Harvard Journal of Law & Technology
Volume 32, Number 2 Spring 2019

A REPLACEMENT FOR JUSTITIA'S SCALES?: MACHINE LEARNING'S ROLE IN SENTENCING

Michael E. Donohue[*]

TABLE OF CONTENTS

I. INTRODUCTION

> TO BELIEVE THAT THIS JUDICIAL EXERCISE OF JUDGMENT COULD BE AVOIDED BY FREEZING "DUE PROCESS OF LAW" . . . IS TO SUGGEST THAT THE MOST IMPORTANT ASPECT OF CONSTITUTIONAL ADJUDICATION IS A FUNCTION FOR

[*] Harvard Law School, J.D. 2019. I would like to thank Professor Jonathan Zittrain of the Berkman Klein Center at Harvard Law School, Professor Joi Ito of the MIT Media Lab and the members of the Spring 2018 Ethics and Governance in Artificial Intelligence class for inspiring this Note. My colleagues Marissa Brodney, Andrew Haag, Jacque Sahlberg, and Jonathan Yang, as well as Nati Hyojin Kim of this Journal, all provided invaluable feedback and encouragement on earlier drafts.

> INANIMATE MACHINES AND NOT FOR JUDGES . . . . EVEN CYBERNETICS HAS NOT YET MADE THAT HAUGHTY CLAIM.
>
> — JUSTICE FELIX FRANKFURTER[1]

Criminal sentencing is one of the most difficult responsibilities of judging.[2] It is a different sort of task than the others that face a judge; unlike deciding upon motions or policing the arguments of counsel, sentencing comes down to a singular moment of moral judgment shared between the robed jurist and the defendant standing before the bench.[3]

The task of sentencing is hard because judges face multiple and conflicting instructions from the legislature and society. The sentence must exact proportional retribution for the wrong committed. It must deter the defendant from offending again, as well as others from offending in the first place. The sentence must be long enough to protect society from danger. And, perhaps, the sentence must be of a suitable length and type to rehabilitate the defendant for re-entry into society after punishment.[4] Only occasionally do these instructions point in the same direction, and one judge's interpretation of where they point will differ from others', threatening uniformity across chambers and jurisdictions. As an additional complicating factor, the judge, often a lawyer by training, has limited information about the defendant and the crime in question. At the time of sentencing, the judge will have only experienced a handful of hearings, including, if there is no plea agreement, a trial focused on determining guilt; the judge has even less information on the impact of any possible sentence.[5]

To ease this process — and to ensure to some degree that the judiciary acts as an agent of the legislature's will — legislatures have created a number of tools to quantify the punishment any given defendant deserves. Some have promulgated guidelines as a framework — or mandate — for judges to use in sentencing,[6] and researchers have recommended evidence-based sentencing practices to better understand which defendants are most likely to pose a future danger to society.

[1] Rochin v. California, 342 U.S. 165, 171 (1952).

[2] *See* MARVIN E. FRANKEL, CRIMINAL SENTENCES: LAW WITHOUT ORDER 15–16 (1972); Irving R. Kaufman, *Sentencing: The Judge's Problem*, ATLANTIC MONTHLY, Jan. 1960, at 40, *available at* https://www.theatlantic.com/past/docs/unbound/flashbks/death/kaufman.htm [https://perma.cc/YAB4-79MT].

[3] *See* KATE STITH & JOSÉ A. CABRANES, FEAR OF JUDGING: SENTENCING GUIDELINES IN THE FEDERAL COURTS 80–81 (1998).

[4] For a more robust discussion of the four sentencing philosophies, see generally Mike C. Materni, *Criminal Punishment and the Pursuit of Justice*, 2 BRIT. J. AM. LEGAL STUD. 263 (2013).

[5] *See* Nancy Gertner, *Supporting Advisory Guidelines*, 3 HARV. L. & POL'Y REV. 261, 264 (2009) ("[J]udges receive[] very little training about how to exercise their considerable discretion. Law schools typically d[o] not offer courses on the subject . . . .").

[6] For a discussion of the once-mandatory U.S. Sentencing Guidelines, see *infra* Section III.A.

Some have sought to apply the latest capabilities in data analysis and processing — machine learning — to this task.[7] Despite the promises of these techniques and technologies, however, all have met with criticism from both defendants and judges.[8]

What explains the criticism for these tools, especially the ones based on machine learning? After all, they have the capability to dispassionately apply the law in every case. They can be, at least theoretically, programmatically blinded to factors that are impermissible to consider.[9] Legislatures can imbue these tools with precise weights and algorithms for consideration of the facts. Moreover, once we determine why we recoil from using these capabilities, what do we do about it?

We are uncomfortable with using these capabilities because we are uncomfortable with how the tools we create interfere with and replace the discretion of human judges. But these tools hold great promise if we can find ways for them to assist judges in their exercise of discretion, rather than usurp them. This Note advances that argument by analyzing objections to two different attempts to control judges' discretion: one a creature of computer code and the other a creature of committee. Part I discusses how machine-learning-based recidivism-risk scores in sentencing can manipulate judges by authoritatively anchoring them to only the single sentencing philosophy of incapacitation, a phenomenon this Note terms "philosophy anchoring." This Note further argues that such philosophy anchoring poses a greater threat than a related phenomenon, "starting-point anchoring," which existed in simpler sentencing tools before the emergence of machine-learning-based instruments. Part II explores a different form of algorithmic control through an analysis of the U.S. Sentencing Guidelines during the period in which they were mandatory. The Part finds that the U.S. Sentencing Commission's attempts to come up with a comprehensive and mandatory set of sentencing instructions were met with criticism because they acted as master over judges by removing human discretion entirely.[10]

This Note concludes by suggesting several methods for how these tools could act as mentor and partner to the judiciary. Part III endorses a new attempt at creating a common law of sentencing, using machine-learning-powered data entry and analytics to inform judges of the outcomes and reasoning behind their colleagues' sentencing decisions.

[7] *See infra* Section II.A.

[8] *See infra* Sections II.B and III.B.

[9] *But see, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/6M2C-8HAC] (alleging that algorithms that do not use race as a variable still show bias against racial minorities).

[10] As Part III will discuss, after *United States v. Booker*, 543 U.S. 220 (2005), the Supreme Court ruled that the Guidelines must be optional to be constitutional. *See infra* note 95 and accompanying text.

Then, the Part more ambitiously proposes researching ways to create artificial-intelligence-infused assistants for judges to actively combat cognitive biases and create instantaneous dialog among stakeholders.

II. THE MACHINE AS MANIPULATOR

One of the more recent applications of machine-learning-based systems to criminal justice is the use of recidivism-risk scores to provide input into sentencing decisions. Although the use of machine learning in the development of risk scores does give rise to several objections, many of those objections are not unique to the use of machine learning or even to the use of risk scores.[11] One objection — anchoring on a computationally determined measure of a philosophy of punishment — does pose a unique concern as it risks a particularly troubling interaction of judge and machine.

