ARTIFICIAL INTELLIGENCE, DUE PROCESS, AND CRIMINAL SENTENCING

John Villasenor[*] *& Virginia Foggo*[**]

2020 MICH. ST. L. REV. 295

TABLE OF CONTENTS

[*] Professor of engineering, law, and public policy, UCLA; co-director of the UCLA Institute for Technology, Law, and Policy; non-resident senior fellow, the Brookings Institution. The authors thank Steven Goode, Aziz Huq, Richard Re, and Alec Walen for providing feedback and/or information in relation to this Article. The opinions expressed herein are the authors' own.

[**] Research Associate, UCLA School of Law.

INTRODUCTION

Artificial intelligence (AI) is rightly viewed as a transformative technology with the potential to bring an extraordinary range of benefits. But the very attributes that make AI so powerful, including its ability to learn from data and therefore to evolve over time without any explicit human input, also give rise to challenges in light of the increasing role AI is certain to play in the criminal justice system. In this Article, we focus on a set of solutions to address those challenges in the context of risk assessments used in relation to criminal sentencing.

We identify a set of three principles that we believe should govern the use of AI in this context and that in combination will help ensure due process for defendants. The first principle is auditability. Whenever an AI algorithm is used to generate a risk assessment, a snapshot of the algorithm and of the data it considered as input should be acquired and preserved thus guaranteeing its availability for potential later examination. The second principle is transparency. This means that a company that makes risk assessment software should not be able to use trade secret law as a mechanism to block access to information about the algorithm and data. Through the use of tools such as protective orders, this access can be enabled in a way that both respects and preserves a company's trade secret rights while also giving courts, defendants, and other parties with standing a mechanism to access the complete set of information used in computing a particular risk assessment. The third principle is consistency, which refers to ensuring that an AI system does not produce materially inconsistent risk assessments at different times for defendants with substantially identical profiles.

In framing the application of AI in criminal justice and other fields, it is important to emphasize the difference between algorithms generally and algorithms that use AI. Put simply, all AI involves algorithms, while most algorithms do not involve AI. The problems that can arise with the use of non-AI-based algorithms in relation to criminal sentencing (as well as to the criminal justice system more broadly and to policing) are well-recognized and have attracted significant attention in recent years, spurred in part by several highly visible developments.

In 2016, ProPublica published a series of reports raising concerns about racial bias in software used to assess criminal risk.[1] Also in 2016, the Wisconsin Supreme Court issued a ruling in *State v. Loomis* rejecting a defendant's claim that the use of algorithmic risk assessment at sentencing was a violation of due process.[2] Both the ProPublica reports and the *Loomis* decision spurred coverage in the broader press.[3] Another factor in the increased attention to algorithmic risk assessments is the American Law Institute's inclusion in the 2017 proposed final draft of the *Model Penal Code: Sentencing* of a section on *Evidence-based sentencing* endorsing the development of "actuarial instruments or processes, supported by current and ongoing recidivism research, that will estimate the relative risks that individual offenders pose to public safety through their future criminal conduct."[4] The proposed final draft also states that "[w]hen these instruments or processes prove sufficiently reliable, the commission may incorporate them into the sentencing guidelines."[5]

[1] *See* Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/V3XP-733V]; Julia Angwin & Jeff Larson, *Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say*, PROPUBLICA (Dec. 30, 2016, 4:44 PM), https://www.propublica.org/article/bias-incriminal-risk-scores-is-mathematically-inevitable-researchers-say [https://perma.cc/RFS7-EDFL]. The ProPublica publication also spurred rebuttals both from Northpointe and others taking issue with ProPublica's claims. *See, e.g.*, William Dieterich et al., NORTHPOINTE INC. RESEARCH DEPT., COMPAS RISK. SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY 1 (2016) https://assets.documentcloud.org/documents/2998391/ProPublica-Commentary-Final-070616.pdf [https://perma.cc/H9HU-KF4K]; *see also* Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*," 80 FED. PROBATION 38, 38 (2016) (arguing that ProPublica's studies contain errors).

[2] 881 N.W.2d 749, 753 (Wis. 2016).

[3] *See* Danielle Citron, *(Un)Fairness of Risk Scores in Criminal Sentencing*, FORBES (Jul. 13, 2016, 3:26 PM), https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#68eaf1784ad2 [https://perma.cc/CL2M-ZJKA]; 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, 5:00 AM), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/Y22E-9X6H].

[4] MODEL PENAL CODE: SENTENCING § 6B.09 (AM. LAW. INST., Proposed Final Draft 2017).

[5] *Id.*

Unsurprisingly in light of these developments, there have been numerous publications in both the legal academic press and in the popular press addressing the issue of algorithms to assess criminal risk.[6] For example, in a comprehensive editorial introducing a June 2019 special issue of the *International Journal of Law in Context* including an article titled *Law, Liberty and Technology: Criminal Justice in the Context of Smart Machines*, Roger Brownsword and Alon Harel note the concerns raised by "a strategy for crime control that relies, first, on a new generation of smart machines that form the infrastructure for the risk assessment of individuals and groups and, then, on the technological management of that risk[.]"[7]

In a 2017 article titled *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing* published through the Responsive Communities project at the Harvard Berkman Klein Center for Internet and Society, Danielle Kehl, Priscilla Guo, and Samuel Kessler noted that while risk assessment algorithms "have the potential to improve sentencing accuracy in the criminal justice system and reduce the risk of human error and bias," they can also "reinforce or exacerbate existing biases and . . . undermine certain basic tenets of fairness that are central to our justice system."[8]

In a 2019 law review article, Sandra Mayson noted the problems inherent to prediction, writing with respect to attempts to address racial bias in risk assessments that

> [t]he deep problem is the nature of prediction itself. All prediction looks to the past to make guesses about future events. In a racially stratified world, any method of prediction will project the inequalities of the past into the

[6] *See* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 1 (2006); Erin Collins, *Punishing Risk*, 107 GEO. L.J. 57, 57 (2018); Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADV. 1, 1 (2018); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 110 (2019); Anne L. Washington, *How to Argue with an Algorithm: Lessons from the COMPAS ProPublica Debate*, 17 COLO. TECH. L.J. 131, 131 (2018); Charlotte Hopkinson, Note, *Using Daubert to Evaluate Evidence-Based Sentencing*, 103 CORNELL L. REV. 723, 724 (2018).

[7] Roger Brownsword & Alon Harel, *Law, Liberty and Technology: Criminal Justice in the Context of Smart Machines*, 15 INT'L. J.L. CONTEXT, 107, 113 (2019) (citations omitted).

[8] Danielle Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, *in* RESPONSIVE COMMUNITIES PROJECT, BERKMAN KLEIN CENTER FOR INTERNET AND SOCIETY, HARVARD LAW SCHOOL 36 (2017).

> future. This is as true of the subjective prediction that has long pervaded criminal justice as it is of the algorithmic tools now replacing it.[9]

There have also been many news articles and commentary pieces in the popular press on criminal risk assessment algorithms.[10]

Not all of the most relevant legal scholarship dates from after the 2016 *Loomis* ruling and the ProPublica reports. In a 2008 law review article titled *Technological Due Process*, Danielle Citron addressed the broad (not specifically in the criminal justice context) consequences of using algorithms to make decisions, observing that "[a]utomation generates unforeseen problems for the adjudication of important individual rights. Some systems adjudicate in secret, while others lack recordkeeping audit trails, making review of the law and facts supporting a system's decisions impossible."[11]

While some of the publications quoted or cited above mention AI, they do not generally focus on AI-specific due process challenges. Yet AI, because it can involve dynamic, computer-created algorithms that can evolve without any direct human input or oversight, creates an *additional* set of policy and legal issues over and above those arising in non-AI contexts. Thus, a core contribution of the present Article is to explore the due process challenges that will arise from AI-

[9] Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2218 (2019).

[10] *See* Ellora Thadaney Israni, *When an Algorithm Helps Send You to Prison*, N.Y. TIMES (Oct. 26, 2017), https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html [https://perma.cc/NM4J-KHKX]; Adam Liptak, *Sent to Prison By a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html [https://perma.cc/SH5J-QVJR]; Frank Pasquale, *Secret Algorithms Threaten the Rule of Law*, MIT TECH. REV. (June 1, 2017), https://www.technologyreview.com/s/608011/secret-algorithms-threaten-the-rule-of-law/ [https://perma.cc/XLV2-SZKY]; Matthias Spielkamp, *Inspecting Algorithms for Bias*, MIT TECH. REV. (June 12, 2017), https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/ [https://perma.cc/ZWJ6-AB8P]; Jason Tashea, *Courts Are Using AI to Sentence Criminals. That Must Stop Now*, WIRED (Apr. 17, 2017), https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/ [https://perma.cc/3X63-K896]; John Villasenor & Virgnia Foggo, *Algorithms and Sentencing: What Does Due Process Require?*, BROOKINGS INSTITUTION (Mar. 21, 2019), https://www.brookings.edu/blog/techtank/2019/03/21/algorithms-and-sentencing-what-doesdue-process-require/ [https://perma.cc/3K2N-TE55]; Rebecca Wexler, *When a Computer Program Keeps You in Jail*, N.Y. TIMES (June 13, 2017), https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html [https://perma.cc/9YZ5-N92B].

[11] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U.L. REV. 1249, 1253 (2008).

enabled criminal risk assessments and to present solutions that can help to mitigate them.

The remainder of this Article is organized as follows. Part I provides a brief primer on AI, including a discussion of the historical context and of the dramatic advances in recent years that have led to its increasing adoption by governments and companies.[12] Part II addresses risk assessments in criminal sentencing and presents an explanation of the spectrum of ways in which AI will be used to assess criminal risk.[13] Part III provides a broad (non-AI-specific) overview of key cases that have established the contours of what due process requires in relation to information used by a court at sentencing.[14] For example, as explained in that section, the protections that govern the admissibility of evidence at trial are often lacking at sentencing.[15] Part IV considers the case law arising from challenges to the use of algorithm-based risk assessment.[16] Though none of these cases are specifically focused on *AI-based* algorithms, they will play an important role in shaping how AI-based criminal risk assessments get used. Part V introduces the three guiding principles for AI-based criminal risk assessment noted briefly above, explaining how each would work and why each is essential to due process.[17] A final Part presents conclusions.

I. ARTIFICIAL INTELLIGENCE: A PRIMER

While specific definitions of AI vary, most definitions state that AI systems learn from experience. For instance, *Brittanica*'s online entry for *Artificial Intelligence* explains that AI is the "ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience."[18] Professional services firm, PwC, states in a 2017 report that "AI is a

[12] *See infra* Part I.
[13] *See infra* Part II.
[14] *See infra* Part III.
[15] *See id.*
[16] *See infra* Part IV.
[17] *See infra* Part V.
[18] *Artificial Intelligence*, ENCYCLOPAEDIA BRITANNICA, https://www.britannica.com/technology/artificial-intelligence [https://perma.cc/73QR-Q4NS] (last visited Mar. 27, 2020).

collective term for computer systems that can sense their environment, think, learn, and take action in response to what they're sensing and their objectives."[19]

While AI has been a topic of academic research for many decades, its emergence as a profoundly important technology for companies and governments is much more recent. To contextualize the issues arising from the use of AI in risk assessments conducted in relation to sentencing, this Part gives a brief explanation of how AI contrasts with algorithms, provides some background on the history of AI, explains some of the data and investment drivers of the rapid changes in recent years, and discusses current and emerging applications.

A. The Difference Between Algorithms and AI

An algorithm is a sequence of steps to move towards a goal.[20] Some algorithms, such as the procedure for adding two numbers, are relatively straightforward. Others, such as the algorithm used to form magnetic resonance imaging (MRI) images from electromagnetic signals measured in an MRI machine, are extremely complex.[21] However, while these two examples involve algorithms that differ greatly in complexity, neither involves artificial intelligence, because there is no learning occurring by a computer in either case. A computer that adds a thousand pairs of numbers does not become more efficient as a result of this experience, and it does not change the algorithm it uses to accomplish the task at hand. Similarly, the computer in a hospital that generates MRI images does not become more efficient with each use, nor does the algorithm used to generate these images typically evolve with the production of subsequent images.

By contrast, an AI algorithm learns from experience, leading to changes over time in the nature of the algorithm itself.[22] As noted

[19] ANAND S. RAO & GERARD VERWEIJ, SIZING THE PRIZE: WHAT'S THE REAL VALUE OF AI FOR YOUR BUSINESS AND HOW CAN YOU CAPITALISE? ii (PWC 2017).

[20] *See* John Villasenor, *In Defense of Algorithms*, SLATE (Dec. 1, 2015, 8:30 AM), https://slate.com/technology/2015/12/in-defense-of-the-algorithms-that-guide-tasks-technical-and-mundane.html [https://perma.cc/V2HR-JRT5].

[21] *See id.*

[22] While the details are beyond the scope of this Article, we note that within AI there are multiple ways in which an algorithm can learn. *See, e.g.*, Isha Salian, *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?*, NVIDIA (Aug. 2, 2018), https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning [https://perma.cc/E5QL-ZJCF].

earlier, all AI involves algorithms, while not all algorithms involve AI. This distinction is important generally as well as in the context of this Article. As far as we are aware, AI algorithms that adapt and evolve in the field are not yet in widespread use in criminal risk assessments.[23] However, this is certain to change in the future as companies create and market products that seek to leverage the additional potential benefits that can be obtained by allowing algorithms to learn, making on-the-fly adaptations based on data they have encountered.[24]

B. Historical Context

One of the earliest pioneers in AI was the British mathematician and computer scientist, Alan Turing, who in a 1947 lecture to the London Mathematical Society said, "What we want is a machine that can learn from experience."[25] In 1950, Turing published a paper titled *Computing Machinery and Intelligence* in which he wrote, "I propose to consider the question, 'Can machines think?'"[26] While the computers of the day were far too rudimentary to take actions that could be reasonably characterized as "thinking," Turing understood that future advances would make this question increasingly relevant.[27]

The second half of the twentieth century saw important progress in computing technology generally as well as its application to AI. However, throughout that period AI remained largely a focus of academic research with little tangible impact outside the laboratory. As a 1988 *New York Times* article titled *Setbacks for Artificial Intelligence* noted, "Although computers often appear to be intelligent in their everyday applications, they generally perform repetitive tasks following rigid rules set down by programs. They do not learn or make

[23] The "as far we are aware" caveat is necessary because, due to trade secret protections, it is not possible to know the full inner workings of the risk assessment algorithms that are in current use in the criminal justice system. However, we have seen no indication that AI algorithms that adapt on their own have yet experienced widespread adoption in this field.

[24] *See, e.g.*, Babak Hodjat, *Evolutionary Algorithms Are the Living, Breathing AI of the Future*, VENTUREBEAT (Feb. 13, 2018, 12:20 PM), https://venturebeat.com/2018/02/13/evolutionary-algorithms-are-the-living-breathing-ai-of-the-future [https://perma.cc/H7CD-AWX3].

[25] Alan M. Turing, Lecture to the London Mathematical Society (Feb. 20, 1947).

[26] Alan M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433, 433 (1950).

[27] *See id.*

cognitive decisions, as humans do."[28] The article also highlighted the (as of that date) "failure of artificial intelligence to quickly live up to its promise of making machines that can understand English, recognize objects or reason like a human expert."[29]

One of the most dramatic illustrations of how far both computing technology and AI have advanced in recent decades can be found in a comparison of chess at two points in time: 1997 and 2017.[30] In 1997, an IBM computer named Deep Blue beat Garry Kasparov, who was at the time the world's top ranked player, in a six game match.[31] Deep Blue was not using true artificial intelligence, as it relied instead on a "brute force" algorithm that, as an August 1997 article in *MIT Technology Review* explained, "looks as far ahead as it can at all possible moves and evaluates the strength of each position according to preprogrammed rules."[32] By exploring "200 million positions each second," Deep Blue had sufficient computing power to examine many suboptimal moves in the process of searching for a strong move.[33]

Deep Blue was the product of years of work by a team of researchers at IBM.[34] In addition, IBM had hired several chess grandmasters as consultants to help provide input on how to program Deep Blue.[35] Thus, the algorithm running on Deep Blue was developed to exploit not only speed but also the knowledge obtained from highly expert players. This knowledge was reflected in how the program would evaluate the strength of each potential move.

