (2017).

203. Lori Ayre & Jim Craner, *Algorithms: Avoiding the Implementation of Institutional Biases*, PUB. LIBR. Q. 341, 343 (2018) (internal citations omitted).

204. Sharad Goel et al., *Combatting Police Discrimination in the Age of Big Data*, 20 NEW CRIM. L. REV. 181 (2017).

discrimination that permeates the present sentencing regime. Programs and algorithms need to be designed so that they do not include any integers that contain implicit bias. Once the programs and algorithms have been developed, there would be no scope for extraneous, racial considerations to have an impact on computerized sentencing decisions. As long as the data and the algorithm are transparent, then we can ensure greater consistency and fairness in judicial decision-making and can eradicate discrimination.

Algorithms do not tire and the analytical routines they generate can run endlessly with no reduction in performance.[205] In fact, machine learning exploits the power of repetition and the ability to run its own hypothetical tests on evolving data sets to learn from its own mistakes. The simplistic, but common, criticism that "using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past"[206] relies on an equally simplistic conception of how modern predictive modelling algorithms function. This criticism is also generally misplaced because the weakness it seeks to focus on is due more to the use of inadequate data sets, failure to recognize existing biases in the data used, improper matching of algorithm to task or other problems with design and application (none of which ought to be insurmountable) rather than the process itself. By contrast, human decision making is notoriously susceptible to fatigue depletion. A decision maker such as a judge, who is called upon to make multiple high stakes discretionary determinations in relatively short periods of time, almost always invariably begins to make less rational decisions as fatigue begins to set in.

According to Danziger et al., "Prior research suggests that making repeated judgments or decisions depletes individuals' executive function and mental resources, which can, in turn, influence their subsequent decisions."[207] In the Danziger study, the authors' hypothesis was that as judges' work through a list of parole applications on a given day, the order of which was determined by someone else, that they would be more likely to make decisions which conformed with the default position—that is, to deny the applications. Noting the existing neurological literature which suggested that decline in executive function caused by fatigue may be alleviated by remedial strategies such as taking short rests, increasing blood glucose levels by snacking or by engaging in mindfulness exercises, they set out to analyze the patterns of ruling favorably in relation to the times at which the rulings were given, and when they

205. *Id.*

206. Karen Hao, *AI is Sending People to Jail – and Getting it Wrong*, TECH. REV. (Jan. 21, 2019), https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/.

207. Shai Danziger et al., *Extraneous Factors in Judicial Decisions*, 108 PROC. OF THE NAT'L ACAD. OF SCI. 6889 (2011).

occurred relative to scheduled breaks and meal times. They found that "the likelihood of a favorable ruling is greater at the very beginning of the workday, or after a food break than later in the sequence of cases."[208]

In applying algorithms to predict the likelihood of offending, researchers need to have access to the largest possible datasets and have access to the outcomes from existing widely used risk and needs assessment tools in order that they can evaluate the variables which are most relevant to accurately predicting recidivism. These can be used as a starting point in developing more accurate algorithms, which are nuanced to the particular offender and offense profiles. This type of approach has been successfully applied in a manual manner in the New Zealand Department of Corrections study on the risk of imprisonment and risk of reconviction.[209] In that study, the researchers used a very large dataset collected by the NZ Department of Corrections, detailing the criminal lives of 133,000 offenders. They were able to demonstrate a statistical model that predicted the probabilities of offenders re-offending and their likelihood of going to prison for the offense. This kind of dataset and model can be easily used by machine learning systems to generate meaningful outputs that can provide immediate guidance in sentencing, bail and parole decisions.

A key component of a fair offense predictive model is transparency. As we have seen, commercial interests often preclude the dissemination of the coding used to develop and run algorithms. In the context of the criminal justice system, however, commerciality cannot be used a basis for limiting full transparency. The criminal justice system is the forum where society, through its courts, acts in its most coercive manner against individuals. It is a public and democratic demonstration and utilization of power and results in the deliberate infliction of suffering against offenders. The commercial interests of individuals or corporations cannot undercut the public nature of the criminal justice system. The integrity of the system commands total transparency. Thus, algorithms which are used to determine future offending should be developed by public institutions. Alternatively, if they are developed by the private sector and adopted by the criminal justice system, then the government must purchase all legal and commercial interests in the programs in order that

208. The authors controlled for potentially confounding factors in a number of ways. For example, the possibility that offenders who had not completed rehabilitation programs while in custody were more likely to appear before breaks or earlier in the lists, was unlikely given that judge both determines the timing of breaks and has no knowledge of the content of cases on the daily list. Furthermore, the position in which a case is actually heard in the list is virtually always determined by time at which the prisoner's attorney arrives at the court. *Id.* at 6890.