A. Recidivism-Risk Tools

Tools that attempt to measure the likelihood an offender would violate the law again were first used to determine which inmates to release on parole in the 1930s.[12] These tools used a basic regression model based on race, ethnicity, education, intelligence, and background for their predictions.[13] Equivant, the developers of the machine-learning-based Correctional Offender Management Profiling for Alternative Sanctions tool ("COMPAS"), classifies this as the second generation of risk assessment.[14]

Through the middle of the twentieth century, social scientists developed more sophisticated frameworks to assess recidivism risk, called "third generation" tools.[15] These tools — like the Level of Service Inventory-Revised ("LSI-R") — used dozens of variables and depended on the services of a professional assessment officer. The officer would both collect data on the offender and conduct an interview. Topics included the offender's social network, family history, and neighborhood.[16] After collecting this data, the officer would produce a risk score.[17]

[11] *See infra* Section III.B.

[12] *See* BERNARD E. HARCOURT, AGAINST PREDICTION 77–78 (2007).

[13] *See, e.g.*, Howard G. Borden, *Factors for Predicting Parole Success*, 19 J. CRIM. L. & CRIMINOLOGY 328, 328–30 (1928).

[14] *See* Tim Brennan et al., *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAVIOR 21, 22 (2009). First-generation assessments are those that only use the intuition of a single human. *See id.* at 21.

[15] *Id.* at 22.

[16] *See* HARCOURT, *supra* note 12, at 78–80.

[17] *See id.* at 78–81; Kelly Hannah-Moffat et al., *Negotiated Risk*: *Actuarial Illusions and Discretion in Probation*, 24 CAN. J.L. & SOC'Y 391, 399–400 (2009).

Equivant terms the most sophisticated tools, including its own COMPAS, as "fourth generation."[18] These tools use machine learning in their modeling, link directly to government databases, and provide unified computer interfaces for examiners. And, unlike the third-generation tools, they can output an explicit forecast, rather than a score.[19] The tools process a training dataset of inputs (that is, offender characteristics) and outputs (that is, whether the offender offended again), and then, depending on the precise method used, create a model into which new inputs can be entered to generate a forecast for any given offender.[20] A major difference from third-generation tools is that when a forecast is generated, it can be difficult to understand precisely what led to the system's determination.[21]

These forecast models are not designed to be used in determining post-trial incarceration.[22] Rather, they are designed to be used in determining which defendants should be granted bail during pre-trial proceedings or to be released on parole.[23] But starting with Virginia in 1994, many states have permitted or mandated their use in sentencing, and the judges who pass down the sentence are aware of the risk scores developed during preliminary proceedings as well.[24]

B. The Risk of Anchoring on a Recidivism Score

Although federal courts have declined to rule on the use of risk scores in sentencing, several state supreme courts have affirmed their

[18] Brennan et al., *supra* note 14, at 22.

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

[20] *See id.* at 223–24.

[21] *See id.* at 225. For a broader review of machine-learning approaches to recidivism forecasting and comparison with more traditional methods, see generally Richard A. Berk & Justin Bleich, *Statistical Procedures for Forecasting Criminal Behavior*, 12 CRIMINOLOGY & PUB. POL'Y 513 (2013) and RICHARD BERK, CRIMINAL JUSTICE FORECASTS OF RISK: A MACHINE LEARNING APPROACH (2012).

[22] *See* State v. Loomis, 881 N.W.2d 749, 756 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017) (relating expert testimony suggesting COMPAS "should not be used for decisions regarding incarceration because [it] was not designed for such use"); *id.* at 755 ("It is very important to remember that risk scores are not intended to determine the severity of the sentence or whether an offender is incarcerated.") (quoting a Wisconsin pre-sentence investigation report) (internal quotations and emphasis removed); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 812 (2014) ("[The LSI-R] 'was never designed to assist in establishing a just penalty.'") (quoting an Indiana training manual).

[23] *See* Erin Collins, *Punishing Risk*, 107 GEO. L.J. 57, 83–84 (2018).

[24] Starr, *supra* note 22, at 809 n.11 (2014); *see also Loomis*, 881 N.W.2d at 759 nn.23–24. *See generally* Stephen L. Chanenson & Jordan M. Hyatt, *The Use of Risk Assessment at Sentencing* (Vill. Univ. Charles Widger Sch. of Law, Working Paper No. 193, 2016) (reviewing the usage of risk assessments in four states and associated survey data).

use under state law, state constitutions, and the federal Constitution.[25] The defendant in *State v. Loomis* presented to the Wisconsin Supreme Court the most comprehensive challenge to date. There, the defendant articulated three objections to the use of COMPAS in deciding his sentence: (1) the proprietary nature of the product prevented him from challenging its accuracy; (2) the product was based on group, rather than individualized data; and (3) the product used unconstitutional inputs.[26] The court rejected each of these criticisms in turn, writing that (1) the defendant could verify COMPAS's inputs and argue against them;[27] (2) the risk score merely guided the discretion of a human decision-maker;[28] and (3) there was no indication the sentencing judge had been swayed by any unconstitutional information.[29] Although the court cautioned trial judges in their use of COMPAS, forbidding them from relying solely on COMPAS when deciding on incarceration and that they be informed of the tool's limitations, the court ultimately upheld the defendant's sentence.[30]

The *Loomis* defendant's objections to the use of COMPAS can be grouped into two main categories: first, that the use of group statistical data to sentence him was unfair given that the factors were unrelated to his crime and outside his control, and second, that the lack of transparency into the inner workings of the algorithms deprived him of the ability to challenge the methodology.[31] These two objections are not necessarily unique to the use of risk scores generally or COMPAS specifically. A third objection is hinted at obliquely in *Loomis* and covered elsewhere in the literature: that the use of "precise" machine-learning-based risk scores will "anchor" judges on a single sentencing philosophy and lead to sentences unbalanced by other mitigating factors.

1. Objections to Population-Level Input Data

The first of this Note's critiques is articulated in *Loomis* as an objection to the use of population-level, actuarial data as an input into his sentence.[32] The *Loomis* defendant also objected to calculations made based on gender — a characteristic over which he has no control.[33]

[25] *See Loomis*, 881 N.W.2d at 753 (upholding the use of COMPAS under state and federal due process); Malenchik v. State, 928 N.E.2d 564, 573–74 (Ind. 2010) (upholding the use of LSI-R under state law).

[26] *Loomis*, 881 N.W.2d at 757.

[27] *See id.* at 761–64.

[28] *See id.* at 764–65.

[29] *See id.* at 765–67.

[30] *See id.* at 767–70. For more comprehensive coverage of the court's reasoning, see Note, *State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*, 130 HARV. L. REV. 1530, 1530–33 (2017).

[31] *See Loomis*, 881 N.W.2d at 757.