[28] Andrew Pollack, *Setbacks for Artificial Intelligence*, N.Y. TIMES (Mar. 4, 1988), https://www.nytimes.com/1988/03/04/business/setbacks-for-artificial-intelligence.html [https://perma.cc/RP3C-H6JA].

[29] *Id.*

[30] *See, e.g.*, Bruce Weber, *Swift and Slahing, Computer Topples Kasparov*, N.Y. TIMES (May 12, 1997), https://www.nytimes.com/1997/05/12/nyregion/swift-and-slashing-computer-toppls-kasparov.html [https://perma.cc/Y5BE-ACQT]; *see generally* David Silver et al., *Mastering Chess and Shogi By Self-Play with a General Reinforcement Learning Algorithm*, ARIXIV (2017) (providing an overview and implications of the AlphaGo Zero algorithm).

[31] Weber, *supra* note 30.

[32] *How the Chess Was Won*, MIT TECH. REV. (Aug. 1, 1997), https://www.technologyreview.com/s/400089/how-the-chess-was-won [https://perma.cc/5VR3-S86A].

[33] *Id.*

[34] *See Deep Blue*, IBM, https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/ [https://perma.cc/8X2N-FU2M] (last visited May 31, 2020) (stating that "IBM computer scientists had been interested in chess computing since the early 1950s").

[35] *See* Steven Levy, *What Deep Blue Tells Us About AI in 2017*, WIRED (May 23, 2017, 12:00 AM), https://www.wired.com/2017/05/what-deep-blue-tells-us-about-ai-in-2017 [https://perma.cc/F9V2-N8V3].

C. Artificial Intelligence Today

Fast forward to twenty years later, in 2017, when a Google computer named AlphaZero used artificial intelligence to teach itself to play chess at a level sufficient to beat the world's top chess program in a matter of hours.[36] As the researchers behind AlphaZero wrote, "Starting from random play, and given no domain knowledge except the game rules, AlphaZero achieved within 24 hours a superhuman level of play in the games of chess and shogi (Japanese chess) as well as Go, and convincingly defeated a world-champion program in each case."[37] While the comparison is not truly apples-to-apples—Deep Blue was up against the top human player of 1997, while AlphaZero played a top computer chess program of 2017—it nonetheless provides a stunning illustration of how advanced AI has become in recent years.[38] Deep Blue, for all of its (for the era) speed, was implementing an approach that was in a sense a human/machine collaboration.[39] The machine's contribution was to explore hundreds of millions of moves per second.[40] The human contribution was reflected in the algorithm used by the program running on Deep Blue to assign metrics to each of those possible moves.[41]

With AlphaZero, by contrast, the human input was far more tenuous. AlphaZero was given nothing but the rules of chess and from there it learned on its own how to play extraordinarily well through the process of playing many games, learning from its mistakes as it did so. This is analogous to what a human does—but a human needs years to play the many thousands of games necessary to become a top chess player. AlphaZero could learn by experience in hours because it had the computational power to both play and learn much faster than a human. AlphaZero was a true example of AI, in which a computer used its own experience to automatically enhance (and in this case, develop essentially in its entirety from scratch) an algorithm.[42]

Mastery of games such as chess requires a skill level that is both extremely high and very domain-specific. But the AlphaZero story illustrates a rate of progress in AI that has implications for many domains. PwC's 2017 report titled *Sizing the Prize: What's the Real*

[36] David Silver et al., *supra* note 30.

[37] *Id.*

[38] *See id.*

[39] *See id.*

[40] *See id.*

[41] *See id.*

[42] *See id.*

Value of AI for Your Business and How Can You Capitalise?, identified "nearly 300 use cases" for AI.[43] AI is ideally suited to applications in which there is a lot of data that can be analyzed to identify correlations, patterns, and dependencies that can—in theory at least—enable more accurate decisions. And in today's world, we are awash in data. To give some idea of the staggering numbers involved, consider that according to one estimate, on an average day there are 500 million tweets sent, 5 billion searches performed, and four petabytes (including 350 million photos) created on Facebook.[44] The total amount of digital data is expected to reach forty-four zettabytes in 2020, which is ten times as large as the amount of digital data that existed in 2013.[45] Of course, quality is different from quantity. Sometimes by accident and sometimes by design, data can be incomplete, deceptive, unrepresentative, and inaccurate. This presents one of the key challenges with any technology—including AI—that aims to effectively analyze and harness data.

The availability of large amounts of data that can be potentially harnessed using intelligent, adaptive algorithms is one reason why AI has moved beyond the laboratory and is now used in the commercial sector by an increasing array of companies.[46] Another reason is the large and growing investment in AI. As a *Forbes* article in early 2019 explained, "Since 2013, VC investments in AI startups had regularly increased over the following four years, with a compound annual growth rate (CAGR) of about 36%."[47] The rate of growth was even higher in 2018: According to the Q4 2018 *MoneyTree Report* from

[43] Rao & Verweij, *supra* note 19, at 3.

[44] *See* Jeff Desjardins, *How Much Data Is Generated Each Day?*, VISUAL CAPITALIST (Apr. 17, 2019), https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/ [https://perma.cc/P9VN-EACQ]. A petabyte is one million gigabytes. *Id.*

[45] *See id.* (noting that the "accumulated digital universe of data" is projected to be 44 zettabytes in 2020, up from 4.4 zettabytes in 2013). A zettabyte is one million petabytes. *Id.*

[46] While it is generally the case that more data leads to better algorithmic outcomes, there are exceptions. Under some circumstances, a smaller data set can actually lead to better decisions. *See, e.g.*, Cathy O'Neil, *Bigger Data Isn't Always Better Data,* BLOOMBERG (Mar. 1, 2017, 7:00 AM), https://www.bloomberg.com/opinion/articles/2017-03-01/bigger-data-isn-t-always-better-data [https://perma.cc/EH6J-AULL].

[47] Jean Baptiste Su, *Venture Capital Funding for Artificial Intelligence Startups Hit Record High in 2018,* FORBES (Feb. 12, 2019, 2:58 PM), https://www.forbes.com/sites/jeanbaptiste/2019/02/12/venture-capital-funding-for-artificial-intelligence-startups-hit-record-high-in-2018 [https://perma.cc/K3VM-PV6Q].

PwC and CB Insights, in the United States, "AI-related companies raised $9.3B in 2018, a 72% increase compared to 2017."[48] This level of investment is spurred by expectations about the economic impact of AI. PwC's *Sizing the Prize* report predicted that "AI could contribute up to $15.7 trillion to the global economy in 2030."[49] In a survey of 1,900 company executives across seven countries (Australia, Canada, China, France, Germany, the United Kingdom, and the United States) published in May 2019, Deloitte reported that 63% of respondents "say that AI technologies are 'very' or 'critically' important to their business success today" and that in the next two years that number will reach 81%.[50]

Today, AI underlies Siri's answers, purchase recommendations from Amazon and Netflix, driving decisions made by Tesla, interactions with smart speakers such as Alexa, playlists suggested by Pandora,[51] route mapping in Google Maps, and matching of drivers and riders by Uber and Lyft.[52] AI will impact an essentially endless list of additional applications as well, in fields including agriculture; climate modeling; economic forecasting; supply chain optimization; defense; education; and policing and criminal justice. Different fields will experience adoption of AI at different rates. For some applications, such as driverless cars, AI is a core technological enabler, and has long been a consideration at the forefront of the minds of system designers.[53] For applications that have historically been addressed using non-AI methods—including risk assessment in relation to policing and criminal justice—the biggest AI-related impacts lie largely in the future.

[48] PwC & CB Insights, MoneyTree Report: Q4 2018, at 20 (2019), https://www.pwc.com/us/en/moneytree-report/moneytree-report-q4-2018.pdf [https://perma.cc/Y7UJ-73TJ].

[49] Rao & Verweij, *supra* note 19, at 3 (noting that "estimated values are expressed in real terms at 2016 prices").

[50] *See* Jeff Loucks et al., *Future in the Balance? How Countries Are Pursuing an AI Advantage*, Deloitte (May 1, 2019), https://www2.deloitte.com/insights/stateofAI-global.html [https://perma.cc/ELD7-ZKAD].

[51] *See* R. L. Adams, *10 Powerful Examples of Artificial Intelligence in Use Today*, Forbes (Jan. 10, 2017), https://www.forbes.com/sites/robertadams/2017/01/10/10-powerful-examples-of-artificial-intelligence-in-use-today [https://perma.cc/W3AB-5YTX].

[52] *See* Gautam Narula, *Everyday Examples of Artificial Intelligence and Machine Learning*, Emerj, https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/ [https://perma.cc/5A7B-Z9MZ] (last updated Mar. 20, 2020).

[53] *See, e.g.*, Michael Krödel & Klaus-Dieter Kuhnert, Autonomous Driving Through Intelligent Image Processing and Machine Learning 712 (2001).

That said, AI-enabled risk assessment is already in (controversial) use for predictive policing, which aims to identify people or places deemed likely to be associated with future crimes.[54] In any field where there is lots of data—and criminal justice certainly qualifies as such a field—people will seek opportunities to use AI to extract as much useful information as possible from that data.

II. Criminal Risk Assessment

Although we focus here on ensuring due process in relation to the use of AI in presentence risk assessments, some of the observations we present here will also be relevant to other algorithmic (including but not limited to AI-based) risk assessments that might be performed within the criminal justice system. This includes those done at the pretrial stage, which are used to "classify defendants based on their flight risk and their threat to community safety," as well as those performed in relation to decisions regarding whether to grant parole.[55]

A. Background

Risk assessment in the American criminal justice system has a long history and has been covered in many other publications.[56] As Monahan and Skeem explain in a 2016 article titled *Risk Assessment in Criminal Sentencing*, during the second half of the nineteenth century following the adoption of parole and probation statutes in New York and Massachusetts respectively, "The explicit assessment of an offender's risk soon became a central component of criminal sanctioning in numerous American jurisdictions."[57] Alicia Solow-Niederman, YooJun Choi, and Guy Van den Broeck write that "[f]or much of the twentieth century, choices about human liberty depended on obviously subjective factors" and that "in making bail and sentencing determinations, 'clinical predictions,' or 'the largely

[54] *See* Randy Rieland, *Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?*, Smithsonian Mag., (Mar. 5, 2018), https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/ [https://perma.cc/CWP4-33K2].

[55] Charles Summers & Tim Willis, *Pretrial Risk Assessment: Research Summary*, Bureau of Justice Assistance, at 1 (Oct. 18, 2010), https://www.bja.gov/Publications/PretrialRiskAssessmentResearchSummary.pdf [https://perma.cc/CUC4-KGX2].

[56] *See* John Monahan & Jennifer L. Skeem, *Risk Assessment in Criminal Sentencing*, 12 Ann. Rev. Clin. Psychol. 489, 490 (2016).

[57] *Id.*

unstructured clinical judgment of skilled practitioners,' was used to assess the likelihood of recidivism."[58] Solow-Niederman *et al.* also note that in recent years, algorithmic solutions have gained momentum, spurred in significant part by state legislation requiring the use of risk assessments.[59] And, they observe that "[t]his apparent state enthusiasm for algorithmic solutions, however, has met mounting public and scholarly debate about the ethical and legal propriety of these tools."[60]

The evolution of risk assessment from the *ad hoc* approaches that were common in the twentieth century to the algorithmic solutions that have become widespread today is often described in terms of generations.[61] Brandon Garrett and Monahan explain that:

> [Risk assessment] instruments have evolved from first generation tools, consisting in clinical judgment and experience of a decisionmaker, to second-generation tools relying on static risk factors (such as criminal history, age, and gender), to third generation instruments both looking at risks and needs, and both static and dynamic risk factors such as educational status, employment; and fourth generation instruments, that provide individualized plans based on assessment of static and dynamic factors. A fifth generation of these tools may use machine learning techniques to predict recidivism in real-time and using far more complex analysis.[62]

In the Berkman Klein Center report on *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing* discussed earlier, Kehl *et al.* describe a similar development trajectory, starting with a "first generation, where risk assessment was conducted on a case-by-case basis by correctional staff and clinical professionals working in prisons."[63] Kehl *et al.* write that the "second generation of risk assessment tools, which emerged in the 1970s, primarily embraced static factors for measuring risk," and "[t]he third generation of risk assessments attempted to solve for the shortcomings of static risk factors by considering static and dynamic factors in tandem with one another."[64] They further note that the fourth generation "builds off of the third generation but . . . embraces a more 'systematic and comprehensive' approach to

[58] Alicia Solow-Niederman et al., *The Institutional Life of Algorithmic Risk Assessment*, 34 BERKELEY TECH. L.J. 705, 710–11 (2019).

[59] *See id.* at 715.

[60] *Id.*

[61] *See id.* at 706–07.

[62] Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439, 451 (2020).

[63] Kehl, *supra* note 8, at 8.

[64] *Id.* at 9.

measuring recidivism and treating offenders based on their specific risk factors and characteristics."[65]

This progression is also reflected in the increasing number of states that have adopted criminal justice policies relating to the use of Risk and Needs Assessment (RNA) tools. According to a September 2017 publication from the Center for Sentencing Initiatives, which is a project of the National Center for State Courts (NCSC), "[o]ver the last 10 years at least 18 states have adopted, by statute, administrative policy or judicial policy, a state-wide use of RNA at the sentencing phase. There are also local jurisdictions using RNA at sentencing in at least 5 other states."[66]

For example, Delaware law provides that "upon conviction of any person for any crime and before sentencing, the court may" order a presentence investigation that "should include administration of an objective risk and needs assessment instrument."[67] Under Kentucky law, "[s]entencing judges shall consider . . . the results of a defendant's risk and needs assessment included in the presentence investigation . . . and [t]he likely impact of a potential sentence on the reduction of the defendant's potential future criminal behavior."[68] In Louisiana, "[e]ligibility for presentence investigation assessment shall be limited to an adult felony defendant who is otherwise eligible for probation or reentry after adjudication of guilt" and who has signed a "complete and irrevocable written acknowledgment of the purpose of the assessment and waiver of confidentiality over the global risk scores contained in the presentence investigation validated risk and needs assessment tool."[69] The Louisiana statute further provides that "[t]he presentence investigation validated risk and needs assessment tool and evaluation report may be utilized by the sentencing court prior to

[65] *Id.*

[66] NAT'L CTR. FOR STATE COURTS, USE OF RISK AND NEEDS ASSESSMENT INFORMATION IN STATE SENTENCING PROCEEDINGS 3, 6 n.10 (Sept. 2017), https://www.ncsc.org/~/media/Microsites/Files/CSI/EBS%20RNA%20brief%20Sep%202017.ashx [https://perma.cc/92FU-NBLD] [hereinafter NCSC 2017 RNA Report] (stating that "the following states have adopted a formal policy to provide RNA information to inform sentencing decisions: AL, AK, AZ, AR, DE, ID, IN, KY, LA, MT, NE, ND, OH, OK, TN, UT, WV, and WI. Some jurisdictions in other states (e.g., CA, CO, IA, TX, and OR) also provide this information to the court"); *see also State Policies and Legislation*, NAT'L CTR. FOR STATE COURTS (June 2018), https://www.ncsc.org/microsites/csi/home/In-the-States/State-Activities/State-Policies-and-Legislation.aspx [https://perma.cc/3CWJ-KFEW] (outlining RNA statutes by state).

[67] DEL. CODE ANN. tit. 11, § 4331 (West 2012).