209. *Risk of Reconviction: Statistical Models which Predict Four Types of Re-Offending*, N.Z. DEP'T OF CORRECTIONS (1999), https://www.corrections.govt.nz/resources/research_and_statistics/risk-of-reconviction [https://perma.cc/A6KG-S3BK].

their workings can be made public. And, as we have seen, transparency is the key requirement for negating or curtailing algorithmic aversion. Moreover, the splendor of this approach is that it will facilitate the testing of the algorithms and provide scope for their continual evaluation, refinement, and improvement.

There is also a profound benefit that would stem from developing an accurate offense predicting algorithm that is not sufficiently underscored by the literature. The main reason in favor of developing such a system is to enhance community safety. But the corollary of injecting greater accuracy in relation to such decisions is that it would greatly enhance the plight and well-being of thousands of offenders, who as a consequence of the current flawed approach to predicting re-offending are wrongly assessed as presenting a risk of offending. These "false positives" result in individuals either being sentenced to unnecessary terms of imprisonment or longer than appropriate terms. The avoidance of this gratuitous suffering presents a powerful incentive to improve the rectitude of decision-making in this area. This reality provides another strong reason for pursuing the recommendations in this Article.

CONCLUSION

People have an innate aversion to human judgements being supplanted by artificial intelligence processes. There are several reasons for this, including an instinctive belief that people are more likely to make more accurate decisions than machines, especially when the matter involves a large number of complex and nuanced variables. This aversion has shown to be unjustified. Properly designed machine processes are more accurate and efficient in making decisions than humans in many fields. Despite this, the aversion to algorithms continues. The bias against machine learning is one reason that there has been a slow and patchy uptake of computer facilitated decision-making in the criminal justice area. This is despite the fact that ostensibly this field is a fertile area for the use of algorithms. The key consideration that informs sentencing, bail, and parole decisions is whether the offender will reoffend, and it has been shown that properly designed algorithms are better at making these assessments than judges or other criminal justice officials.

The bias against algorithm usage in the criminal justice system is especially paradoxical given that the most forceful criticism against it relates to matters that research shows are key failings of the current (human) decision-making process. Thus, we see that human decision-making has resulted in poor people and those from some racial minorities being disproportionality adversely affected in sentencing and bail determinations. While there is some evidence of certain algorithms also

computing decisions which suggest racial bias, the significance of algorithms is that they operate in a binary manner. Computers do not have actual or subconscious biases. The simply provide answers which are driven by the code that is given to them. If the code has no inappropriate variables, then appropriate answers will follow. Thus, flaws of this nature are, at least in theory, readily fixable. The key to this solution is to identify express and implied sources of biases in the integers that drive the algorithms. Another important requirement is for the coding to be transparent. As we have seen, this will assist in overcoming the innate distrust of algorithms and provide the vehicle for ongoing testing, refinement and improvement of the algorithms.

The key to reform in this area is to improve the rectitude of the decision-making. The one undeniable advantage of computer decision-making over human process is efficiency. Computers can synthesize thousands of variables almost instantaneously. By contrast, the same processing can take humans weeks. Yet, efficiency gains alone are not enough to encourage the greater use of criminal justice algorithms.

Thus, the design part of the algorithmic process is essential. There are numerous algorithms which are currently in use in the criminal justice process. Their greatest use is in the context of parole. On balance, they have been shown to be more accurate than decisions made by people. However, this evidence is not unwavering, and some algorithms are compromised in their accuracy. It is important to understand where these failings have occurred to ensure that they are negated in other models.

Another important consideration is the manner in which algorithms are deployed. At the one extreme they can be used as a substitute for human decision-making. At other end, they can be used as an optional aid by judges and administrators in the criminal justice system. Their ideal position can only be ascertained from a calculus the takes into the account their accuracy, efficiency and receptivity, and confidence in these instruments. Ideally, judges should not be able to simply ignore their calibrations. New systems should be integrated into existing systems methodically and slowly to ensure appropriate acceptance and not to undermine confidence in the criminal justice process. Thus, we recommend that at least initially they serve as aids to the human decision-making in this area, whereby decision-makers are not required to give reasons when they do not adopt the algorithmic decision. This, in time, can evolve to a situation, where judges are required to adopt computer decisions unless reasons are provided for not rejecting the outcomes of AI decision-making. The reforms in this Article will improve transparency, coherency and the accuracy of decisions regarding whether individuals will commit offenses in the future. This will make society safer and reduce the criminal justice burden on countless offenders who are now

wrongly evaluated as being likely to commit criminal offences in the future.