[32] *See id.* at 764.

[33] *See id.* at 765.

Critics contend that the use of this generalized data serves to exacerbate existing racial and socioeconomic disparities in the prison system.[34] Sometimes this data harms a defendant because it is based on immutable characteristics, and sometimes the data harms a defendant because it is based on past events skewed by those immutable characteristics.[35]

This objection is not unique to machine-learning-based systems, however, and it is also not unique to risk assessment tools. These types of statistical generalizations have been used for parole purposes since early cases in which statistics were applied.[36] Non-machine-learning-based instruments (like the LSI-R) and sentencing guidelines, use statistical generalizations, as well.[37] Indeed, *judges* use such data at an intuitive, albeit unconscious, level whenever they make sentencing decisions.[38] After all, they are exposed to the entirety of a defendant's proceedings and bring their own professional experiences to the final handing down of a punishment.[39]

The use of population-level data by a machine-learning-enabled tool could pose more of a threat than its use in a simpler instrument or by individual judges. Perhaps this data could have unanticipated second-order effects due to the vagaries of training sets or the complexity of the underlying systems.[40] Or perhaps the usage of the data in a sophisticated software tool could give it a greater veneer of legitimacy than when used as a heuristic by an individual judge.[41]

But the first threat is more of a concern over the implementation of the tool than the use of the data in a machine-learning-powered system at all. Clearly any usage of population data needs to be monitored for those second-order effects through a more open tool than used now, as

[34] *See* Starr, *supra* note 22, at 837; *see also* Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1025 (2017) (reviewing FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015)) (urging that we "design our algorithms for a world permeated with the legacy of discriminations past and the reality of discriminations present").

[35] *See* Anna Maria Barry-Jester et al., *Should Prison Sentences Be Based on Crimes That Haven't Been Committed Yet?*, FIVETHIRTYEIGHT (Aug. 4, 2015), https://fivethirtyeight.com/features/prison-reform-risk-assessment/ [https://perma.cc/EG7L-8ZAA] (quoting then Attorney General Eric Holder: "By basing sentencing decisions on static factors and immutable characteristics . . . they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.").

[36] *See, e.g.*, Borden, *supra* note 13, at 328–30; *see also* HARCOURT, *supra* note 12, at 77–78.

[37] *See* HARCOURT, *supra* note 12, at 78–81.

[38] *See* Iñigo De Miguel Beriain, *Does the Use of Risk Assessments in Sentences Respect the Right to Due Process?*, 17 LAW, PROBABILITY & RISK 45, 48 (2018).

[39] *But see* Starr, *supra* note 22, at 865–66 (arguing that the decision being made here is of a different kind from what a judge may make on her own).

[40] *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 680–81 (2017).

[41] *See* Starr, *supra* note 22, at 867–70 (using an original experiment to suggest that "quantified risk assessments might affect the weight placed on different sentencing considerations").

discussed in the following subsection. But that is an argument for better designed systems, rather than no system at all. And, at its core, the second threat is really more directed at the anchoring phenomenon that is discussed below in Section II.B.3.

2. Objections to an Opaque and Proprietary Tool

The second critique centers on the lack of transparency inherent in a proprietary system like COMPAS. In *Loomis*, the defendant argued that his inability to independently verify the accuracy of the COMPAS risk assessment infringed upon his right to due process.[42] Other critics have picked up this same objection in the context of *Loomis* itself[43] and against algorithmic inputs into sentencing more generally.[44] Using non-public risk-assessment tools leaves open the possibility that those tools are operating with meaningful technical flaws in the software.[45] And without broader transparency into the collective inputs and outputs of a system, it is difficult to evaluate the system's effects.[46]

Although such opacity may be unique to machine-learning-based systems like COMPAS when compared to simpler tools like LSI-R,[47] it is not unique when compared to human-driven sentencing. In the federal system, judges need not go into depth when explaining their reasoning — they only need to note that they considered each of the relevant factors.[48] Further, the decision is reviewed under an abuse of discretion standard.[49] The judge is quite nearly the same black box that the algorithm is, and in both cases the parties can check each other on

[42] *See* State v. Loomis, 881 N.W.2d 749, 760 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[43] *See* Beriain, *supra* note 38, at 48–51.

[44] *See* Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 106–10 (2017); Alyssa M. Carlson, Note, *The Need for Transparency in the Age of Predictive Sentencing Algorithms*, 103 IOWA L. REV. 303, 322–24 (2017); *see also* Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to an Explanation*, 7 INT'L DATA PRIVACY L. 233, 233 (2017) (arguing that the European General Data Protection Regulation ("GDPR") includes a "right to explanation" for automated decision-making).

[45] *See* Carlson, *supra* note 44, at 323 (discussing how many risk tools have been put into practice without validation by the executing agency).

[46] *See* Chander, *supra* note 34, at 1039.

[47] *See* HARCOURT, *supra* note 12, at 78–81, 85 (describing how an LSI-R score is calculated).

[48] *See, e.g.*, United States v. Fernandez, 443 F.3d 19, 29–30 (2d Cir. 2006) (noting that judges need not "precisely identify" the factors that led to a sentence and that it is presumed that sentencing judges "faithfully discharge[]" their duty to consider the proper factors) (emphasis removed); *see also* Ryan W. Scott, *The Skeptic's Guide to Information Sharing at Sentencing*, 2013 UTAH L. REV. 342, 380.

[49] *See Fernandez*, 443 F.3d at 27.

the accuracy of the inputs into the sentencing decision, even if they cannot fully vet the decision-making process itself.[50]

This objection is directed at the use of current systems like COMPAS, rather than at machine learning generally. An open-source and vigorously validated set of tools could assuage some of these worries.[51] Plus, as errors are determined or enhancements identified at a system-wide level, fixes can be pushed to all installations of the system.[52] The black-box nature of disparate judges, though, is difficult to fix at a system-wide level, especially given that to fix it, judges would need to engage in more extensive and regular documentation of their reasoning.[53]

3. Objections to Anchoring on a Single Philosophy of Punishment

The third critique, however, is unique to machine-learning-based tools: the risk of what this Note calls "philosophy anchoring." Anchoring as a general term refers to the phenomenon whereby a human decision-maker heavily weighs a piece of tangible and available evidence, potentially in a way that does not serve the decision well.[54] In the case of sentencing, the COMPAS score is presented as three bar charts, each displaying to the sentencing judge a score from one to ten.[55] This clear quantitative measure may outweigh more qualitative factors in the judge's mind — the countervailing human intuitions must be strong to "override" the score.[56]

A related observation has been directed at non-machine-learning-based tools, as well, including the federal sentencing guidelines. This Note calls this related observation "starting-point anchoring" to distinguish it from the philosophy anchoring of recidivism-risk algorithms.

[50] *See* Kroll et al., *supra* note 40, at 657–60. For a similar discussion in the GDPR context, see Lilian Edwards & Michael Veale, *Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18 (2017).