[68] KY. REV. STAT. ANN. § 532.007(3)(a)–(b) (West 2011).

[69] LA. STAT. ANN. § 15:326(b) (2014).

determining an appropriate sentence, in order to evaluate the defendant's risk of committing future offenses and to reduce the recidivism of the defendant."[70]

The Ohio Department of Rehabilitation and Correction (DRC) contracted in 2006 with the Center for Criminal Justice Research and the University of Cincinnati to develop the Ohio Risk Assessment System (ORAS).[71] The DRC website explains that "[t]he ORAS tools can be used at pretrial, prior to or while on community supervision, at prison intake, and in preparation for reentry just prior to release from prison."[72] Ohio law provides that DRC "shall select a single validated risk assessment tool for adult offenders" to be used, among other things, for sentencing in municipal, common pleas, and county courts.[73] Under Tennessee law, the sentencing court must consider "[t]he result of the validated risk and needs assessment conducted by the department and contained in the presentence report," and the sentence must be based on factors including "the validated risk and needs assessment."[74] At the federal level, enactment of the First Step Act of 2018 led to the development of a new RNA tool to be used by the Department of Justice.[75] The Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN) "contains static risk factors (e.g. age and crime of conviction) as well as dynamic items (i.e. participation or lack of participation in programs like education or drug treatment) that are associated with either an increase or a reduction in risk of recidivism."[76]

One of the contradictions in the policy discussion around risk assessments arises in relation to the question of how they impact sentencing decisions. For example, the NCSC's Q&A-format website titled *Use of Risk and Needs Assessment Information* asks the question, "Are risk and needs assessment scores used by the courts to

[70] *Id.* § 15:327(a).

[71] *Ohio Risk Assessment System*, OHIO DEP'T OF REHAB. AND CORR., https://drc.ohio.gov/oras [https://perma.cc/9CQS-X8SQ] (last visited on Mar. 27, 2019).

[72] *Id.*

[73] OHIO REV. CODE ANN. § 5120.114 (West 2019).

[74] TENN. CODE ANN. § 40-35-210(b)(8), (f) (2019).

[75] First Step Act of 2018, Pub. L. No. 115-391, § 101, 132 Stat. 5194, 5195 (2018).

[76] Press Release, U.S. Dep't of Justice, Department of Justice Announces the Release of 3,100 Inmates Under First Step Act, Publishes Risk and Needs Assessment System (July 19, 2019), https://www.justice.gov/opa/pr/department-justice-announces-release-3100-inmates-under-first-step-act-publishes-risk-and [https://perma.cc/R6HU-WMT2].

make decisions about the appropriate severity of punishment?"[77] The initial portion of the answer is, "No. State courts have disapproved the use of such scores as aggravating factors in determining the severity of punishment."[78] Yet later on in the same answer, the document states that "RNA information may also be a factor in determining whether imprisonment should be avoided or scaled back for low risk offenders."[79] As a matter of logic, if a "good" risk score can lead a sentencing court to issue a lower sentence than it otherwise would have, then it cannot be true that the scores have no impact on the severity of the punishment. After all, such a system would unfairly punish a defendant who is mischaracterized as high risk, and therefore fails to receive the shorter sentence that would have accompanied a more accurate risk assessment.

Furthermore, as the citations from various state laws above make clear, laws in multiple states *specifically require* consideration of risk scores when determining a sentence.[80] Our point is not that consideration of risk is inappropriate at sentencing—in fact, done properly, it can clearly be beneficial, as it can reduce bias and sentencing disparities, as well as increase predictive accuracy.[81] Rather, it is that the policy discussion will be better advanced when it is framed with recognition that risk scores *are* impacting sentencing, and therefore that the due process issues that arise in connection with ensuring the accuracy of those scores merit increased attention.[82]

B. AI-enabled Risk Assessment: A Spectrum of Approaches

In the future, risk assessments used in relation to criminal sentencing will likely be increasingly reliant on algorithms that embody AI. As an AI-based algorithm could take many forms, simply categorizing an algorithm as involving AI leaves room for wide variations.[83] Risk assessment tool taxonomies will need to be updated to reflect these impending changes. Rather than attempting to

[77] *See, e.g.*, NCSC 2017 RNA Report, *supra* note 66 at 1, 3.

[78] *Id.* at 3.

[79] *Id.*

[80] *See* KY. REV. STAT. ANN. § 532.007 (West 2011); OHIO REV. CODE ANN. § 5120.114 (West 2019); TENN. CODE ANN. § 40-35-210 (1989).

[81] Solow-Niederman et al., *supra* note 58, at 711 (describing historical recognition of the need to make risk assessment less biased and subjective).

[82] *See id.* at 714 (arguing that "[t]he practical stakes" of using risk scores "are high" because of their increased use and impact on sentencing in the United States).

[83] *See* Salian, *supra* note 22.

oversimplify things by defining multiple non-overlapping and seemingly contrasting categories (i.e. static vs. dynamic) for AI-enabled risk assessment tools, it is more flexible and technologically accurate to describe them in relation to a spectrum, the two ends of which raise different policy and due process concerns.[84]

Approaches at one end of the spectrum can be called "static," in the sense that while AI is used *in advance* of a product's release during the process of developing and tuning a risk assessment tool, once the algorithm has been finalized and placed into service, it remains static.[85] Put another way, with static algorithms, the AI is used by the manufacturer of the system to help *design* an algorithm that does not, after the design is complete, evolve further on its own.[86] This means that the algorithm is knowable to the manufacturer, making it at least potentially accessible for analysis pursuant to a due process claim.[87] A manufacturer might nonetheless attempt to block access by citing trade secret rights, but that would be a legal obstacle, not a technological obstacle.[88]

At the other end of the spectrum are solutions that can be called "dynamic."[89] This refers to approaches in which AI is built into the system so that the algorithm can continuously evolve on its own in

[84] *See* Aaron D. Kirk, *Artificial Intelligence and the Fifth Domain*, 80 A.F.L. REV. 183, 194 (2019) (explaining that AI tasks "exist along spectrums that vary across several dimensions").

[85] *See* Yavar Bathaee, *The Artificial Intelligence Blackbox and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 898 (2018) ("On the most inflexible end of the spectrum are AI that make decisions based on preprogrammed rules from which they make inferences or evaluate options.").

[86] *See id.* (describing static AI programs that deal with "preprogramed rules" and do not develop).

[87] How easy it would be for the manufacturer to know the algorithm would depend on multiple factors, including the extent to which the manufacturer had taken affirmative steps to design the system to support outputting a clear, human-readable description of the algorithm, including any changes to the algorithm that had occurred after it was initially written by human programmers. *See id.* at 898–99 (describing static AI by using the examples of AI that can evaluate chess moves according to a scoring formula or school admissions AI that uses a mathematical formula, and how, potentially, both the chess scoring formula and school admissions formula could be accessed for analysis).

[88] *See* Jessica M. Meyers, *Artificial Intelligence and Trade Secrets*, 11 LANDSLIDE 17, 20 (2019) ("AI technologies are particularly well suited to protection by trade secret.").

[89] *See* Kirk, *supra* note 84, at 194 (describing the spectrum of AI programs as including static and dynamic ends).

response to data that it collects.[90] This could occur over very short time scales.[91] For example, consider an AI-based presentencing risk assessment system that continuously monitors nationwide news feeds and arrest records for information regarding new arrests, looking for recidivism examples. If, due to this monitoring, the AI system identifies previously unknown statistical correlations that appear to impact the recidivism rate for people convicted of a particular crime, it might use that information to modify its algorithm for computing risk scores for future people convicted of a similar crime. This evolution could happen very quickly. Even a single new recidivism incident, when pulled into the AI system's database and used to recalculate statistical correlations, could result in an adjustment of the numerical parameters that impact risk score computation.

We are not suggesting that this is necessarily a *good* way for an algorithm to evolve. In fact, there are multiple reasons why it could be problematic—including the implicit assumption that people who are convicted of "similar" crimes share enough in common that the past recidivism rate of some individuals within that group can help predict the future recidivism risk of others in that group.[92] Rather, we are providing an example of what an AI system nonetheless might do, either because its original human programmers designed it that way or because it made a non-human-influenced decision on its own to monitor arrest records and news stories for use as a source of data.

Because this evolution can occur completely autonomously, it is possible that no one—not even the manufacturer of the system—has a snapshot of the algorithm in the form that it exists at the moment when it is used to calculate a particular risk assessment score. Relative to static algorithms, dynamic algorithms pose a more complex set of due process issues, since the information regarding the algorithm used to compute a person's score may no longer exist by the time a request for it is made many weeks or months after the score was computed.[93] Even if any legal challenges to obtaining this information were successfully

[90] *See* Bathaee, *supra* note 85, at 898 ("For this sort of AI, there are no pre-programmed rules about how to solve the problem at hand, but rather only rules about how to learn from data.").

[91] *See id.* (describing dynamic AI evolving over the course of a chess game).

[92] *See* Citron, *supra* note 3 (describing how Wisconsin's risk scores are based on group probabilities and not individual likelihood of recidivism).

[93] *See* Bathaee, *supra* note 85, at 899 (noting that "the defining characteristic of [dynamic AI] is their ability to learn from data" and thus, the information used by the AI is ever-changing).

overcome by a person raising a due process challenge, the technological barrier could be insurmountable.

Of course, not all AI-enabled risk assessment will be firmly at one of these two ends of the "static" to "dynamic" spectrum. An algorithm that is updated every six months via an AI-designed software upgrade from the manufacturer is not truly static, but it is close to the static end of the spectrum. An algorithm that is hosted on a cloud server managed by the manufacturer and updated weekly based on AI-driven enhancements derived from the collective data from all of the manufacturer's customers is not truly dynamic, but it is close to the dynamic end of the spectrum. The broader point is that, with respect to due process, asking whether a risk assessment tool uses AI will yield an answer that only tells part of the story. If the answer is "yes," there will need to be further questions to identify how AI is used, the frequency and extent to which the algorithm is updated, and crucially, what steps, if any, are in place to enable snapshots of the algorithm to be obtained every time a risk assessment is performed. In the present Article, we focus (though not exclusively) on dynamic AI, as algorithms that automatically evolve over time raise particularly timely and important due process challenges.

III. DUE PROCESS AND INFORMATION USED AT SENTENCING

While the issue of due process specifically in relation to AI-based presentence risk assessments has not been tested in the courts, there is a wealth of case law directly relevant to the broader issue of due process in relation to information used at sentencing.[94] We review some of the key case law and its application to AI-based systems here, focusing on four themes: (1) accuracy and admissibility of information used at sentencing, (2) access by a defendant to information used at sentencing, (3) the scientific validity of AI-based presentencing risk assessment methods, and (4) the extent to which such approaches might inadvertently consider impermissible factors such as race.[95]

[94] *See, e.g.*, Townsend v. Burke, 334 U.S. 736, 741 (1948); *see also* United States v. Cook, 550 F.3d 1292, 1296 (10th Cir. 2008); Stewart v. Erwin, 503 F.3d 488, 495 (6th Cir. 2007).

[95] *See infra* Sections III.A–D. The discussion herein focuses on the aspects of sentencing most likely to be impacted by AI. Even without the additional factor of AI, sentencing is a complex area of law characterized, among other things, by varying levels of discretion. There is an additional set of Sixth Amendment and other questions that could be asked in light of the combination of various sentencing frameworks with different potential approaches to introducing AI-based risk assessments into the sentencing process.

A. Accuracy and Admissibility of Information Used at Sentencing

A defendant has a due process right to be sentenced based on information that is not materially inaccurate.[96] The landmark 1948 Supreme Court ruling in *Townsend v. Burke* is a key reference point on this issue.[97] As the Sixth Circuit explained when citing *Townsend* in a 2007 decision, "*Townsend* and its progeny are generally viewed as having established a due process 'requirement that a defendant be afforded the opportunity of rebutting derogatory information demonstrably relied upon by the sentencing judge, when such information can in fact be shown to have been materially false.'"[98] The Tenth Circuit put it similarly in a 2008 ruling, noting that "the due process clause protects a defendant's right not to be sentenced on the basis of materially incorrect information."[99] And, as the DC Circuit explained in 1983, "[t]he requirements of due process are not suspended with the pronouncement of guilt, but continue to operate in the sentencing process. Thus, the sentencing judge may not rely on mistaken information or baseless assumptions."[100] A closely related concern arises when courts consider unproven allegations, whether true or not, at a sentencing hearing.[101] This has led federal appeals courts to issue rulings vacating sentences and remanding for resentencing in cases including *United States v. Juwa* (in the Second Circuit) and *United States v. Weston* (in the Ninth Circuit). [102]

However, not all instances in which a sentencing court cites unproven allegations have been found to violate a defendant's due process rights.[103] *Williams v. New York* was a 1949 Supreme Court decision issued one year after *Townsend*.[104] *Williams* arose from a death sentence imposed at a sentencing proceeding in which the judge, in addition to referencing the burglary and murder for which the

[96] *See Townsend*, 334 U.S. at 741 (overturning a criminal conviction based on materially untrue information).

[97] *Id.*

[98] *Stewart,* 503 F.3d at 495 (quoting Collins v. Buchkoe, 493 F.2d 343, 345 (6th Cir. 1974)).

[99] *Cook,* 550 F.3d at 1296.

[100] United States v. Lemon, 723 F.2d 922, 933 (D.C. Cir. 1983) (internal citations omitted).

[101] *See* United States v. Juwa, 508 F.3d 694, 702 (2d Cir. 2007) (vacating a sentence due in part to unproven allegations).

[102] *See id.*; United States v. Weston, 448 F.2d 626, 634 (9th Cir. 1971).

[103] *See, e.g.*, Williams v. New York, 337 U.S. 241, 252 (1949).

[104] *See generally id.* (finding no due process violation during a sentencing proceeding).

defendant had recently been convicted, also mentioned thirty other burglaries allegedly committed by the defendant, but for which there had been no conviction.[105]

After receiving the sentence, the defendant asserted that his due process rights had been violated by the judge's consideration of unproven crimes.[106] The Supreme Court disagreed, writing that

> [i]n determining whether a defendant shall receive a one-year minimum or a twenty-year maximum sentence, we do not think the Federal Constitution restricts the view of the sentencing judge to the information received in open court. The due-process clause should not be treated as a device for freezing the evidential procedure of sentencing in the mold of trial procedure. So to treat the due-process clause would hinder if not preclude all courts—state and federal—from making progressive efforts to improve the administration of criminal justice.[107]

The *Williams* Court also cited the historical distinctions between the procedural protections at trial and at sentencing, noting that

> [t]ribunals passing on the guilt of a defendant always have been hedged in by strict evidentiary procedural limitations. But both before and since the American colonies became a nation, courts in this country and in England practiced a policy under which a sentencing judge could exercise a wide discretion in the sources and types of evidence used to assist him in determining the kind and extent of punishment to be imposed within limits fixed by law.[108]

Williams is sometimes portrayed as having narrow scope because, as the Ninth Circuit put it in citing *Williams* in its *Weston* ruling, Williams had "urged that sentencing should be turned into a second trial" and also failed to object to the judge's citation at the sentencing hearing of the unproven crimes.[109] But the extensive *dicta* in *Williams* underscoring the notion that due process protections are lower at sentencing hearings than at trial is consistent with multiple more recent lower court decisions.

For instance, consider the exclusionary rule, which (subject to a "good faith" exception) prohibits the introduction at trial of evidence obtained in violation of a defendant's constitutional rights.[110] By

[105] *See id.* at 244.

[106] *See id.* at 245.

[107] *Id.* at 251.

[108] *Id.* at 246.

[109] United States v. Weston, 448 F.2d 626, 631 (9th Cir. 1971).

[110] *See* Weeks v. United States, 232 U.S. 383, 393 (1914) ("If letters and private documents can thus [without a warrant] be seized and held and used in evidence against a citizen accused of an offense, the protection of the 4th Amendment, declaring his right to be secure against such searches and seizures, is of no value, and,

contrast, multiple federal appeals courts have concluded that the exclusionary rule has little force at all at sentencing. This includes the Tenth Circuit in 2001 in *United States v. Ryan*,[111] the Seventh Circuit in 2000 in *United States v. Brimah*,[112] and the First Circuit in 2002 in *United States v. Acosta*.[113]

Even when law enforcement has *intentionally* violated a defendant's constitutional rights in order to obtain information intended to secure a longer sentence, courts have sometimes been unwilling to apply the exclusionary rule.[114] In *Acosta*, the First Circuit noted that multiple other circuits had "left open the possibility that the exclusionary rule will still apply if there is an indication that the police violated the defendant's Fourth Amendment rights with the intent to secure an increased sentence."[115] Against that backdrop, the First Circuit went no further, writing "[w]e leave open the question of whether the exclusionary rule would bar the use of evidence when police intentionally act in violation of the Fourth Amendment in order to increase a defendant's sentence."[116]

On the other side of the ledger it is possible to cite the Ninth Circuit's 1968 *Verdugo v. United States* decision, in which the court concluded that "where, as here, the use of illegally seized evidence at sentencing would provide a substantial incentive for unconstitutional searches and seizures, that evidence should be disregarded by the

so far as those thus placed are concerned, might as well be stricken from the Constitution."); *see also* Illinois v. Krull, 480 U.S. 340, 349 (1987) ("The application of the exclusionary rule to suppress evidence obtained by an officer acting in objectively reasonable reliance on a statute would have as little deterrent effect on the officer's actions as would the exclusion of evidence when an officer acts in objectively reasonable reliance on a warrant."); United States v. Leon, 468 U.S. 897, 922 (1984) ("We conclude that the marginal or nonexistent benefits produced by suppressing evidence obtained in objectively reasonable reliance on a subsequently invalidated search warrant cannot justify the substantial costs of exclusion."); Mapp v. Ohio, 367 U.S. 643, 651 (1961) (extending the exclusionary rule to state courts).

[111] 236 F.3d 1268, 1271 (10th Cir. 2001) ("[T]he extension of the exclusionary rule to sentencing would, in the ordinary case, have a minimal deterrent effect on the police.").

[112] 214 F.3d 854, 858 (7th Cir. 2000) ("Against the backdrop of the traditionally broad sentencing inquiry, and the congressional mandate in U.S.S.G. § 3116 that courts consider all relevant information in sentencing, the detrimental effects of applying the exclusionary rule at sentencing are obvious.").

[113] 303 F.3d 78, 86 (1st Cir. 2002) ("[W]e hold that the exclusionary rule does not bar the use of evidence seized in violation of a defendant's Fourth Amendment rights in sentencing.").

[114] *See, e.g.*, *id.* at 85.

[115] *Id.*

[116] *Id.* at 86.

sentencing judge."[117] Yet in *United States v. Kim* in 1994, the Ninth Circuit acknowledged the "factual similarities" with *Verdugo* but arrived at the opposite conclusion.[118]

Yet another reference point in assessing the sort of information that is admissible at sentencing is the Supreme Court's 1972 *United States v. Tucker* decision.[119] Tucker had been sentenced at a proceeding in which "the District Judge conducted an inquiry into the respondent's background, and . . . gave explicit attention to . . . three previous felony convictions."[120] However, several years after the sentencing, two of the earlier convictions were found to be "constitutionally invalid."[121] Tucker then challenged the validity of the sentence.[122] The Supreme Court agreed, writing that "if the trial judge . . . had been aware of the constitutional infirmity of two of the previous convictions, the factual circumstances of the respondent's background would have appeared in a dramatically different light at the sentencing proceeding."[123]

As the above case citations make clear, while there are some defendant-favorable decisions like *Tucker* and *Verdugo*, there is also a substantial body of case law underscoring the lower level of protection at sentencing hearings.[124] This is also reflected in rules of evidence at the federal level and in multiple states.[125] The Federal Rules of Evidence contain an exception stating that (other than with regard to privilege) they do not apply to sentencing proceedings.[126] In addition, the Supreme Court has "authoriz[ed] the consideration of offender-specific information at sentencing without the procedural

[117] 402 F.2d 599, 613 (9th Cir. 1968).

[118] *See* United States v. Kim, 25 F.3d 1426, 1434 (9th Cir. 1994) (noting that, in contrast with Verdugo, "the opposite result . . . is proper in the instant case").

[119] 404 U.S. 443, 443 (1972).

[120] *Id.* at 444.

[121] *Id.* at 445.

[122] *See id.* (stating that Tucker challenged the validity of his sentence).

[123] *Id.* at 448.

[124] *Compare* Verdugo v. United States, 402 F.2d 599, 614 (9th Cir. 1968), *and Tucker*, 404 U.S. at 445, *with* United States v. Ryan, 236 F.3d 1268, 1271 (10th Cir. 2001), United States v. Brimah, 214 F.3d 854, 858 (7th Cir. 2000), *and* United States v. Acosta, 303 F.3d 78, 86 (1st Cir. 2002).

[125] *See* ILL. R. EVID. 1101(b)(3) (underscoring the lower level of protection at sentencing hearings); IND. R. EVID. 101(d)(2) (underscoring the lower level of protection at sentencing hearings); WIS. STAT. § 911.01(4)(c) (1973) (underscoring the lower level of protection at sentencing hearings).

[126] *See* FED. R. EVID. 1101(d)(3) (identifying sentencing as a proceeding to which "these rules—except for those on privilege—do not apply").

protections attendant at a criminal trial."[127] And, commentary in the U.S. Sentencing Commission's 2016 Guidelines Manual states that "[i]n determining the relevant facts, sentencing judges are not restricted to information that would be admissible at trial."[128]

At the state level there is significant variation on the extent to which evidentiary protections that apply at trial are also used at sentencing.[129] For example, in Texas, the rules of evidence apply at sentencing hearings.[130] In Wisconsin, Illinois, and Indiana, they do not.[131] This can have direct consequences in relation to risk assessments. In *Malenchik v. State* in 2010, the Indiana Supreme Court considered a challenge to "the admissibility of assessment tool results at sentencing on grounds of alleged lack of scientific reliability under Indiana Evidence Rule 702."[132] The court then noted that the "Indiana Rules of Evidence, except with respect to privileges, do not apply in trial court sentencing proceedings" and explained that "sentencing proceedings are exempted from the rules of evidence 'to provide the trial judge with the widest range of relevant information in reaching an informed decision.'"[133]

In Virginia, the application of the rules of evidence is "mandatory" in some types of sentencing proceedings (including capital murder sentencing hearings) and "permissive" in others.[134]

[127] *See* Witte v. United States, 515 U.S. 389, 400 (1995).

[128] *See* U.S. SENTENCING GUIDELINES MANUAL § 6A1.3 cmt. (U.S. SENTENCING COMM'N 2004) [hereinafter U.S.S.G. MANUAL].

[129] *Compare* TEX. R. EVID. 101(e) (suggesting that the rules of evidence apply during sentencing), *with* ILL. R. EVID. 1101(b)(3) (stating that the rules of evidence do not apply during sentencing).

[130] *See* TEX. R. EVID. 101(e) (indicating that sentencing is not among the exceptions to which the rules of evidence—other than in relation to privilege—do not apply).

[131] *See* WIS. STAT. § 911.01(4)(c) (stating that the rules are inapplicable—except in relation to privilege and some narrow admissibility carveouts such as results of an HIV test—to sentencing proceedings); ILL. R. EVID. 1101(b)(3) (noting that—except in relation to privilege—the rules are inapplicable in sentencing proceedings); IND. R. EVID. 101(d)(2) (noting that—except in relation to privilege—the rules are inapplicable in sentencing proceedings).

[132] 928 N.E.2d 564, 573 (Ind. 2010); *see also* IND. R. EVID. 702 (addressing "Testimony by Expert Witnesses").

[133] *Malenchik*, 928 N.E.2d at 573, 574 (citing Dumas v. State, 803 N.E.2d 1113, 1121 (Ind. 2004)).

[134] *See* VA. SUP. CT. R. 2:1101 (noting that "[e]xcept as otherwise provided by statute or rule, adherence to the Rules of Evidence (other than with respect to privileges) is permissive, not mandatory, in the following situations: (1) Criminal proceedings other than (i) trial, (ii) preliminary hearings, (iii) sentencing proceedings before a jury, and (iv) capital murder sentencing hearings").

Washington State's rules of evidence contain a provision stating that they "need not be applied" at sentencing proceedings.[135] In New Jersey, the rules of evidence (other than with respect to privilege) "may be relaxed . . . to admit relevant and trustworthy evidence in the interest of justice . . . [in] proceedings in a criminal or juvenile delinquency action in which information is presented for the court's use in exercising a sentencing or other dispositional discretion."[136]

The upshot is that at the federal level and in many states, the rules of evidence disfavor the defendant in relation to sentencing, leaving significant uncertainty about what, if any, evidentiary protections actually apply. And, at both the federal and state levels, the jurisprudence regarding the admissibility and accuracy of information at sentencing hearings is inconclusive. Advocates of defendants' rights can cite cases such as *Townsend*, *Tucker*, and their progeny.[137] And—importantly with respect to the specific issue of presentence risk assessments—there are the procedural protections of Rule 32 of the Federal Rules of Criminal Procedure, which requires that a defendant be given the presentence report (and thus any risk assessment contained within the report) "at least 35 days before sentencing unless the defendant waives this minimum period" and states that "[w]ithin 14 days after receiving the presentence report, the parties must state in writing any objections, including objections to material information . . . contained in or omitted from the report."[138] And while Rule 32 states that the court "may accept any undisputed portion of the presentence report as a finding of fact," it also says that the court "must—for any disputed portion of the presentence report or other controverted matter—rule on the dispute or determine that a ruling is unnecessary either because the matter will not affect sentencing, or because the court will not consider the matter in sentencing."[139]

On the other hand, there are numerous rulings, including some cited above, that highlight the *lack* of protection against the introduction of unreliable or improperly obtained information at

[135] *See* WASH. R. EVID. 1101(c) (stating that "[t]he rules (other than with respect to privileges, the rape shield statute and ER 412) need not be applied" to (as identified in (c)(3)) sentencing).

[136] *See* N.J. R. EVID. 101(a)(2), (a)(2)(C).

[137] *See generally* United States v. Tucker, 404 U.S. 443 (1972) (implementing a defendant favorable decision); Townsend v. Burke, 334 U.S. 736 (1948) (implementing a defendant-favorable decision).

[138] FED. R. CRIM. P. 32(e)(2), 32(f)1.

[139] *Id.* at 32(i)(3)(A)–(B).

sentencing. In the aggregate, the case law and the patchwork state-by-state variations in the applicability of the rules of evidence at sentencing (and the lack of such applicability in the federal courts) leave no doubt that the level of due process protection is lower at sentencing than at trial, while also creating very little clarity regarding how *much* lower it is.

So, what does this all mean in terms of the presentence risk assessments that are output by AI-based systems? Defendants clearly have a due process right not to be sentenced based on a risk assessment that is materially inaccurate. But how should that determination be made? Risk assessments are inherently probabilistic—for example, classifying a defendant's level of risk on a numerical scale representing highest risk at one end of the scale and lowest risk at the other. This means that it would be possible to explore inaccuracy over a sufficiently large group and given a sufficiently long amount of time. For example, if there is a group of 100 defendants identified by a particular risk assessment algorithm as having high risk to recidivate (while not incarcerated), and after many years only five of them actually do, then it is easy to conclude that the algorithm's designation of "high risk" was inaccurate. But when considering only the evaluation of a single individual defendant, without the broader context of the overall statistical performance of the algorithm, no such conclusion can be drawn.

The goal of avoiding material inaccuracies also raises important definitional challenges of what is meant by "materially inaccurate" in relation to the use of AI. The large data sets used to both train and run AI algorithms will rarely be completely accurate.[140] That raises the question of whether there should be any sort of metric applied to the input data to indicate its accuracy, and even if such a metric could be developed and successfully applied, what standards should be used to determine materiality of any defects. There are also additional complexities introduced when data from multiple data sets are combined, perhaps in ways that might either attenuate or amplify the impact of inaccuracies in the data.[141]

[140] *See, e.g.*, SHARAD GOEL, RAVI SHROFF, JENNIFER SKEEM, & CHRISTOPHER SLOBOGIN, THE ACCURACY, EQUITY, AND JURISPRUDENCE OF CRIMINAL RISK ASSESSMENT 7 (2018) (noting that "[m]any have expressed skepticism that risk assessments can ever be fair, as the training data necessarily contain inaccuracies, some of which arise through biases in past human actions").

[141] Mathematically, one way to indicate the potential presence of inaccuracies is through confidence intervals. However, this raises multiple challenges. First, it will not always be clear how to calculate confidence intervals, leading to what

Another challenge is that the burden of proof to show inaccuracy lies with the defendant.[142] In the case of a static risk assessment algorithm—that is, an algorithm that remains unchanged over many years and is used to evaluate a large number of defendants over that time span—a defendant with the benefit of years of prior data might be able to show that the historical outputs from that algorithm have been statistically inaccurate. But with AI—or at least the subset of AI systems in which the algorithm is constantly evolving—this becomes impossible. The algorithm used today might be different from the one that was used last year, or last week. This makes it very hard to do any sort of historical accuracy testing in relation to probabilistic risk determinations, even when those determinations might in fact be, probabilistically speaking, inaccurate.

A defendant not only has a due process interest in the accuracy of the algorithm's output, but also in the accuracy of the information used to generate that output. This includes not only the defendant-specific data (e.g., the specific crime for which he or she has been convicted) that is input to the algorithm, but also the algorithm itself. To take an unrealistic but illustrative example, consider an algorithm that has a defect in the software such that any defendant who is identified as having committed a crime in a zip code ending in the number "3" is automatically bumped up to a higher risk category.[143] Clearly, this would be manifestly unfair, and would be a blatant violation of due process. It might also be an algorithmic defect that was unknowable to the defendant due to the secrecy considerations we discuss next.

B. Secrecy of Information Used at Sentencing

What are the due process issues arising when a sentencing proceeding makes use of secret information—that is, information to

would likely be highly divergent opinions on what those intervals should be. Second, a risk assessment is already a probabilistic statement, and we are not convinced it would simplify things to overlay yet another layer of complexity by adding confidence intervals.

[142] Courts will only adopt risk assessment algorithms that they believe to be accurate. Stated another way, the choice by a court to adopt a risk assessment algorithm of necessity conveys that the court has concluded, correctly or otherwise, that it is accurate. Neither the prosecution nor the court have any incentive to challenge the accuracy of the algorithm. Such a challenge, if it is raised, will come from the defendant.

[143] In constructing in this example, we are assuming that the last digit of the zip code is a completely random number with no correlation to recidivism.

which the defendant is not given access? Given that, as discussed in the prior section, a defendant has a right to be sentenced based on information that is not materially inaccurate, logic would seem to dictate that the secrecy question should be largely moot. After all, to exercise the right to challenge information as materially inaccurate, a defendant must be able to access that information. Unfortunately, the case law regarding a defendant's right to access information used at sentencing is only partially consistent with this conclusion.[144]

In exploring this issue further, it is instructive to start with *Gardner v. Florida*, a 1977 Supreme Court decision that, while unrelated to algorithms, directly addressed the issue of a defendant's right to access information used at sentencing.[145] In January 1974, Gardner was sentenced to death by a Florida state court after being convicted of first-degree murder.[146] The judge's decision to impose the death penalty—rather than the life sentence recommended by the jury—was based in part on information included in a presentence investigation report.[147] Though Gardner's counsel had been given a copy of some portions of the report, other parts were considered confidential, and were not disclosed to his counsel before sentencing.[148] After the Florida Supreme Court affirmed the sentencing court's decision, Gardner filed a petition with the U.S. Supreme Court, which vacated Gardner's sentence and remanded the case for resentencing.[149]

Writing for the plurality in *Gardner*,[150] Justice Stevens acknowledged the constitutional challenges raised by "a capital-sentencing procedure which permits a trial judge to impose the death sentence on the basis of confidential information which is not disclosed to the defendant or his counsel."[151] The plurality also observed that "[f]rom the point of view of society, the action of the sovereign in taking the life of one of its citizens . . . differs dramatically from any other legitimate state action" and that "the

[144] *See generally* Gardner v. Florida, 430 U.S. 349 (1977) (addressing the issue of a defendant's right to access information used at sentencing).