[51] Kroll et al. extensively discuss how to implement such improvements in their article. *See generally* Kroll et al., *supra* note 40, at Parts III, IV, and V.

[52] *See id.* at 701 (urging government systems to have provisions for over-the-air updates). For a discussion of how such continuously updated systems may be implemented in a different configuration than a recidivism-risk tool, see *infra* Section IV.A.

[53] *See* Scott, *supra* note 48, at 382–83 (warning of the danger in only having detailed sentencing opinions for extreme cases).

[54] *See* Note, *supra* note 30, at 1536.

[55] State v. Loomis, 881 N.W.2d 749, 754 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[56] *See* ANGÈLE CHRISTIN ET AL., COURTS AND PREDICTIVE ALGORITHMS 8 (2015), http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf [https://perma.cc/9VU3-376Z]; Note, *supra* note 30, at 1536. Indeed, a risk assessment poses a risk of exacerbating existing and impermissible biases in a judge's mind. *See generally* Ben Green & Yileng Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*, 82 CONF. ON FAIRNESS, ACCOUNTABILITY, TRANSPARENCY (2019), http://yiling.seas.harvard.edu/wp-content/uploads/19-fat_1.pdf [https://perma.cc/MXN3-87NN].

Most prominently, Justice Sonia Sotomayor argued that post-*Booker*, advisory U.S. Sentencing Guidelines act as "anchors" on the district judges who apply them, casting the anchoring effect as a way to ensure uniformity in sentencing.[57] Even with the ability to depart, a guidelines range establishes the starting point for any sentencing analysis.[58]

Three differences separate the philosophy anchoring of machine-learning-based tools from starting-point anchoring and perhaps make the former more dangerous. First, unlike tools through which a human inputs data and manually calculates a score,[59] machine-learning-based tools produce a single definitive answer.[60] Second, the sophistication inherent in a machine-learning-based tool could decrease the likelihood that a judge would fully "override" the decision suggested by the tool,[61] unlike advisory guidelines, from which judges depart frequently.[62] Third, of the four justifications for punishment — retribution, deterrence, incapacitation, and rehabilitation — machine-learning-based tools are focused on optimizing only incapacitation.[63] Indeed, they are designed to predict those offenders most likely to re-offend and keep them incarcerated.

Put another way, these tools function as an overly persuasive input that manipulates the process, rather than, as is the case with anchoring criticisms directed at advisory guidelines regimes, a total substitute for the discretion of a decision-maker trying to balance the rationales behind punishment.[64] Regardless of which justification of punishment a tool is trained to optimize — deterrence, retribution, or even rehabilitation — that tool will anchor its users on the chosen philosophy of punishment, to the exclusion of the others that may counsel a different result.

[57] *See* Peugh v. United States, 569 U.S. 530, 531, 541, 549 (2013) (describing the U.S. Sentencing Guidelines as "anchor[ing] both the district court's discretion and the appellate review process for the federal sentencing process").

[58] *See Peugh*, 569 U.S. at 541; Mark W. Bennett, *Confronting Cognitive "Anchoring Effect" and "Blind Spot" Biases in Federal Sentencing*, 104 J. CRIM. L. & CRIMINOLOGY 489, 520–21 (2014).

[59] *See* Hannah-Moffat et al., *supra* note 17 and accompanying text.

[60] *Cf. id.* at 405–06 (suggesting that risk assessment tools are used as a guide in practice, rather than a device that simply outputs a score).

[61] *See* CHRISTIN ET AL., *supra* note 56, at 8.

[62] *See* UNITED STATES SENTENCING COMMISSION, QUARTERLY DATA REPORT: 3RD QUARTER RELEASE 11, 12–13 (2018) (indicating about half of sentences are not within the applicable Guidelines range).

[63] *See* CHRISTIN ET AL., *supra* note 56, at 9; *see also* James Franklin, *Discussion Paper: How Much of Commonsense and Legal Reasoning Is Formalizable?*, 11 LAW, PROBABILITY & RISK 225, 235 (2012) (discussing the challenges an algorithm might have with judicial balancing); Starr, *supra* note 22, at 867–70 (using an original experiment to suggest that "quantified risk assessments might affect the weight placed on different sentencing considerations").

[64] *Cf.* Peugh v. United States, 569 U.S. 530, 531 (2016) (noting that the advisory Guidelines are meant to achieve sentencing uniformity throughout the federal courts).

Fittingly, this third critique arises from a strength that a machine-learning system brings to bear on the challenge of sentencing. Only a system as well-trained and as complex as a machine-learning model could build the expertise and quickly process the data necessary to produce a single score that predicts recidivism to some accuracy.[65] And perhaps it is the fact that the output is a single score — rather than a narrative or some other less determinate measure — that creates much of the discomfort here.[66] The discomfort inherent in this third critique can be framed like so: these systems use a machine to take advantage of human vulnerabilities to influence the final and moral decision of punishment.

III. THE MACHINE AS MASTER

A natural response to critics objecting to a tool that focuses on a single philosophy of punishment might be to propose a tool that optimizes on *all* the philosophies of punishment: deterrence, retribution, rehabilitation, *and* incapacitation. Such a system might resemble a comprehensive system of sentencing guidelines, not dissimilar from those created by the U.S. Sentencing Commission in the 1980s and treated as mandatory in the federal courts until 2003. Indeed, those guidelines were algorithms, even though they were made through committees rather than code. An analysis of the Guidelines' history and implementation indicates, however, that criticism was not primarily founded on their accuracy; rather, it was focused on their withdrawal of discretion from the human judges making up the federal judiciary.

A. The U.S. Sentencing Guidelines

Congress passed the Sentencing Reform Act in 1984 to reduce disparities in federal sentencing.[67] To do so, it created the U.S. Sentencing Commission, which had a mandate to develop a comprehensive set of sentencing guidelines.[68] These U.S. Sentencing Guidelines ("Guidelines") were required to produce "narrow" sentencing ranges, "with the maximum of any guideline range being no more than 25% of the minimum of such range . . . ."[69] In drafting the Guidelines, the Commission

[65] *See* Barry-Jester et al., *supra* note 35 (reviewing literature indicating risk assessment tools "predict behavior better than unaided expert opinion").

[66] For a discussion of the difficulties in programming explainability into a machine-learning model, see Edwards & Veale, *supra* note 50, at 59–65.

[67] Pub. L. No. 98-473, 98 Stat. 1987 (1984) (codified as amended in 18 U.S.C. §§ 3551–3742 (2012) and 28 U.S.C. §§ 991–998 (2012)); Brent E. Newton & Dawinder S. Sidhu, *The History of the Original United States Sentencing Commission, 1985–1987*, 45 HOFSTRA L. REV. 1167, 1183–84 (2017).

[68] Newton & Sidhu, *supra* note 67, at 1184.

[69] *Id.* at 1186.

was to keep in mind all four philosophies of punishment — retribution, deterrence, incapacitation, and rehabilitation.[70] Despite this explicit instruction, the legislative history of the Act indicated that rehabilitation was considered the least important of the four philosophies,[71] and the Act itself expressed the view that current sentences were not stiff enough for many offenses.[72] In developing the Guidelines, Congress directed the Commission to take into account, but not be bound by, the current sentencing practices of the federal judiciary.[73]

In March 1986, the Commission created two committees charged with creating initial drafts. The first, the "Just Deserts" Committee, was led by Paul H. Robinson, a law professor at Rutgers University and a noted retributivist.[74] The second, the "Crime Control" Committee, was headed by a leader in the then-nascent law and economics school, Michael K. Block, a professor of economics and management at the University of Arizona.[75] The two teams had very different views, but the Commission hoped a synthesis of the approaches would help triangulate its work.[76]

The Crime Control Committee, in contrast to its retributivist peer, sought to create a single system that would ably represent the utilitarian goals of the justice system — namely incapacitation and deterrence.[77] Although it had at its disposal a dataset describing much of the federal criminal justice system, its inputs, and its outputs, the Committee quickly realized the task it had set for itself was too enormous for the methods available and the timescale it faced.[78] The Committee never produced a workable draft.

The Just Deserts Committee was the only committee to create an output. Its draft was complex, seeking to list every harm an offense could cause and quantify it.[79] For example, its July 1986 proposal en-

[70] *Id.* at 1185.

[71] *See id.* at 1183.

[72] *See id.* at 1186 ("The Commission had to 'insure that the guidelines reflect the fact that, in many cases, current sentences do not accurately reflect the seriousness of the offense.'" (quoting 28 U.S.C. § 994(m) (1988))).

[73] *See id.*

[74] *Id.* at 1189, 1226.

[75] *Id.* at 1189, 1227.

[76] An anecdote from the Commission's first meeting illustrates this divide: "Commissioner Robinson observed that the Department is called 'Department of Justice,' not the 'Department of Maximizing Social Utility,' to which Commissioner Block responded that the [Sentencing Reform Act] is part of the Comprehensive Crime Control Act, not the 'Comprehensive Justice and Fairness Act.'" *Id.* at 1227.

[77] *See id.* at 1226.

[78] *See id.* at 1230–31.

[79] *See id.* at 1228.

couraged its followers to use a calculator to find the fourth-root of property damage to calculate the relevant guidelines.[80] Initial reactions from the Commission chairman, judges, and prosecutors were positive,[81] but the other Commissioners, including then-Chief Judge of the First Circuit, Stephen G. Breyer, criticized it as unworkable.[82] Although the Just Deserts concept was eventually abandoned — and Robinson dissented from the promulgation of the Guidelines — it heavily influenced the final product.[83]

The eventual Guidelines produced were akin to the draft of the Just Deserts Committee, if less complex and ambitious in scope.[84] Unlike Robinson's draft — which used his own views on appropriate retribution for certain crimes[85] — the final draft drew more from "past practice" data.[86] The Commission eventually adopted a "modified real offense" approach, whereby the charged offense provided a starting point and the characteristics of the offense — to be proven at the sentencing phase, rather than in trial — would then cause "departures" up or down from the initial target.[87]

The Guidelines as described are algorithms, albeit deterministic and static ones.[88] While the Guidelines were mandatory, the judge's only task was to input data into the tables created by the Commission, resulting in a required and narrow sentencing range.[89] And it is this rote process, not the source of the Guidelines, that engendered the core of the criticism from the judiciary.[90]

B. The Trouble with Replacing Discretion

Although Congress allowed the newly developed guidelines to go into force, the general reaction was far from positive. Defense advocates called foul, arguing that the Guidelines resulted in more severe punishments for their clients and too much clout for prosecutors, especially when paired with mandatory minimum sentences that Congress

[80] *See id.* app. C at 1307; *see also id.* at 1228 (discussing the process by which "harm values" were calculated and converted to "sanction units," which were then adjusted further before proscribing a specific punishment type and severity).

[81] *See id.* at 1228–29.

[82] *See id.* at 1229.

[83] *See id.* at 1231, 1300 n.920.

[84] *See id.* at 1231.

[85] *See id.* at 1229.

[86] *Id.* at 1235, 1269–72.

[87] *Id.* at 1253–61.

[88] *Cf.* Berkeley J. Dietvorst et al., *Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them*, 64 MGMT. SCI. 1155 (2018) (suggesting that customizable algorithms are more palatable to users).

[89] *See* STITH & CABRANES, *supra* note 3, at 83.

[90] *See id.* at 82.

had laid upon federal drug crimes.[91] A wide array of district judges ruled the Guidelines unconstitutional in the years immediately following their promulgation.[92] Although the Supreme Court upheld the Guidelines' constitutionality in 1989,[93] criticism continued into the twenty-first century, culminating in *United States v. Booker*.[94] In *Booker*, the Court ruled that the Guidelines were unconstitutional in that they violated the Sixth Amendment; imposing mandatory sentences based on facts proven by a preponderance of the evidence to a judge violated the right to a jury trial.[95] The once-mandatory Guidelines became advisory.[96]

In the pre-*Booker* period, the federal judiciary acutely felt the loss of discretion that came with mandated and narrow sentencing ranges. Instead of "deliberation and moral judgment," judges were called to conduct "complex quantitative calculations that convey the impression of scientific precision and objectivity."[97] Under the mandatory Guidelines, judges could not take into account either the true suffering felt by the victim or the holistic background of the defendant.[98] And judicial options were constrained; because the top of a sentencing range could be no more than twenty-five percent more than the minimum, a non-prison sentence was an option available only for the most minor offenses.[99]

The diagnosis of Professor Stith and Judge Cabranes gets right to the point: "The federal Sentencing Guidelines as they are now constructed [that is, pre-*Booker*] seek not to *augment* but to *replace* the knowledge and experience of judges."[100] Congress asked the Sentencing Commission to build a machine that ensures consistency in sentencing and adheres better to Congress's wishes than the previous sentencing scheme. The Guidelines delivered it. Even though the Guidelines eventually became based upon past practice data, the lack of rationale behind the rules grated on jurists.[101] Under the pre-*Booker* system, judges had become discretion-less "accountants" in a scheme

[91] *See generally* ERIK LUNA, CATO INST., MISGUIDED GUIDELINES: A CRITIQUE OF FEDERAL SENTENCING (2002).

[92] Newton & Sidhu, *supra* note 67, at 1193 (collecting cases).

[93] *See* Mistretta v. United States, 488 U.S. 361, 374 (1989).

[94] 543 U.S. 220 (2005).

[95] *See id.* at 243–44.

[96] *See id.* at 245.

[97] STITH & CABRANES, *supra* note 3, at 82.