[145] *See id.*

[146] *See id.* at 353.

[147] *See id.*

[148] *See id.* at 351.

[149] *See id.* at 362.

[150] The plurality opinion was written by Justice Stevens and joined by Justices Powell and Stewart. *See id.* at 351. In addition, Chief Justice Burger and Justices Blackmun, Brennan, and White issued concurrences, and Justices Marshall and Rehnquist issued dissents. *See id.* at 362–71.

[151] *Id.* at 358.

sentencing process, as well as the trial itself, must satisfy the requirements of the Due Process Clause."[152]

Noting that "if [confidential information] is the basis for a death sentence, the interest in reliability plainly outweighs the State's interest in preserving the availability of comparable information in other cases," the plurality concluded "that petitioner was denied due process of law when the death sentence was imposed, at least in part, on the basis of information which he had no opportunity to deny or explain."[153] Furthermore, the plurality noted that

> [e]ven if it were permissible to withhold a portion of the report from a defendant, and even from defense counsel, pursuant to an express finding of good cause for nondisclosure, it would nevertheless be necessary to make the full report a part of the record to be reviewed on appeal.[154]

Justice Stevens also wrote that

> consideration must be given to the quality, as well as the quantity, of the information on which the sentencing judge may rely. Assurances of secrecy are conducive to the transmission of confidences which may bear no closer relation to fact than the average rumor or item of gossip, and may imply a pledge not to attempt independent verification of the information received. The risk that some of the information accepted in confidence may be erroneous, or may be misinterpreted, by the investigator or by the sentencing judge, is manifest.[155]

Gardner was clearly hostile to secrecy, but there is also a question regarding its scope. It could be argued that *Gardner* is limited to capital cases. A proponent of this argument might cite the Court's 1972 opinion in *Morrissey v. Brewer*, which considered "whether the requirements of due process in general apply to parole revocations."[156] The Court not only concluded that due process did apply to that particular action, but also wrote more generally that "[o]nce it is determined that due process applies, the question remains what process is due."[157] The Court explained that "the concept of due process is flexible," reflecting "a recognition that not all situations calling for procedural safeguards call for the same kind of procedure."[158]

[152] *Id.*
[153] *Id.* at 359, 362.
[154] *Id.* at 361.
[155] *Id.* at 359.
[156] Morrissey v. Brewer, 408 U.S. 471, 481 (1972).
[157] *See id.*
[158] *Id.*

Thus, *Morrissey* makes clear that due process can mean different things in different contexts. And it is beyond any doubt that capital cases merit the highest level of due process protections. But that is not the end of the story. After all, it would belie logic to assert that the *Gardner* plurality's statement that "the sentencing process, as well as the trial itself, must satisfy the requirements of the Due Process Clause" is relevant *only* to capital cases.[159] Due process is clearly a requirement of sentencing proceedings in non-capital cases, though what exactly that means is open to debate.

Gardner can be juxtaposed with the multiple circuit court rulings (in non-capital cases) that have underscored that a defendant does *not* have a right to access all information used at sentencing. For instance, in *United States v. Headspeth* in 1988, the Fourth Circuit addressed a defendant's assertion that "his due process rights were violated when he was denied access to the portion of the presentence report that contained the probation officer's sentencing recommendation."[160] The Fourth Circuit disagreed, writing that

> [w]hile a convicted defendant retains a due process right not to be sentenced on the basis of materially false or inaccurate information, access to the sentencing recommendation, which is nothing but a subjective judgment made on the basis of facts contained elsewhere in the report, is not necessary to vindicate that interest.[161]

In *United States v. Baldrich* in 2006, the Ninth Circuit reached a similar conclusion.[162] Baldrich "argue[d] that the district court violated his right to due process at sentencing by denying his motion to disclose the probation officer's confidential sentencing recommendation" and "that Rule 32(e)(3) of the Federal Rules of Criminal Procedure is unconstitutional to the extent it allows the court to withhold the recommendation."[163] The court ruled against Baldrich, writing that "Rule 32's requirement that all facts relevant to the defendant's sentence be provided to the defendant for adversarial testing clearly extends to the factual information underlying a probation officer's confidential sentencing recommendation, even though the

[159] *Gardner*, 430 U.S. at 358.

[160] 852 F.2d 753, 755 (4th Cir. 1988).

[161] *Id.* (internal citation omitted).

[162] 471 F.3d 1110 (9th Cir. 2006).

[163] *Id.* at 1111. Rule 32(e)(3) of the Federal Rules of Criminal Procedure states: "By local rule or by order in a case, the court may direct the probation officer not to disclose to anyone other than the court the officer's recommendation on the sentence." FED. R. CRIM. P. 32(e)(3).

recommendation itself need not be disclosed."[164] The court also found that

> all of the facts in the confidential sentencing recommendation were discussed in the presentence report or in open court at the sentencing hearing. Therefore, the district court's decision not to disclose the confidential recommendation to Baldrich did not violate Rule 32 or Baldrich's due process rights. We also reject Baldrich's argument that the district court must disclose the probation officer's confidential analysis and opinions.[165]

Thus, *Headspeth* and *Baldrich* (and Rule 32(e)(3)) indicate that a defendant does not have a due process right to obtain information about *recommendations* and *opinions* presented to a court in relation to sentencing.[166] But what about the *facts* underlying those recommendations and opinions? Unfortunately, at least some federal appeals courts have suggested that an incomplete disclosure of the facts might be acceptable. As the Ninth Circuit wrote in 2015 in *United States v. Eyraud*, "[t]o date, no circuit . . . has concluded that the Due Process Clause requires full disclosure of all the information relied on by a court at sentencing."[167] A similar view was articulated by the Sixth Circuit in 2007:

> [W]e do not read *Townsend* and its progeny as having clearly established the considerably broader principle . . . that *all* information relied upon by a sentencing court must be disclosed to the defendant, whether or not it is later determined to be materially false. To the contrary, the federal appellate courts that have considered this issue have uniformly concluded that *Townsend* and *Tucker* do *not* recognize such a federal due process right to full disclosure.[168]

The secrecy question is particularly important in relation to AI-based presentence risk assessment algorithms, which will often be proprietary, as the companies that design and sell them want to retain a competitive advantage in the market. By definition, a proprietary AI algorithm is secret, including—at least unless and until a court orders otherwise—from the defendant it is being used to evaluate. A risk assessment algorithm—whether it uses AI or not—takes in data and then performs mathematical analysis of that data in order to provide a

[164] *Baldrich*, 471 F.3d at 1114.

[165] *Id.* at 1114–15.

[166] *See* FED. R. CRIM. P. 32(e)(3) ("By local rule or by order in a case, the court may direct the probation officer not to disclose to anyone other than the court the officer's recommendation on the sentence.").

[167] 809 F.3d 462, 471 (9th Cir. 2015).

[168] Stewart v. Erwin, 503 F.3d 488, 495 (6th Cir. 2007) (emphasis in original).

risk assessment. AI adds the additional wrinkle that the form of mathematical analysis that gets conducted can change over time.

For each of these factors—the input data, the nature of the analysis performed on it, the way that analysis evolves over time, and the resulting risk assessment—the question could be asked regarding a defendant's right of access under due process. The case law discussed above provides at least a reasonable (though certainly not ironclad) argument that the defendant has a due process right to access *facts* considered at sentencing to ensure that they are not materially inaccurate, while also indicating that the defendant may have little or no right to *recommendations* derived from those facts. The input data (e.g., the crime for which the defendant has been convicted, any previous convictions, the defendant's age, etc.) constitute facts. Thus, these data fall in the category of information the defendant should have a right to know, and to challenge if there are inaccuracies.

But what about the algorithm itself, i.e., the nature of the analysis performed on the input data? And what about the AI aspects of the system, which govern the way that analysis evolves over time? Or the risk assessment output by the algorithm and presented to the court at sentencing? Our view is that blocking any of this information from a defendant who seeks it is a violation of due process, as it prevents the defendant from evaluating the potential impact of flaws in the analysis that could lead to a significant overstatement of risk.

As a counterargument to our conclusion, it could be suggested that if an AI algorithm is fair,[169] then the details of how it operates no longer matter. In other words, this counterargument would hold that if a defendant is informed that an algorithm is operating in a way that upholds due process, then he or she no longer has any need to know its inner workings. We are skeptical of this counterargument, for at least the reason that it would require the defendant to rely on the dubious assumption that the person or entity providing the assurances had performed an analysis free from any oversights that might have led to an overly optimistic conclusion regarding algorithm fairness. Another concern is that the dynamic nature of AI means that even an algorithm that was problem-free (however that might be defined) at one point in time might no longer be so in the future. In addition, in the context of predictions used in criminal justice, there is a moral

[169] Of course, the issue of what it means for an algorithm to be "fair" is an entire field of inquiry on its own. Our point here is that if, according to some agreed-upon measure of fairness, an algorithm can be shown to be fair, the argument could be made (though it would suffer from the problems we discuss in the text) that it does not matter how it works internally.

imperative not to foreclose access to the information that can shed light on why a particular prediction is made.

C. Scientific Validity

How valid, accurate, and widely accepted will AI-based presentence risk assessments be? And how does that relate to due process? A case relevant to answering these questions is *Daubert v. Merrell Dow Pharmaceutical, Inc.*, which was a 1993 Supreme Court ruling that considered testimony by expert witnesses.[170] *Daubert* was particularly focused on the question of "whether the expert is proposing to testify to (1) scientific knowledge that (2) will assist the trier of fact to understand or determine a fact in issue."[171] As the Court recognized, this "entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue."[172] While stopping short of offering a "definitive checklist," the Court identified four factors relevant to the inquiry: (1) "determining whether a theory or technique is scientific knowledge that will assist the trier of fact" by examining "whether it can be (and has been) tested," (2) whether it "has been subjected to peer review and publication," (3) the "known or potential rate of error . . . and the existence and maintenance of standards controlling the technique's operation," and (4) whether the technique has experienced general acceptance within the scientific community.[173]

As Justice Blackmun wrote in the first sentence of his opinion, *Daubert* addressed "the standard for admitting expert scientific testimony in a federal trial."[174] *Daubert* was primarily focused on interpreting Federal Rule of Evidence 702, which the *Daubert* court recognized as "a specific Rule that speaks to the contested issue."[175] As noted above, the Federal Rules of Evidence do not apply at sentencing.[176] With respect to techniques used to perform presentencing risk assessments, that means a defendant asserting a

[170] 509 U.S. 579, 579 (1993).

[171] *Id.* at 592.

[172] *Id.* at 592–93.

[173] *See id.* at 593–94.

[174] *See id.* at 582.

[175] *See* FED. R. EVID. 702; *Daubert*, 509 U.S. at 588.

[176] *See* discussion *supra* Section III.A (providing that FED. R. EVID. 1101(d)(3) identifies sentencing as a proceeding to which "these rules—except for those on privilege—do not apply").

right to a strict application of the *Daubert* factors would likely not be successful. But it is nonetheless instructive to examine AI-based risk assessment against that standard to see how it measures up.

In relation to the first factor ("whether it can be (and has been) tested"), AI-based presentencing risk assessment comes up well short.[177] AI-based (as opposed to the more general non-AI algorithm-based) risk assessment is still new. In theory, of course, it is testable, but as far as we are aware there have not yet been any large-scale third-party tests to assess its accuracy or reliability. In relation to the second factor (whether it has been subjected to peer review and publication), AI-based presentence risk assessment again comes up short. The number of peer-reviewed publications on this topic, while not zero, is still low.[178] In noting this, we are in no way suggesting that the peer-reviewed papers that have been published on this topic are not of high quality; rather, the point is that the volume of such papers is still too low to constitute a substantive body of work. In relation to the third factor (the error rate), AI-based presentence risk assessment is a question mark. We simply do not know how error prone it will be, or what sorts of factors might either exacerbate or help to minimize errors. There is also the issue that in algorithmic predictions, "error rate" can take on multiple meanings. For instance, if the prediction is binary, measures of error rate include the false positive rate and the false negative rate. An additional issue is that optimizing an algorithm to reduce an error rate could come at the cost of performance according to other statistical measures such as positive predictive value.[179] Finally, with respect to the fourth factor (general acceptance within the scientific community), AI-based presentence risk assessment has not yet matured to the point where it has experienced anything near general acceptance.

[177] *See Daubert*, 509 U.S. at 593.

[178] *See generally* Han-Wei Liu et al., *Beyond* State v. Loomis*: Artificial Intelligence, Government Algorithmization, and Accountability*, 27 INT'L J.L. & INFO. TECH. 122 (2019) (discussing AI-based presentence risk assessment).

[179] *See* Melissa Hamilton, *Debating Algorithmic Fairness*, 52 U.C. DAVIS L. REV. ONLINE 261, 269 (2019) (explaining that "when base rates between groups differ, the algorithm cannot achieve equal false positive rates and equal positive predictive values at the same time"). In a binary predictor, the positive predictive value refers to the probability that a prediction regarding a "positive" outcome is correct. For example, if an algorithm predicts that a particular set of fifty students will pass a test, and if after the test is administered, forty in that group do in fact pass, the positive predictive value is forty divided by fifty, which is 80%. *See, e.g.*, Virginia Foggo, John Villasenor, & Pratyush Garg, *Algorithms and Fairness*, 17 OHIO ST. TECH. L.J. (forthcoming 2020).

What is clear from the above is that as things currently stand, AI-based presentence risk assessment falls well short of the requirements of every one of the four *Daubert* factors. But does this matter given the lack of applicability of Federal Rule of Evidence 702 to sentencing in federal courts (and, in many states, an analogous lack of evidentiary protection at sentencing)? While the evidentiary rules may be largely irrelevant at sentencing, the *Daubert* factors also relate to accuracy, which is highly relevant at all phases of a criminal proceeding, including sentencing.[180]

As a thought experiment, if prosecutors were to attempt to introduce testimony at a trial from a person claiming to be a mind reader capable of knowing the thoughts in the defendant's mind, the defense would obviously be able to use a *Daubert* motion to exclude the testimony.[181] If the same self-professed mind reader sought to submit a risk assessment for consideration at a sentencing hearing, a defendant would just as clearly have a right to block that information, though he or she would probably need to object on grounds other than *Daubert*. This example matters because risk assessment algorithms (including but not limited to those based on AI) will span a quality spectrum. At one end of the spectrum they could in theory be very good and produce risk assessments that are highly accurate. At the other end of the spectrum they could be flawed—so flawed, in fact, that in the (theoretically) extreme case they might be producing random outputs, just like the purported mind reader.

In practice, no AI-based risk assessment algorithms are likely to be so poorly designed as to produce what amount to random outputs. But that is not the end of the question because it is certainly possible that an algorithm might be pretty badly designed, just not so badly designed as to be completely random. The bottom line is that the due process right to not be sentenced based on materially inaccurate information ought to confer at least some level of protection against the use of underperforming risk assessment algorithms. That, in turn, means that a defendant should have a right to question the scientific validity of the specific risk assessment algorithm used to generate his or her risk scores. Such questions, when they are asked, will then force consideration of whether there should be some standard applied to determine prediction algorithm admissibility.

[180] *See* discussion *supra* Section III.A.

[181] *See* discussion *supra* Section III.C.