[98] *See id.* at 94.

[99] *See* Newton & Sidhu, *supra* note 67, at 1239.

[100] STITH & CABRANES, *supra* note 3, at 82.

[101] *See id.* at 95 (referring to the Guidelines as *diktats* from the Commission).

set up by others.[102] Judges wanted to balance society's needs one defendant at a time,[103] but the Guidelines forced them down a path that focused on a single balancing of the needs of justice for all cases.

Judges' critique of the Guidelines arose because they could not individualize the cases in front of them. They saw aggravating and mitigating factors before them that they could not apply, threatening their ability to deliver proportional sentences.[104] They feared that the arcane parsing of the Guidelines in open court could reduce any cathartic effect that sentencing could have for the community and risk making the proceeding seem completely arbitrary, rather than driven by any sense of due process.[105] And they saw the Guidelines system as one that elevated the probation officer and the Commission higher in the process than was appropriate.[106]

***

The examples of the Guidelines and the recidivism-risk algorithms show two different approaches to dealing with judges' discretion: in the former, quantifying discretion as precisely as possible and then mandating a result; in the latter, attempting to create perfectly informed discretion thereby manipulating the result because of the way the information is presented. Professor Fennell describes this force to cabin jurists' discretion as "The Machine," drawing on the opinion from Justice Frankfurter excerpted in the epigraph.[107] The human judge is better

[102] *See* Douglas A. Berman, *Sentencing Guidelines*, *in* 4 REFORMING CRIMINAL JUSTICE 95, 103 (Erik Luna ed. 2017); *see also* Luna, *supra* note 91, at 4.

[103] *See* STITH & CABRANES, *supra* note 3, at 84 ("[T]he Guidelines threaten to transform the venerable ritual of sentencing into a puppet theater in which defendants are not persons, but *kinds* of persons . . . .").

[104] *See id.* at 82–83; Robert W. Sweet et al., *Towards a Common Law of Sentencing: Developing Judicial Precedent in Cyberspace*, 65 FORDHAM L. REV. 927, 936–37 (1996); Matthew Van Meter, *One Judge Makes the Case for Judgment*, ATLANTIC (Feb. 25, 2016), https://www.theatlantic.com/politics/archive/2016/02/one-judge-makes-the-case-for-judgment/463380/ [https://perma.cc/95ZF-XWQS].

[105] *See* STITH & CABRANES, *supra* note 3, at 85.

[106] *See generally id.* at 85–91, 97–103. I do not attempt in this Note to determine whether the Guidelines and their state analogs are just when mandatorily implied. Indeed, some evidence suggests that disparities increased after *Booker* made the federal Guidelines advisory. *See* Crystal S. Yang, *Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from* Booker, 89 N.Y.U. L. REV. 1268, 1275 (2014). However, the criticisms that resulted from the discretion-less system of the late twentieth century would likely apply to any attempt to use techniques more sophisticated than used by the Crime Control and Just Deserts committees, such as machine learning, if those techniques resulted in anything mandatory.

[107] Lee Anne Fennell, *Between Monster and Machine: Rethinking the Judicial Function*, 51 S.C. L. REV. 183, 193 (1999); *cf.* Ryan Calo, *Robots as Legal Metaphors*, 30 HARV. J.L. & TECH. 209, 216–23 (2016) (describing use of "robot" as a metaphor by jurists when referring to mechanical jurors or witnesses).

able to deal with the humans that pass through her court than the mechanical Machine can, partially because of her intuition and understanding of the human condition, but also because sometimes novel situations arise that the Machine is not programmed to handle.[108]

This Note suggests that the recidivism and Guidelines cases share the same failing: they attempt to apply the Machine in ways that conflict with, rather than complement, the human judge. So, in the Guidelines case, an algorithm was applied to an individual's sentence when only a human could fully take into account all of the factors. And, in the recidivism-score case, an algorithm attempted to optimize a decision that is incapable of optimization at the individual level. Both cases seek to take away the individualization of the sentencing process. That process — of one robed judge and one convicted defendant in conversation — has moral value in and of itself, and the addition of an interloping machine can cheapen that.[109]

IV. The Machine as Mentor

If sophisticated mandatory and holistic guidelines take away too much discretion, and if singular machine-derived inputs hold too much weight, what is the place of the powerful tools at our disposal? It is in *augmenting* discretion through partnership, rather than *replacing* it.[110] Researchers call this model "Human Agent Robot Teamwork."[111] For example, machines "are not 'aware' of the fact that the model of the world is itself in the world," so they need people to ensure their model remains aligned with reality.[112] And humans are highly sensitive to changes around them, but they use machines to "align and repair their perceptions."[113] In this spirit, this Part will detail two proposals for teamwork between jurists and machines: first, the creation of modern and user-friendly Sentencing Information Systems ("SIS") to implement a new common law of sentencing; and second, the development of judicial cognitive assistants to function as full partners to the judiciary and create a more dialogic method of sentencing.

[108] *See* Fennell, *supra* note 107, at 194–95.

[109] *See* Stith & Cabranes, *supra* note 3, at 82 ("This solemn confrontation was predicated on the fundamental understanding that only a person can pass moral judgment, and only a person can be morally judged.").

[110] *Cf. id.*

[111] *See* Jeffrey M. Bradshaw et al., *Human-Agent-Robot Teamwork*, IEEE Intelligent Sys., Mar.–Apr. 2012, at 9.

[112] *Id.* at 11.

[113] *Id.*

A. An SIS-Enhanced Common Law of Sentencing

Ever since the introduction of mechanical and deterministic inputs into sentencing, commentators who favor greater judicial discretion have been advocating for a common-law approach.[114] In this type of system, judges would issue detailed written opinions when they give out sentences, and those sentences would be subject to substantive appellate review.[115] Supporters argue that such a system is better in that it places more discretion in the hands of judges and that it is better able to address novel fact situations than a single set of sentencing guidelines.[116]

Key to these proposals is readily available data to guide judges in determining what previous jurists have done in similar circumstances. Transcripts alone, often spare and not easily disseminated, are not enough.[117] Accordingly, some have proposed Sentencing Information Systems as a way to fill in the gap. These systems allow judges to review past sentencing decisions on a number of criteria, including narrative details of the case.[118] They "echo traditional common law systems in some ways" and are sometimes used in jurisdictions without a structured guidelines system.[119] Because they reflect the practices of judges as a collective, the systems can evolve as standards change.[120] Some systems provide summary statistics and organize data around different sentencing rationales.[121] Most powerfully, they allow an informed inquiry into similar situations, as defined by the investigating judge. For example, a judge attempting to pass sentence on a young man who robbed a shop with a knife and without causing injury could investigate "similar" cases on any number of dimensions, including the characteristics of the defendant and the characteristics of the crime.[122]

Adoption in the United States, however, has been limited. Much data-entry work is placed on the sentencing judge and the judge's staff, and critics consider the data that *is* entered as unrepresentative.[123] Also,

[114] *See, e.g.*, Gertner, *supra* note 5, at 262, 278–79; Douglas A. Berman, *A Common Law for This Age of Federal Sentencing: The Opportunity and Need for Judicial Lawmaking*, 11 STAN. L. & POL'Y. REV. 93, 94 (1999); Sweet et al., *supra* note 104, at 939–43; Norval Morris, *Towards Principled Sentencing*, 37 MD. L. REV. 267, 275 (1977).

[115] *See* Sweet et al., *supra* note 104, at 939, 944–45.

[116] *See* Morris, *supra* note 114, at 274–75.

[117] *See* Sweet et al., *supra* note 104, at 939–40; *see also* Scott, *supra* note 48, at 362–63.

[118] *See* Marc L. Miller, *A Map of Sentencing and a Compass for Judges: Sentencing Information Systems, Transparency, and the Next Generation of Reform*, 105 COLUM. L. REV. 1351, 1370–71, 1379 (2005).

[119] *Id.* at 1379–80. *But cf. id.* at 1375 (noting the efforts of one Oregon judge to create an SIS for his jurisdiction); Starr, *supra* note 22, at 811–12 (noting criticism of this judge's use of racial data).

[120] *See* Miller, *supra* note 118, at 1381.

[121] *See id.* at 1375 (describing the system in New South Wales, Australia).

[122] *Id.* at 1373–74 (relating an illustration from a Scottish source).

[123] *See* Scott, *supra* note 48, at 362–66.

a system modeled on the common law could increase inequality in sentences, contrary to what algorithmic inputs and sentencing guidelines seek to achieve.[124] Finally, if human judges are to decide on what data to include in their opinions, and if those same judges are to decide what to take from others' decisions, fears of implicit bias become important.[125] Unlike technology-driven fears that can be solved centrally through carefully constructed design principles,[126] judge-driven implicit bias is diffuse and difficult to rectify at scale.

Modern technology can help in two ways. First, modern software development practices could make the consumption of data easier for the judiciary.[127] For example, a new effort assisted by the U.S. Digital Service or the General Services Administration's 18F could professionally deliver and maintain a product that would actually be used and updated.[128]

Second — and more directly related to this discussion of machine learning — a modern SIS could overcome the data-entry problems that have plagued previous iterations. It could, for example, use new voice recognition and language-parsing technologies to directly review sentencing transcripts and court documents to develop databases of factual circumstances and judicial reasoning.[129] It could also use machine-learning-based tools to analyze that database and identify trends that would not be visible to the casual judicial observer.[130] These trends could then be picked up by appellate courts, the Sentencing Commission, or third-party organizations.[131]

This second application of technology to the idea of a common law of sentencing holds particular promise because it places the power of the human judge in the driver's seat in an area where the human judge

[124] *But see* Gertner, *supra* note 5, at 278–79 (suggesting judges could be trained on when and how to appropriately depart from sentencing guidelines).

[125] *See* Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245, 1270 (2016).

[126] *See* Vyacheslav Polonski, *Mitigating Algorithmic Bias in Predictive Justice: 4 Design Principles for AI Fairness*, TOWARDS DATA SCI. (Nov. 23, 2018), https://towardsdatascience.com/mitigating-algorithmic-bias-in-predictive-justice-ux-design-principles-for-ai-fainess-machine-learning-d2227ce28099 [https://perma.cc/T98L-QW2A].

[127] *See* Scott, *supra* note 48, at 393 (describing how difficult an SIS can be to use).

[128] *See* Trisha Thadani, *U.S. Digital Service, Obama's White House 'Startup,' Finding Its Way Under Trump*, S.F. CHRON. (Jun. 19, 2018), https://www.sfchronicle.com/business/article/U-S-Digital-Service-Obama-s-White-House-13005321.php [https://perma.cc/4DGG-XQPB].

[129] *See, e.g.*, Jesse Jarnow, *Transcribing Audio Sucks — So Make the Machines Do It*, WIRED (Apr. 26, 2017, 7:00 AM) https://www.wired.com/2017/04/trint-multi-voice-transcription/ [https://perma.cc/A8HR-BK4U].

[130] *See* NICOLAUS HENKE ET AL., MCKINSEY GLOB. INST., THE AGE OF ANALYTICS: COMPETING IN A DATA-DRIVEN WORLD 75–76 (2016). For a more technical discussion of how machine learning could be precisely applied to predictive analytics in sentencing, see Vincent Chiao, *Predicting Proportionality: The Case for Algorithmic Sentencing*, 37 CRIM. JUST. ETHICS 238 (2018).

[131] *Cf.* Sweet et al., *supra* note 104, at 947–49 (noting that users of an SIS would include groups beyond just trial courts).

is strong — that is, analogizing the facts in front of the judge, and it leverages the power of computers where computers are strong — that is, in meticulously categorizing and inputting data. Further, many of the components of such an application are already in use. "Big Data" has revolutionized business analytics, and user-friendly analytics tools are regularly used as well.[132] And legal research tools have already begun to heavily incorporate artificial intelligence ("AI") and analytics into their offerings.[133]

This, however, is a modest step. A modern SIS is really little removed from a guidelines-driven model of sentencing. This common-law approach is path-dependent and descriptive by nature. The actions of past judges will guide future judges, subject to interventions by legislature through the passage of statutes.[134] A human judge's thought process will always be the ultimate arbiter of a defendant's sentence, and the human judge will always decide how much to depend on others' past practices. Distinguishing cases is easy after all, and trial judges receive great deference from appellate courts on findings of fact.[135]

The power of the tools at our fingertips suggests we can be more ambitious in how we integrate AI and machine learning into the judicial process. Here, we have the potential to create not just a database or set of inputs. We can attempt to create a tool that functions as a true partner to the jurist.

B. A Machine-Learning-Powered Dialog

Another reform concept is that of dialogic sentencing, in which sentencing commissions act as expert analysts, giving feedback on which sentences "work" and which do not, as well as other impacts of jurists' sentencing philosophies.[136] Judges then can use the commission's research in future sentencing decisions. The commissions then look to the actual practices of judges and update their own side of the conversation.[137] In essence, the commission acts as a partner and mentor to judges, informing the judges of the impact of their sentencing decisions on both an individual and population model.

COMPAS and similar risk assessment models seek to play a similar role when used in sentencing. Using sophisticated research, they

[132] *See* HENKE ET AL., *supra* note 130, at 21.

[133] *See, e.g.*, Jason Tashea, *Thomson Reuters Announces Westlaw Edge, Increasing Role of AI and Analytics*, A.B.A. J. (Jul. 12, 2018, 7:00 AM), http://www.abajournal.com/news/article/thomson_reuters_announces_westlaw_edge_increasing_role_of_ai_and_analytics/ [https://perma.cc/64CC-J6DE].