D. Consideration of Impermissible Factors

At sentencing, consideration of factors such as race or nationality is prohibited. In *United States v. Onwuemene* in 1991, the Eighth Circuit noted that "[c]onsideration [at sentencing] of Onwuemene's alien status . . . violated his constitutional rights" and ordered that the case be remanded for resentencing "[b]ecause we cannot say that the district court would have imposed the same sentence absent this impermissible consideration."[182] Moreover, the court noted that guidelines from the U.S. Sentencing Commission "state unequivocally that race, sex, national origin, creed, religion, and socioeconomic status 'are not relevant in the determination of a sentence.'"[183]

The protections against the introduction of impermissible factors are no less strong even when the problematic assertions are introduced through the actions of the defendant's own attorney. In 2017, the Supreme Court issued its decision in *Buck v. Davis*, which arose from a Texas case in which Buck was found guilty of murder.[184] At the penalty phase of the trial, in relation to the question of whether Buck would likely commit future acts of violence, Buck's own attorney called a psychologist as an expert witness.[185] As described by the Court, "[t]he psychologist testified that Buck probably would not engage in violent conduct. But he also stated that one of the factors pertinent in assessing a person's propensity for violence was his race, and that Buck was statistically more likely to act violently because he is black."[186] After the jury sentenced Buck to death, Buck filed an appeal, asserting that his attorney's decision to call a witness who cited his race as an indicator of propensity for future violence "violated his Sixth Amendment right to the effective assistance of counsel."[187] In finding in favor of Buck, the Court noted that Buck's attorney had "specifically elicited testimony about the connection between Buck's race and the likelihood of future violence" and concluded that Buck had "demonstrated . . . ineffective assistance of counsel."[188]

[182] 933 F.2d 650, 652 (8th Cir. 1991).

[183] *Id.* at 651 (citing U.S.S.G. MANUAL §5H1.10 (Nov. 1, 1987)).

[184] 147 S. Ct. 759, 767 (2017).

[185] *See id.*

[186] *Id.*

[187] *Id.*

[188] *Id.* at 775, 780.

State courts have also recognized the importance of avoiding the introduction of impermissible factors at sentencing. For example, in *State v. Harris* in 2010, the Wisconsin Supreme Court held that

> [d]iscretion is erroneously exercised when a sentencing court actually relies on clearly irrelevant or improper factors, and the defendant bears the burden of proving such reliance by clear and convincing evidence. It is beyond dispute that race and gender are improper factors; they may not be relied upon—at all—in the imposition of a sentence.[189]

This issue is of particular concern because even when a risk assessment algorithm is designed to specifically avoid considering factors such as race, socioeconomic status, etc., such factors can end up being implicitly included due to the data relied upon to perform the assessment.[190] For example, the Iowa Risk Revised (IRR) tool utilizes inputs regarding (among other things) the defendant's employment and housing status, current conviction, and any previous convictions—all factors that are highly impacted by historical and continuing patterns of racial discrimination.[191] After all, it is well established that African Americans are disproportionately targeted both in policing and in the criminal justice system.[192] As a result, the number of previous felony convictions is correlated with race.[193]

It also acts, to some extent, as a proxy for socioeconomic status (which itself also correlates with race). Whether a person is convicted of a crime for which they have been arrested depends on a multiplicity of factors, including the resources they can put into hiring counsel for his or her defense. To take one example, because public defenders face extremely high caseloads, they can face significant pressure to convince their clients to agree to plea bargains.[194] We have no doubt

[189] *Id.* at 411.

[190] *See generally* BETH SKINNER ET AL., 2016 IOWA STATE BAR ASS'N ANNUAL MEETING - EVOLVING TRENDS IN IOWA'S CORRECTIONAL PRACTICES, IOWA DEPT. CORR. 16 (IOWA DEPT. CORR. 2016) (discussing how seemingly neutral factors can be discriminatory).

[191] *See id.*

[192] *See* Radley Balko, *There's Overwhelming Evidence that the Criminal-Justice System Is Racist. Here's the Proof*, WASH. POST (Sept. 18, 2018), https://www.washingtonpost.com/news/opinions/wp/2018/09/18/theres-overwhelming-evidence-that-the-criminal-justice-system-is-racist-heres-the-proof [https://perma.cc/PX7V-XSHC].

[193] *See* Timothy Williams, *Black People Are Charged at a Higher Rate Than Whites. What if Prosecutors Didn't Know Their Race?*, N. Y. TIMES (June 12, 2019), https://www.nytimes.com/2019/06/12/us/prosecutor-race-blind-charging.html [https://perma.cc/VBV6-CENB].

[194] *See, e.g.*, Jaeah Lee, Hannah Levintova, & Brett Brownell, *Charts: Why You're in Deep Trouble if You Can't Afford a Lawyer*, MOTHER JONES (May 6, 2013),

that the companies that develop and sell AI-based systems for presentencing risk assessment will endeavor to avoid explicit consideration of prohibited factors such as race, nationality, or citizenship status. But given the extent to which this information is indirectly reflected through correlations with data that might be used by such systems, there is a very real concern that these data might act as an inadvertent back door through which prohibited factors might be considered, thereby raising due process concerns. As Huq has written, "the primary reason for concern with racial equity in the algorithmic criminal justice context is that efforts to suppress crime entrench wider social patterns of racial stratification."[195] These patterns, Huq writes, are in significant part attributable to "the asymmetrical spillovers from criminal justice for minority but not majority populations."[196]

IV. Algorithmic Risk Assessments: Some Key Cases

Having provided the broader context for the case law relating to information used at sentencing, we now discuss some of the cases in which the use of risk assessment algorithms was itself the reason for a due process challenge. Importantly, at least as far as the public record indicates, the cases discussed in this section involved algorithms, but it did not involve *artificial intelligence* algorithms. As noted earlier, the dynamic nature of AI adds a significant level of both technical and legal complexity. That said, it is still very instructive to consider the case law on non-AI-based algorithms, and to consider what it might mean for the future when the due process challenges specific to AI come before the courts.

A. *State v. Loomis*

State v. Loomis is (at least to date) the most on-point case involving secrecy of risk assessment algorithms used in relation to

https://www.motherjones.com/politics/2013/05/public-defenders-gideon-supreme-court-charts/ [https://perma.cc/MW2G-K37Q] (stating that "many well-meaning defense lawyers are sucked into a 'meet 'em and plead 'em' routine (PD parlance for meeting clients just a few minutes or hours before their hearings and then encouraging them to admit guilt just to get rid of the case). It's a large reason why 90 to 95 percent of their clients plead guilty, says Tanya Greene, an ACLU attorney and capital public defender. 'You've got so many cases, limited resources, and there's no relief,' she says. 'You go to work, you get more cases. You have to triage.'").

[195] *See* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1055 (2019).

[196] *See id.*

sentencing.[197] *Loomis* arose from a drive-by shooting that occurred in La Crosse County, Wisconsin in February 2013.[198] Prosecutors asserted that Eric Loomis was the driver of the car and charged him with five criminal counts.[199] Loomis denied being at the scene of the crime, stating that he had driven the car after, but not during the shooting.[200] Nonetheless, he pled guilty to two of the counts: "Attempting to Flee or Elude a Traffic Officer" and "Operating a Motor Vehicle without Owner's Consent."[201] The plea agreement provided that the "other counts will be dismissed and read in for sentencing."[202]

The trial court accepted Loomis's pleas, and in July 2013, the Wisconsin Department of Corrections submitted a presentence investigation report (PSI) that included the output from the COMPAS (Correctional Offender Management Profiling for Alternative Sanction) risk assessment tool.[203] As the Wisconsin Supreme Court later explained,

> COMPAS is a risk-need assessment tool designed by Northpointe, Inc. to provide decisional support for the Department of Corrections when making placement decisions, managing offenders, and planning treatment. The COMPAS risk assessment is based upon information gathered from the defendant's criminal file and an interview with the defendant.
>
> A COMPAS report consists of a risk assessment designed to predict recidivism and a separate needs assessment for identifying program needs in areas such as employment, housing and substance abuse. The risk assessment portion of COMPAS generates risk scores displayed in the form of a bar chart, with three bars that represent pretrial recidivism risk, general recidivism risk, and violent recidivism risk. Each bar indicates a defendant's level of risk on a scale of one to ten.[204]

Notably, COMPAS uses a proprietary algorithm.[205] Thus, while the output scores are available to a court, prosecutors, and the defendant, the specific ways in which input data were combined to create that output are not included in the report.[206] At the August 2013

[197] *See generally* State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

[198] *See* Petition for Writ of Certiorari, *Loomis*, 137 S. Ct. 2290 (No. 16-6387) [hereinafter *Loomis cert.*].

[199] *See id.*

[200] *See id.*

[201] *See id.*

[202] *Loomis*, 881 N.W.2d at 754.

[203] *See id.*, *cert. denied*, 137 S. Ct. 2290 (2017).

[204] *Loomis*, 881 N.W.2d at 754 (internal citations omitted).

[205] *See id.* at 753.

[206] *See id.*

sentencing hearing, the state specifically referenced the COMPAS output, arguing that "the COMPAS report that was completed in this case does show the high risk and high needs of the defendant. There's a high risk of violence, a high risk of recidivism, high pre-trial risk; so all of these are factors in determining an appropriate sentence."[207] The judge also referenced COMPAS, saying:

> You're identified, through the COMPAS assessment, as an individual who is at high risk to the community. In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend.[208]

After Loomis was sentenced to a multiyear prison sentence, he filed a motion for post-conviction relief, which the trial court denied.[209] Loomis then appealed to the Wisconsin Court of Appeals, District IV, which in turn certified the appeal to the Wisconsin Supreme Court, specifying the question of whether using COMPAS at sentencing "violates a defendant's right to due process, either because the proprietary nature of COMPAS prevents defendants from challenging the COMPAS assessment's scientific validity, or because COMPAS assessments take gender into account."[210]

In a July 2016 opinion, the Wisconsin Supreme Court ruled against Loomis.[211] The court conceded that "[i]f a COMPAS risk assessment were the determinative factor considered at sentencing this would raise due process challenges regarding whether a defendant received an individualized sentence," but then concluded that "although the circuit court mentioned the COMPAS risk assessment, it was not determinative in deciding whether Loomis should be incarcerated, the severity of the sentence or whether he could be supervised safely and effectively in the community."[212] The court asserted that its "not determinative" conclusion was justified because the trial court "explained that its consideration of the COMPAS risk scores was supported by other independent factors"[213] and because those scores were used by the trial court as "'an observation' to reinforce its assessment of the other factors it considered."[214]

[207] *Id.* at 755.

[208] *Id.*

[209] *See id.* at 756 n.18, 757.

[210] *See id.* at 753.

[211] *See id.*

[212] *Id.* at 764, 771.

[213] *Id.* at 753.

[214] *Id.* at 770.

The Wisconsin Supreme Court also addressed the use of gender information in COMPAS, writing that "there is a factual basis underlying COMPAS's use of gender in calculating risk scores," that "any risk assessment tool which fails to differentiate between men and women will misclassify both genders," and concluding "[t]hus, if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose."[215]

The Wisconsin Supreme Court also considered the more general question of the appropriateness of the use of COMPAS at sentencing, writing:

> [A] sentencing court may consider a COMPAS risk assessment at sentencing subject to the following limitations. As recognized by the Department of Corrections, the PSI instructs that risk scores may not be used: (1) to determine whether an offender is incarcerated; or (2) to determine the severity of the sentence. Additionally, risk scores may not be used as the determinative factor in deciding whether an offender can be supervised safely and effectively in the community.
>
> Importantly, a circuit court must explain the factors in addition to a COMPAS risk assessment that independently support the sentence imposed. A COMPAS risk assessment is only one of many factors that may be considered and weighed at sentencing. Any Presentence Investigation Report ('PSI') containing a COMPAS risk assessment filed with the court must contain a written advisement listing the limitations.[216]

In October 2016, Loomis filed a petition with the U.S. Supreme Court, asking the Court to consider the question:

> Is it a violation of a defendant's constitutional right to due process for a trial court to rely on such risk assessment results at sentencing: because the proprietary nature of COMPAS prevents a defendant from challenging the accuracy and scientific validity of the risk assessment; and . . .because COMPAS assessments take gender and race into account in formulating the risk assessment.[217]

In opposing the petition, the State of Wisconsin identified the question presented as "[d]oes a trial court violate a criminal defendant's due-process rights at sentencing when it considers—but does not base the sentence upon—an evidence-based recidivism-risk assessment?"[218] The State argued that "[t]he use of risk assessments by sentencing courts is a novel issue, which needs time for further percolation" and that in *Loomis*, the trial court "merely used COMPAS

[215] *Id.* at 766.

[216] *Id.* at 769.

[217] *Loomis cert.*, *supra* note 198.

[218] Brief in Opposition at i, Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (No. 16-6387).

to 'corroborat[e]' the evaluation it had already made using the sentencing factors, and was well aware of COMPAS's limitations."[219]

The United States also submitted an *amicus* brief opposing the petition, conceding that a "sentencing court's use of actuarial risk assessments raises novel constitutional questions that may merit this Court's attention in a future case."[220] But the brief also argued that "this case is an unsuitable vehicle for review because the Wisconsin Supreme Court concluded, as the sentencing judge stated, that the trial court would have imposed the same sentence absent any consideration of petitioner's COMPAS risk scores. The petition therefore should be denied."[221] In June 2017, the U.S. Supreme Court denied the petition.[222]

One of the stranger paradoxes illustrated by *Loomis* concerns assertions that due process was not implicated because the risk assessments purportedly had no impact on sentencing. The Wisconsin Supreme Court wrote that "although the circuit court mentioned the COMPAS risk assessment, it was not determinative in deciding whether Loomis should be incarcerated, the severity of the sentence or whether he could be supervised safely and effectively in the community."[223] To suggest that the inclusion of a risk assessment in a PSIR is acceptable only when it has no impact on sentencing leaves open the question of why the risk assessment was used at all. After all, the main goal of a sentencing proceeding is to determine whether and for how long a convicted defendant should be incarcerated—or, alternatively, whether the defendant can be effectively supervised in the community.

B. *Malenchik v. State*

Malenchik v. State was a 2010 Indiana Supreme Court ruling that upheld the use of risk assessment tools at sentencing.[224] Before discussing the specifics of *Malenchik*, it is relevant to note that two years earlier, in *Rhodes v. State*, the Indiana Court of Appeals had upheld a defendant's challenge to the use of the Level of Service Inventory-Revised (LSI-R) risk assessment tool in relation to

[219] *Id.* at 1, 9 (citations omitted).

[220] Brief for the United States as Amicus Curiae at 12, *Loomis*, 137 S. Ct. 2290 (2017) (No. 16-6387).

[221] *Id.* at 13.

[222] *See* Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (denying cert.).

[223] 881 N.W.2d 749, 771 (Wis. 2016).

[224] *See generally* Malenchik v. State, 928 N.E.2d 564 (Ind. 2010).

sentencing.[225] In *Rhodes*, the Indiana Court of Appeals had found that "[t]he use of a standardized scoring model, such as the LSI-R, undercuts the trial court's responsibility to craft an appropriate, individualized sentence" and concluded that it was "an abuse of discretion to rely on scoring models to determine a sentence."[226]

In *Malenchik*, the Indiana Supreme Court cited the Indiana Court of Appeals' conclusion in *Rhodes* and then wrote,

> [w]e disagree. As noted above, there is a growing body of impressive research supporting the widespread use and efficacy of evidence-based offender assessment tools. The results of such testing can enhance a trial judge's individualized evaluation of the sentencing evidence and selection of the program of penal consequences most appropriate for the reformation of a particular offender . . . We defer to the sound discernment and discretion of trial judges to give the tools proper consideration and appropriate weight. We disapprove of the resistance to LSI-R test results expressed by the Court of Appeals in *Rhodes*.[227]

C. *People v. Younglove*

People v. Younglove was a 2019 decision from the Court of Appeals of Michigan and arose from a set of consolidated appeals challenging the use of COMPAS.[228] The defendants asserted "their respective sentencing courts' presumed use of the COMPAS information in the PSIRs when determining their sentences deprived them of the due process of law."[229] More specifically, the defendants asserted that "because COMPAS statistically analyzes data from a general population in making its determinations . . . its use [is] inappropriate in an individualized sentencing decision," and that COMPAS "has discriminatory impacts as concerns race and gender inputs, and that the scores it produces lack transparency."[230] The court rejected the due process claim, writing that

> the references to COMPAS in defendants' PSIRs, despite reflecting a software program's projections about future behavior, are similar to the opinions of probation agents that are routinely included in PSIRs . . . Therefore, we are not persuaded by defendants' arguments that the

[225] *See generally* Rhodes v. State, 896 N.E.2d 1193 (Ind. Ct. App. 2008).