[134] *See* Newton & Sidhu, *supra* note 67, at 1235, 1269–72.

[135] *See* Scott, *supra* note 48, at 373–74.

[136] *See* Eric S. Fish, *Sentencing and Interbranch Dialogue*, 105 J. CRIM. L. & CRIMINOLOGY 549, 581–82 (2016).

[137] *See id.* at 590.

seek to tell jurists the types of people most likely to reoffend. But these models, and even an active and engaged sentencing commission, are not really in conversation with a jurist. This is where an omnipresent assistant could step in.

This partnership or mentorship model is in active development in other fields. The use of "cognitive assistants" is rising in medicine, and some are being developed specifically to help human reasoning. For example, radiologists may soon be able to depend upon cognitive assistants and decision support software to make recommendations to them as they interrogate new data.[138] And researchers have developed natural-language assistants to guide users through exercises in reasoning.[139] Of particular relevance is that system's focus on correcting "problems of common heuristics and biases."[140] Those researchers note that humans can fall prey to cognitive biases and that, to improve desion-making, training systems should encourage the use of "hypothetical thinking and analytic intelligence."[141]

Notable in these models is that they provide, in addition to updated data, a real-time conversation partner. That partner is programmed to shore up areas in which humans show weakness, either in a cognitive bias or in a lack of mental endurance. This is where a proposed cognitive assistant can be most helpful for judges, in particular in state trial courts where the judges do not have dedicated clerks or assistants to be conversation partners.[142] This Part ends by sketching two areas in which an AI assistant could be helpful to judges.

1. Mitigating Cognitive Biases

The fact that humans — and judges — are not fully rational beings is well known.[143] Sometimes their lack of rationality is due to conscious or unconscious biases; indeed, the U.S. Sentencing Guidelines were first developed to reduce racial disparities.[144] Sometimes this lack of rationality is due to a weakness of the human mind, perhaps a lack of endurance or faulty processing. These weaknesses could be ripe for assistance from a sophisticated AI assistant.

[138] *See* Tanveer Syeda-Mahmood, *Role of Big Data and Machine Learning in Diagnostic Decision Support in Radiology*, 15 J. AM. C. RADIOLOGY 569, 573–74 (2018); Munib Sana, *Machine Learning and Artificial Intelligence in Radiology*, 15 J. AM. C. RADIOLOGY 1139, 1139–40 (2018).

[139] *See* Nguyen-Thinh Le & Laura Wartschinski, *A Cognitive Assistant for Improving Human Reasoning Skills*, 117 INT'L J. HUM.-COMPUTER STUD. 45 (2018).

[140] *Id.* at 45.

[141] *Id.* at 46.

[142] *See* Scott, *supra* note 48, at 372–73.

[143] *See* Roth, *supra* note 125, at 1294–95.

[144] *See* Newton & Sidhu, *supra* note 67, at 1180–81, 1184.

An assistant could analyze a judge's past sentencing decisions and actively call out when the judge is departing from his or her own past practices.[145] It could identify unique features of any one case and raise it to the judge's attention while the judge is making the sentencing decision. It could even note areas of regular departure from colleagues or trends that may indicate unconscious biases.[146]

2. Enabling Conversation

In addition, such an assistant could serve as a medium of communication between judges and the communities they serve. By incorporating the assistant into their sentencing decisions, judges could more easily provide data to sentencing commissions and other rule-makers, informing those commissions of issues the judges run into each day. This data, plus a complex analytic engine, could help commissions gain a better grasp of the issues beyond strictly quantifiable outcomes like recidivism. In the reverse, judges could hear more regularly from a sentencing commission that performs community-wide analyses of issues surrounding criminal justice and have those views incorporated into recommendations and prompts.

To be sure, this concept of a judicial cognitive assistant is not without pitfalls. A poor execution — more akin to Microsoft Office's Clippy[147] than an intelligent conversation partner — could be viewed as annoying at best and an impermissible *ex parte* interference with judicial decision-making at worst. And any incorporation of powerful AI systems into human decision-making comes with the risk of implicit bias.[148] But with the right design, perhaps these tools could find a way to be both an informative database and insightful thought-partner.[149]

[145] Daniel Chen suggests that data from machine-learning analytic models could be used as teaching tools for judges at a general level. *See* Daniel L. Chen, *Machine Learning and the Rule of Law* 7–8 (Toulouse Sch. of Econ., Working Paper No. 18-975, 2019).

[146] Tania Sourdin and Richard Cornes outline a similar "Judge Co-Bot" as a quality control mechanism for judges. *See* Tania Sourdin & Richard Cornes, *Do Judges Need to Be Human? The Implications of Technology for Responsive Judging*, *in* THE RESPONSIVE JUDGE 96 (Tania Sourdin & Archie Zariski eds., 2018) (IUS GENTIUM: COMPARATIVE PERSPECTIVES ON LAW AND JUSTICE VOL. 67).

[147] *See* Robinson Meyer, *Even Early Focus Groups Hated Clippy*, ATLANTIC (June 23, 2015), https://www.theatlantic.com/technology/archive/2015/06/clippy-the-microsoft-office-assistant-is-the-patriarchys-fault/396653/ [https://perma.cc/9CPM-N5GB].

[148] *See* Roth, *supra* note 125, at 1270.

[149] For example, Vyacheslav Polonski recommends that designers of machine-learning-enabled sentencing tools keep four principles in mind: (1) using as diverse a training set as possible; (2) building tools to search for and *remove* impermissible bias from systems; (3) ensuring the developers of any tools are themselves diverse; and (4) guarding against malicious actors that might seek to corrupt the machine-enabled process. *See* Polonski, *supra* note 126.

V. CONCLUSION

Since judges began to exercise greater and more nuanced control over sentencing, they have looked to research and technological tools to make their task easier. So, too, have communities looked to many of those same tools to assure that judges follow the communities' will as they sentence. It is only natural that, as our ability to process data, insightfully analyze it, and usefully present it grows, we look more upon the expansion of technology into chambers.

This Note concludes, however, with a note of caution. This technology taken too far could abdicate the role of sentencing to machines, making the sentencing process naught but a mechanical contrivance. The judges of the federal judiciary certainly felt similar pressure when the U.S. Sentencing Guidelines were mandatory, and the Wisconsin Supreme Court foresaw such an eventuality in its own warning to lower courts in *State v. Loomis*.[150] We must be cautious because whatever those machines mete out would not be justice, in the same way that the blindfolded Justitia's scales alone do not tell her how to rule on those before her.[151]

[150] 881 N.W.2d 749, 767–70 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[151] *Cf.* STITH & CABRANES, *supra* note 3, at 79 ("Before [judgment] is exercised, before the sword is raised, Justitia must lift the blindfold . . . . The need is not for blindness, but for insight, for equity . . . . This can occur only in a judgment that takes account of the complexity of the individual case.").