[226] *Id.* at 1195.

[227] *Malenchik*, 928 N.E.2d at 573.

[228] *See* People v. Younglove, No. 341901, 2019 WL 846117, at *1 (Mich. Ct. App. Feb. 21, 2019).

[229] *Id.* at *2.

[230] *Id.*

> inclusion of COMPAS information unfairly influences or replaces a sentencing court's individual sentencing discretion.[231]

V. AI AND CRIMINAL RISK ASSESSMENT: THREE GUIDING PRINCIPLES

In this section we introduce three principles that we believe should guide the use of AI in presentence risk assessments. As discussed above, due process requires that a defendant not be sentenced based on materially inaccurate information. Securing that right requires ensuring that defendants have at least the possibility of pursuing a claim of material inaccuracy associated with either the data used to compute an algorithmic risk score or the algorithm used to compute risk scores based on that data. That, in turn, means that the algorithm and data associated with generating risk assessments must be archived and, as appropriate, made available to defendants. These requirements lead directly to the principles of auditability and transparency, both of which pertain to snapshots of how an AI system operates in relation to *individual* defendants.

There is also the question of how the algorithm operates in relation to defendants as a group, and in particular whether there is evidence that might support a claim of either (1) material inaccuracy or (2) bias arising from impermissible use of factors such as race.[232] This leads to the principle of consistency. The following sections describe each of these three principles in more detail.

A. Principle #1: Auditability

Auditability, as the term is used in this Article, refers to preserving all of the information that was used to perform a risk assessment so that it can potentially be accessed in the event of a due process challenge. This involves three distinct categories of information.[233] The first category is non-defendant-specific data that is

[231] *Id.*

[232] *See, e.g.*, United States v. Onwuemene, 933 F.2d 650, 651 (8th Cir. 1991) (noting that guidelines from the U.S. Sentencing Commission "state unequivocally that race, sex, national origin, creed, religion, and socioeconomic status 'are not relevant in the determination of a sentence'").

[233] There is also a fourth category of information, which is the output of the assessment tool. However, we are assuming that the output would already be available to the defendant. To the extent that assumption does not hold, the output of the assessment tool would need to be recorded along with the other three categories of information described in this section.

used by the algorithm. For instance, this could include the historical recidivism rate for people found guilty of similar crimes in the state where the trial took place. A statistic like this might evolve slowly (e.g., if it is recalculated once a year and conveyed to the AI system through an annual upgrade) or quickly (if the AI system monitors databases of trials and convictions and automatically updates the recidivism rate data in response to information about new convictions).

The second category of information is data specific to the defendant, such as the defendant's arrest history.[234] In providing this example, we are not offering an opinion that the number of prior felony convictions is necessarily a bias-free input to a risk assessment algorithm. In fact, there is a good case to be made that it is problematic, given that it can be influenced by factors (e.g., racial bias, the level of policing in the neighborhood where the defendant lives, etc.) that are not specific to the individual being evaluated.

The third category of information is the algorithm itself—i.e., the specific mathematical and logical processes that were used to combine the data in the first two categories in order to produce an assessment. Since an AI algorithm can evolve very quickly, if there is no specific effort to record the state of the algorithm every time it is used, it could be difficult or impossible to reconstruct it at a later time. In sum, the "snapshot" that would be recorded each time the algorithm is run would comprise a complete tally of all three categories of information, thus enabling the exact same assessment, using the exact same data, to be performed again at a later time, with a guarantee that it would produce the same output.

In relation to auditability, it is also relevant to consider practical issues relating to how and by whom this information would be stored. One initial question is whether the resulting file sizes would be problematic. Our view is that this would not be an issue. As noted above, there are three categories of information. The first category (non-defendant-specific data) and second category (defendant-specific data) would consume only a very small storage volume in light of current and future storage capacities. Many algorithms are likely to use no more than (and often much fewer than) a few dozen inputs in each of these categories.[235] The total data volume needed to

[234] *See, e.g.*, NORTHPOINTE, INC., PRACTITIONER'S GUIDE TO COMPAS CORE 37 (2019) (noting the use of "arrest history").

[235] Practical limits on the number of inputs will be imposed by the need to collect and input all of these data into a risk assessment algorithm. In theory this is a

record those parameters would be negligible. To take a specific example, suppose that a particular algorithm used 100 distinct pieces of information across the two categories of non-defendant-specific data and defendant-specific data. Suppose further that each piece of information required an average of forty bytes to store. This corresponds to a total of four thousand bytes—which is essentially free to store given today's technology environment.

One way to store an algorithm is to store the complete set of code for implementing it.[236] The advantage of storing an algorithm this way is that it can easily be run again either to replicate its performance on inputs that it had used in the past or to explore its performance on new inputs.[237] The downside of storing algorithms in the form in which they are actually implemented on computers is that code can be very difficult for humans to read. This is particularly the case for code that was written not by humans but by other code, which will occur with increasing frequency as AI-based systems become more advanced. This is because AI algorithms will, in some cases, quite literally write their own code, and will of course do so without regard for how human-readable it might be.

An alternative approach is through the use of "pseudocode," which is commonly used in the field of computer science to record algorithms in a form that is human readable and also easy to convert back into computer code should the need arise.[238] Of course, since pseudocode is a *representation* of the actual code, it would be important to ensure that there was no divergence—i.e., that the pseudocode accurately represents what the actual code did. Whether an algorithm was stored directly by simply archiving a literal copy of the code or indirectly by automatically generating and then storing pseudocode, the data volume involved in storing the algorithm would likely be higher than that needed to store the first two categories of information (the non-defendant-specific data and defendant-specific

process that could be automated but in practice the requisite data will often require some level of human intervention to collect.

[236] In referring to the "complete set" of code, we mean all of the programs, and all of the associated parameters, etc. that are needed to make it run.

[237] *See generally* Harry Surden, *Computable Contracts*, 46 U.C. DAVIS L. REV. 629 (2012).

[238] *See, e.g.*, *Designing an Algorithm*, BBC, https://www.bbc.co.uk/bitesize/guides/z3bq7ty/revision/2 [https://perma.cc/R87G-VBA9] (last visited June 1, 2020) (explaining that "[p]seudocode is not a programming language, it is a simple way of describing a set of instructions that does not have to use specific syntax").

data), but still very modest in a world where it is now possible to purchase a four terabyte external disk drive for under $100.[239]

While data volume (and therefore storage capacity or cost) would not be an impediment, there is a separate question about where this information would be stored. One approach is for companies that produce the risk assessment software to also take responsibility for archiving the data and algorithms. This has the advantage of placing responsibility for managing storage with the entity that knows the software and data best, but it also raises complex issues of compliance and verification. Another (non-mutually exclusive) possibility would be for this information to become a standard digital appendix to presentence investigation reports. This would ensure that the information was directly linked to a defendant's file. Companies would undoubtedly object to the storage of algorithm information in unencrypted form, but this objection could be mitigated through the use of encryption and a suitable accompanying plan for managing decryption keys.

Yet another potential issue is what Selbst and Barocas have termed algorithmic "inscrutability," which they describe as "a situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension."[240] It is certainly possible that when AI-based systems (including those used for risk assessment) become advanced enough, they will acquire a complexity that creates substantive challenges to understanding what they are doing. We are cautiously optimistic that, at least in the relatively near time horizon, and in the very narrow domain we are addressing, this will not be an insurmountable challenge. As an initial matter, many AI-enabled risk assessment algorithms will not be so complex as to elude human scrutiny. And for the subset of those that are, it will not always be necessary to understand every single thing that they are doing. It might be sufficient to obtain an understanding that, while not fully complete, is sufficient to confirm or rebut a particular due process concern.

It is also possible to use algorithms to evaluate algorithms. In other words, an algorithm that might appear to be inscrutable under direct inspection by a human might, when viewed by another

[239] For example, as of June 2020, Staples sells a four-terabyte external hard drive for $94.99. *See* STAPLES, https://www.staples.com/Seagate-Expansion-Portable-Hard-Drive-4TB-Black/product_2095130 [https://perma.cc/87DG-XGQE] (last visited June 1, 2020).

[240] Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1094 (2018).

algorithm specifically designed to understand complex systems, be more amenable to analysis. More generally, while recognizing the importance of both research and scholarship relating to the broader issue of highly complex, machine-generated algorithms, our call for auditability is somewhat more modest: We are arguing that the information about an algorithm used to perform a presentence risk assessment should be archived, thus creating at least the *possibility* that it can be analyzed at a later date in relation to due process concerns. And, in the specific context of algorithmic criminal risk assessment, we are confident that for at least the near-term future, the primary obstacles for algorithmic understanding will lie in the practical questions relating to how the algorithm is stored and who has access to it.

B. Principle #2: Transparency

While auditability refers to ensuring that the information used in computing a risk assessment is archived, transparency refers to ensuring that it can be accessed.[241] The broader concept of algorithm transparency is not new, though there are some domain-specific features for how it would need to operate in the criminal justice context. Companies that produce risk assessment software will have very reasonable concerns about preserving the trade secret aspects of their algorithms, as those might be key factors in ensuring an advantage over competitors in the market. Thus, they would legitimately object to a requirement that the audit information simply be published on the internet. Yet it would also be unreasonable for companies to assert that trade secret protection precludes *any* possibility of access by a defendant to the audit information, as this would surely violate a defendant's right to due process. Put simply,

[241] *See* Kartik Hosanagar & Vivian Jair, *We Need Transparency in Algorithms, but Too Much Can Backfire*, HARV. BUS. REV. (July 23, 2018), https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire [https://perma.cc/72FK-WE6D]. By "information used in computing a risk assessment," we are referring not only to the input data but also to the way in which that input data is processed to produce the output risk assessment. *See id.*; *see also* EUROPEAN PARLIAMENTARY RESEARCH SERVICE, A GOVERNANCE FRAMEWORK FOR ALGORITHMIC ACCOUNTABILITY AND TRANSPARENCY (2019); ACM U.S. PUB. POLICY COUNCIL, STATEMENT ON ALGORITHMIC TRANSPARENCY AND ACCOUNTABILITY (2017); Kartik Hosanagar, *People Want to Know About Algorithms—but Not Too Much*, WIRED (Mar. 12, 2019), https://www.wired.com/story/book-excerpt-algorithm-transparency/ [https://perma.cc/2AHT-C4J6].

trade secret rights of companies that make risk assessment software should not come at the expense of due process rights of defendants.[242]

One way to resolve the tension between the due process rights of defendants and the trade secret rights of companies that make risk assessment software is through tools such as protective orders, which are commonly used in relation to trade secrets in civil litigation.[243] While this offers a procedural way forward, mapping the methods used to protect trade secrets in civil litigation into a criminal context would be financially prohibitive for many defendants. For example, it is common in patent infringement cases for attorneys and technology experts working on behalf of the plaintiff to gain access to a defendant's data, source code, and other trade secret information.[244] The protective orders in such cases are often quite elaborate and complying with them can require significant resources.[245] To take one example, protective orders in patent cases where source code is alleged to infringe a patent often require that the defendant produce source code on a standalone, non-internet-connected computer so that experts working on behalf of a plaintiff can inspect or review it. Typically, the dedicated room is located at the offices of the law firm representing the defendant. In many instances the source code computer must be made available for long periods of time (weeks or months). In some cases, at any time when a representative of the plaintiff is in the room with the source code computer, a staff member from the defendant's law firm must be assigned to be physically present in an adjacent room.

[242] Relatedly, as described in a news item published by the Electronic Privacy Information Center, in 2019 "Idaho became the first state to pass a law specifically promoting transparency, accountability, and explainability in pre-trial risk assessment tools." *Idaho Enacts Law Requiring Transparency in Pre-Trial Risk Assessments*, EPIC (Mar. 28, 2019), https://epic.org/2019/03/idaho-enacts-law-requiring-tra.html [https://perma.cc/YK8X-5P9T].

[243] We are not the first to raise the possibility of using protective orders in this context. *See* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1409–10 (2018) ("For trade secret evidence that satisfies the criminal discovery or subpoena requirements, courts can mitigate any risk from disclosure by using protective orders, sealing orders, and limited courtroom closures.").

[244] In patent litigation, persons such as outside counsel and technical experts working on behalf of the plaintiff gain access to the source code and other relevant trade secret information of the company accused of infringement, but employees of the company bringing the complaint typically do not receive access. *See id.*

[245] *See* Elizabeth Miles, *Protective Orders: Does Yours Cover All the Bases?*, NAT'L L. REV. (Sept. 24, 2019), https://www.natlawreview.com/article/protective-orders-does-yours-cover-all-bases [https://perma.cc/Z3MW-6PYQ].

While the costs for the resources described above can be borne by companies engaged in patent litigation, they would clearly be prohibitive for many defendants in criminal cases. But such costs would be within the capacity of civil rights and criminal justice advocacy organizations. One solution could therefore be for such an organization to underwrite the costs of pursuing a due process claim, providing the resources to cover the costs of complying with the protective order. A successful due process claim would benefit not only the particular defendant(s) on behalf of whom the claim was initially brought, but also a much larger group of defendants who were subject to risk assessments using the same or similar algorithm approaches.

We recognize that a requirement to make proprietary information accessible—even with the safeguards conferred by protective orders—will not be welcomed by companies that make risk assessment tools.[246] As Rebecca Wexler wrote in a 2018 law review article, "Developers often assert that details about how their tools function are trade secrets. As a result, they claim entitlements to withhold that information from criminal defendants and their attorneys, refusing to comply even with those subpoenas that seek information under a protective order and under seal."[247] However, just as routinely occurs in certain types of civil litigation involving protective orders to protect trade secrets, a company that sells risk assessment solutions and has trade secrets to protect can be compelled by the court to produce the relevant information.[248] Thus, we are not arguing that trade secret law should be inapplicable in this context; to the contrary, we believe that companies that make risk assessment software have the right to safeguard their trade secret rights, just as do companies in other markets. When proper protections are in place, it is well established in other areas of law that trade secret rights can be protected while still conferring controlled, limited access by outside parties.[249] The same should hold true in relation to algorithms underlying risk assessments.

[246] *See* Wexler, *supra* note 243, at 1349–50.

[247] *See id.*

[248] *See id.* at 1346–47. We note that, in other domains, the company making disclosures under a protective order has substantial latitude to provide input to the court regarding the terms of that order. *See id.* This can help ensure that disclosures are managed in a manner that preserves the non-public status of the relevant information. *See id.*

[249] *See id.* at 1369.

It is also important to address another concern that might arise in relation to calls for transparency: Could the availability of information about an AI-based risk assessment algorithm be used to game the system? In other words, could defendants with foreknowledge of the inner workings of the algorithm attempt to construct a set of input data aimed at generating a more favorable assessment than is objectively merited? In theory, this could occur, but the opportunities for this type of manipulation can be minimized through appropriate algorithm design approaches. For instance, one technique that would help would be to ensure that the required algorithm inputs are forms of data that are not overly subjective and thus more easily susceptible to manipulation.

Finally, it is also important to acknowledge that transparency has limits.[250] Consider a defendant who asks how an algorithm works and is then given tens of thousands of lines of source code. Under some definitions this might satisfy a transparency requirement, but it could still leave the defendant largely in the dark about the operation of the algorithm.[251] As Joshua Kroll *et al.* have observed, while "full or partial transparency can be a helpful tool for governance in many cases . . . transparency alone is not sufficient to provide accountability in all cases."[252] Similarly, Barocas and Selbst have written that "[t]he problem . . . is greater than disclosure; in the absence of the specialized knowledge required to understand source code, disclosure may offer little value to affected parties and regulators."[253] Mike Ananny and Kate Crawford have observed that "transparency is an inadequate way to understand—much less govern—algorithms."[254] Cynthia Stohl *et al*. have written that one consequence of transparency requirements could be "strategic opacity" achieved through releasing very large amounts of information.[255] As a result, "unimportant pieces of information will take so much time and effort to sift through that receivers will be distracted from the central information the actor wishes to conceal."[256] The skepticism about transparency conveyed in the foregoing is

[250] *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 657–58 (2017).

[251] *See id.*

[252] *Id.*

[253] Selbst & Barocas, *supra* note 240, at 1093–94.

[254] Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 1, 2 (2016).

[255] Cynthia Stohl et al., *Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age*, 10 INT'L. J. COMM. 123, 133 (2016).

[256] *Id.* at 133–34.

warranted. However, whatever the shortcomings of transparency may be, it certainly provides much more information than algorithmic secrecy. In short, while it may be possible to imagine circumstances under which due process remains elusive even in the presence of transparency, it is hard to see how due process can be present when a defendant is denied access to the workings of the algorithm.

C. Principle #3: Consistency

The principle of consistency aims to identify and avoid the issue of substantively different prediction outputs arising from similar inputs that occur at different times. This concern arises due to the dynamic nature of AI algorithms, which can automatically change over time as they learn.[257] As a consequence, this means that the same inputs, presented to the system at two different times, might lead to different outputs. To see why this is a potential due process concern, consider two different defendants with identical (with regard to all of the defendant-specific inputs to the risk assessment tool) profiles who are subject to algorithmic risk assessments. The first defendant is evaluated in March, and the second defendant is evaluated six months later in October.

In "traditional" (non-AI-based) algorithmic risk assessments, the risk score in both cases will be the same, because the underlying algorithm will be the same.[258] But with AI-based systems, the algorithm might have evolved. Suppose that due to the learning over the intervening six months, the algorithm improves. Suppose further that the first defendant is scored as high risk and the second defendant, who is scored using the improved version of the algorithm, is classified as medium risk. The first defendant would have a due process claim arising from a material inaccuracy (the "high risk" classification), since he or she was in effect penalized by being evaluated using an earlier, less advanced (and accurate) version of the algorithm. In making this observation, we are not suggesting that improvements in the accuracy of an algorithm are a bad thing—in fact,

[257] *See, e.g.*, Or Shani, *From Science Fiction to Reality: The Evolution of Artificial Intelligence*, WIRED (2015), https://www.wired.com/insights/2015/01/the-evolution-of-artificial-intelligence/ [https://perma.cc/7WJK-DGM3] (explaining that "[w]eather forecasts, email spam filtering, Google's search predictions, and voice recognition, such Apple's Siri, are all examples" of AI, and that "[w]hat these technologies have in common are machine-learning algorithms that enable them to react and respond in real time").

[258] *See* Villasenor, *supra* note 20.

it is obviously beneficial if an algorithm can become more accurate. However, defendants should be able to seek relief if they were assessed by an algorithm that, with the benefit of hindsight, gave them a materially higher risk score than they deserved.

In calling for consistency, we are not suggesting that, every single time any minor algorithmic update occurs, it is necessary to go back and reexamine and recompute every prior risk assessment that was computed by earlier versions of the algorithm. In the limiting case of dynamic AI algorithms that self-update as frequently as every few days, this would lead to a completely impractical result of essentially requiring a continuous recalculation of all prior risk scores. Rather, we are highlighting the importance of being attentive to truly large-scale changes in the algorithms that have highly consequential impacts on the resulting risk assessments—and therefore, potentially on the sentences that were handed down at proceedings that considered those assessments. To take an extreme example, suppose that a major software flaw that substantially overstated risk was identified in an algorithm. Of course, one necessary step would be to fix the flaw. But an equally necessary step would be to examine previous assessments made with the old software to see how those assessments would have been different had the new software been used at the time of sentencing.

We are also cognizant of the point that, as Richard Re has observed, "the courts are heavily motivated by an aversion to 'technological exceptionalism.'"[259] In other words, why should courts make an exception to address technology-induced inconsistencies, particularly in light of the many human-induced inconsistencies that undoubtedly arose in the past in relation to pre-algorithmic presentence risk assessments? We certainly would not argue that courts should be required to track and react to every tweak in a risk-assessment algorithm that might generate results displaying some level of inconsistency with previous outputs. But we would also argue that while algorithmic approaches certainly have potential disadvantages including those discussed herein, one of their clear advantages is that their digital nature makes it far easier to identify inconsistencies. If those inconsistencies become sufficiently glaring, we think it would be a mistake for courts to turn a blind eye to them simply on the grounds that addressing them would represent a departure from approaches developed in a pre-digital era.

[259] E-mail from Richard Re, Professor of Law, UCLA, to John Villasenor (Aug. 11, 2019, 10:43 PDT) (on file with author).

To examine consistency, it is not necessary to record the algorithm. It is only necessary to have a record of the input data and output risk scores for a large number of defendants over time. Such a record should be generated and retained so that it can be examined in the future to screen for some of the potential concerns identified above. Compliance with this principle would not raise the same level of trade secret concerns that arise in relation to the transparency principle discussed above, since it does not require any disclosure of the algorithm. Recording this input/output information will also make it easier to study the impact of potential correlations in the data used to produce risk assessments.

It is important to emphasize that consistency in the sense discussed here is related to but distinct from the concept of fairness. We have not specifically identified "fairness" as one of the principles because it is self-evident that algorithms used in relation to sentencing (and more generally) should be fair. Of course, there is enormous room for debate regarding what it means for an algorithm to be fair and how fairness should be measured. Our call for consistency is more modest, in that it recognizes the problems that would arise if a single set of input data, if applied as input to a prediction algorithm repeatedly on multiple occasions, could give rise to a variety of different output criminal risk scores just based on the randomness of when the prediction is performed.

That said, the same tools to compare large sets of inputs with their corresponding outputs that can be used to evaluate consistency can also be valuable in providing insights into fairness. As Chander has observed, "Even a transparent, facially neutral algorithm can still produce discriminatory results. What we need instead is a transparency of inputs and results, which allows us to see that the algorithm is generating discriminatory impact."[260] Chander has also written that "the problem is not the black box, which is often more neutral than the human decisionmaker it replaces, but the real world on which it operates."[261] We would go further, arguing that the potential for bias exists not only in the data but also in an initially unbiased algorithm that might evolve in the future in a manner that introduces bias into the calculations. We also underscore that this sort of evolution is not inevitable. But the fact that it is possible means that measures to detect it need to be in place.

[260] Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1024–25 (2017) (emphasis omitted).

[261] *Id.* at 1025.

The records created in relation to the consistency principle would also, as an ancillary benefit, provide information that would be useful if there are concerns that an algorithm is making decisions that could run afoul of antidiscrimination frameworks that prohibit disparate treatment and disparate impact. AI uses data to identify correlations—including subtle correlations that can only be teased out through highly detailed mathematical analysis—to exploit in furthering an objective such as the accuracy of a prediction.[262] Correlations observed in the data may at times reflect a true underlying attribute relating to propensity towards recidivism, and therefore may be properly considered in a risk assessment algorithm. Or, they may be deceptive consequences of a combination of factors well outside the scope of data available to the AI system. In a pre-digital context, a caution against an overreliance on purported correlations can be found in the Supreme Court's 1976 *Craig v. Boren* decision, which considered an Oklahoma statute codifying a gender-based differential regarding the minimum age to purchase alcoholic beverages.[263] The Court found the statute to be in violation of the Fourteenth Amendment, holding that "the principles embodied in the Equal Protection Clause are not to be rendered inapplicable by statistically measured but loose-fitting generalities concerning the drinking tendencies of aggregate groups."[264] There are many "loose-fitting generalities" that might be inferred from the massive amounts of data that are now available regarding nearly everyone, including people who are in the criminal justice system. The information recorded under the "consistency" principle can help serve as a bulwark against improper uses of apparent correlations.

D. The Three Principles and Due Process

As the Supreme Court explained in 1976 in *Mathews v. Eldridge*, "Procedural due process imposes constraints on governmental decisions which deprive individuals of 'liberty' or 'property' interests within the meaning of the Due Process Clause of the Fifth or Fourteenth Amendment."[265] *Mathews* arose from a challenge to the constitutionality of terminating Social Security disability benefits without an evidentiary hearing.[266] While this dispute did not arise in a

[262] *See id.* at 1038.

[263] 429 U.S. 190, 190 (1976).

[264] *Id.* at 208–09.

[265] 424 U.S. 319, 332 (1976).

[266] *Id.* at 320.

criminal justice context, the *Mathews* Court articulated a test that applies to due process more broadly: "[O]ur prior decisions," the Court wrote,

> indicate that identification of the specific dictates of due process generally requires consideration of three distinct factors: first, the private interest that will be affected by the official action; second, the risk of an erroneous deprivation of such interest through the procedures used, and the probable value, if any, of additional or substitute procedural safeguards; and, finally, the Government's interest, including the function involved and the fiscal and administrative burdens that the additional or substitute procedural requirement would entail.[267]

The balancing test in *Mathews* clearly weighs in favor of the three proposed principles of "auditability," "transparency," and "consistency" discussed above. With respect to the first *Mathews* factor, in criminal sentencing, the "private interest" at stake is liberty—or, in capital cases, life—which is clearly a private interest of the utmost importance. The second factor addresses the "the risk of an erroneous deprivation of such interest through the procedures used" and the value of additional procedural safeguards.[268] If an inaccurate (in the sense of being erroneously high by a material amount) risk assessment is used at sentencing, the potential that the defendant will receive a more severe sentence as a result is substantial. The three principles can help avoid this injustice and thus, to use the language of *Mathews*, are "procedural safeguards" with high "probable value."[269]

The third factor relates to the burden on the government that would accompany the additional safeguards.[270] Enabling "auditability," "transparency," and "consistency" would place an additional burden on the government, which would need to provide an opportunity for defendants to gain access to information about the methods used to evaluate criminal risk. However, compared with the overall complexity of administering a trial (including all of the associated pre- and post-trial motions, scheduling issues, etc.), the marginal burden would be modest.

In addition, while the third prong of the *Mathews* balancing test relates to the burden on *government*, it is worth noting that adoption of the three principles would also place an additional burden on the

[267] *Id.* at 334–35.
[268] *Id.* at 335.
[269] *Id.*
[270] *See id.*

private companies that provide risk assessment software.[271] For instance, those companies would need to negotiate the language of protective orders and then provide access to the relevant algorithmic information under the resulting protective orders. The costs associated with this compliance would end up getting priced into the products and get therefore passed on to governments. Thus, a fair analysis of the burden on government would also need to recognize the indirect effect of those costs. In short, risk assessment tools would be slightly more expensive—but at the same time, used in a manner much more consistent with the due process. We think that is a good tradeoff to make.

Finally, it is important to address what might be viewed as an inconsistency in the types of information to which access would need to be provided in accordance with the three principles: In relation to "auditability" and "transparency," we highlight the importance of access to the algorithm, while in discussing "consistency," we emphasize the utility of examining input/output information, *without* necessarily knowing the specifics of the algorithm. These differences arise from two different consequences of the need for due process. "Auditability" and "transparency" are a response to the potential problems raised by algorithmic secrecy. In our view, a defendant's right to be sentenced based on information that is not materially inaccurate and in a manner that does not impermissibly rely on factors such as race precludes secrecy (from the defendant).[272] A defendant also has a due process right not to be subject to discrimination (relative to other defendants) based on a factor such as race.[273] To examine that issue, the inquiry can be effective with only the input/output

[271] *Id.* (noting the need to consider "the fiscal and administrative burdens that the additional or substitute procedural requirement would entail").

[272] We write "impermissibly" here to make clear that it is possible to envision scenarios in which explicit consideration of what are normally deemed impermissible factors can be justified. For example, if the input data to a criminal risk prediction algorithm is known to be biased to a particular extent against people of a given race or gender, an argument can be made that it should be acceptable to use the algorithm to correct for that known bias as part of the process of producing an output.

[273] *See, e.g.*, State v. Harris, 786 N.W.2d 409, 416 (Wis. 2010) (writing that "[n]o Wisconsin case has held that defendants have a due process right not to be sentenced on the basis of gender. We now so hold because to do so is in conformity with our understanding of the basic tenets of due process. Everyone agrees, then, that race and gender are improper factors, and that imposing a sentence on the basis of race or gender is therefore an erroneous exercise of discretion").

information (assuming there is a sufficient amount of it to be statistically significant).[274]

CONCLUSION

When used properly, AI has the potential to bring important benefits to criminal risk assessments, including those used in relation to sentencing. At least in theory, AI can operate in a manner free from the many human biases that have impacted risk assessments in the past. AI-based risk assessments also have the potential to be fairer, more consistent, and more objective than methods used in the past.

Yet AI also brings the risk of incorporating and even amplifying the very same biases it could in principle avoid, either through the reliance on data that has bias built into it or through natively generated biases, including those arising from mistaking correlation with causation.[275] And, despite the best intentions of their designers, AI-based systems might at times produce inaccurate risk assessments, either due to problems with the input data or to the algorithms operating on that data.[276]

Due process needs to be a central concern when introducing AI-enabled risk assessments into the criminal justice system. While there is as of yet no case law specific to *AI-enabled* risk assessments, as discussed earlier in this Article, there is a large body of case law on the broader question of what information can be considered at sentencing, as well as a nascent body of case law on algorithmic (though not AI-based) risk assessment. In some respects, the case law is inconclusive, leaving it unclear exactly what specific level of due process applies at sentencing. However, there is also a clear and unambiguous floor: Defendants have *at least* the right not to be sentenced based on materially inaccurate information or on impermissible information such as race.[277]

[274] In examining whether bias is present, useful information can be obtained by examining the algorithm directly as well as by examining input/output information even in the absence of direct access to the algorithm. Bias will sometimes be more easily ascertainable through examination of a large number of inputs and outputs, sometimes through examination of the algorithm.

[275] *See* Chander, *supra* note 260, at 1038.

[276] *See generally* Mayson, *supra* note 9.

[277] Regarding impermissibility of sentencing based on materially inaccurate information, see generally Townsend v. Burke, 334 U.S. 736 (1948). Regarding impermissibility of sentencing based on race, see generally United States v. Onwuemene, 933 F.2d 650 (8th Cir. 1991).

In light of this floor and of the dynamic nature of the algorithms that will be a feature of AI-based risk assessment systems, we have proposed three key principles that can help safeguard due process in sentencing. The first principle, auditability, ensures that a snapshot of the information used to perform the risk assessment is archived. The second principle, transparency, allows that information to be accessed by defendants in a manner that also protects the trade secrets of the manufacturer of the risk assessment system. The final principle, consistency, aims to provide an additional lens through which to examine issues including (1) whether impermissible factors such as race are entering into the analysis, and (2) whether defendants with substantially identical profiles are, at different times, being given substantially different risk scores.

If AI is to be a positive force in promoting equity in sentencing—and we believe that, with proper protections and attention to due process, it can—then it will be important to ensure that risk scores are not corrupted by bias, flawed data, flawed algorithms, or flawed assumptions. That will be much easier to do if the safeguards we have identified here are adopted and upheld.