Fall 2020 www.uclajolt.com Volume 25, Issue 1

UCLA Journal of Law & Technology

THE MORAL (UN)INTELLIGENCE PROBLEM OF ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE: A COMPARATIVE ANALYSIS UNDER DIFFERENT THEORIES OF PUNISHMENT

Alberto De Diego Carreras[†]

Abstract

The use of Artificial Intelligence (AI) and Machine Learning (ML) in criminal justice has been understandably controversial. The recent application of these technologies in the form of risk-needs assessment tools—and their potential future application as AI judges—has raised a myriad of concerns. While some worry that these algorithmic tools serve to perpetuate pre-existing biases, others worry that they raise serious Equal Protection or Due Process concerns.

Still others offer a more esoteric concern; that is, that AI/ML tools are inherently incapable of moral judgment, which they consider necessary for judicial decision-making, especially in the criminal context. Accordingly, they fear that these tools offer an inadequate means of deciding the fate of criminal defendants and are especially ill-suited to replace judges altogether. It is this critique that this Comment seeks to challenge.

This Comment does not presume to argue that these tools *are* in fact capable of such moral judgment. Instead, this Comment challenges the premise. It argues that the degree to which a capacity for moral judgment is central to a judge's role depends largely on the presiding theory of punishment. Under a retributive framing, proponents of the moral judgment concern may well be right. After all, retributivism turns on a judgment about the moral culpability that attaches to a criminal defendant for his or her past acts. But this Comment contends that, under a more utilitarian framing, where concerns over moral culpability largely yield to more forward-facing aims, the purported moral incompetence of these tools is less problematic.

Moreover, this Comment argues that our criminal justice system is (and, indeed, should be) trending away from a retributive framing in favor of a more utilitarian approach instead. Accordingly, at least with respect to their alleged moral incompetence, the use of AI and ML tools in criminal justice may not only be unproblematic, it may indeed be desirable if our preferred theory of punishment is utilitarianism.

[†] JD (University of California, Los Angeles); Winner of the 2020 UCLA Journal of Law and Technology (JOLT) Writing Competition.

TABLE OF CONTENTS

The Moral (Un)intelligence Problem of Artificial Intelligence in Criminal Justice: A Comparative Analysis Under Different Theories of Punishment

Alberto De Diego Carreras

Introduction

> There is a better way. We need to move from anger-based sentencing that ignores cost and effectiveness to evidence-based sentencing that focuses on results, sentencing that assesses each offender's risk and then fits that offender with the cheapest and most effective rehabilitation that he or she needs.[1]

The use of Artificial Intelligence (AI) and Machine Learning (ML) in the criminal justice system has been met with tremendous enthusiasm and harsh opposition all at once.[2] Their recent application as risk-needs assessment (RNA) tools and their potential application as future AI judges have been especially controversial.[3] On the one hand, many celebrate the promise that these technologies hold for making our criminal justice system more impartial, efficient, and inexpensive.[4] On the other hand, some raise a myriad of concerns.[5] Amongst them is one which

[1] Hon. Ray Price, Chief Justice of the Supreme Court of Missouri, State of the Judiciary Address (Feb. 3, 2010) (transcript available at https://www.courts.mo.gov/page.jsp?id=36875).

[2] *See, e.g.*, William S. Isaac, *Hope, Hype, and Fear: The Promise and Potential Pitfalls of Artificial Intelligence in Criminal Justice*, 15 OHIO ST. J. CRIM. L. 543 (2018); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Arthur Rizer & Caleb Watney, *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just*, 23 TEX. REV. L. & POL. 181 (2018).

[3] *See, e.g.*, 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; Derek Thompson, *Should We Be Afraid of AI in the Criminal Justice System?*, ATLANTIC (June 20, 2019), https://www.theatlantic.com/ideas/archive/2019/06/should-we-be-afraid-of-ai-in-the-criminal-justice-system/592084/; *The Platform: When a Bot Is the Judge*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y AT HARV. U. (Dec. 1, 2017), https://cyber.harvard.edu/podcast/when-a-bot-is-the-judge.

[4] *See* Rizer & Watney, *supra* note 2.

[5] *See, e.g.*, 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 (arguing that these tools serve to perpetuate and amplify biases embedded in their training data); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 820–21 (2014) (arguing that these tools raise Equal Protection concerns because they consider broad categories such as gender or socioeconomic status); State v. Loomis, 881 N.W.2d 749, 760–64 (Wis. 2016) (considering, but ultimately rejecting, the idea that reliance on the risk prediction generated by an

this Comment terms the moral handicap problem of AI.[6] Proponents of this argument worry that AI/ML tools are inherently incapable of exercising discretionary moral judgment, which they deem a necessary feature of a judge's decision-making process, especially in the criminal context.[7] Accordingly, they fear that these tools offer an inadequate means for deciding the fate of criminal defendants and that they are especially ill-suited to replace human judges altogether.[8] This Comment seeks to challenge that contention.

To be clear, this Comment does not argue that these tools *are* in fact capable of such discretionary moral judgment. That remains to be seen. Rather, this Comment challenges the premise. It argues that the degree to which a capacity for moral judgment is central to a judge's role depends largely on the presiding theory of punishment. Under a predominantly retributive system, the claim that moral judgment is central to a judge's decision-making process may well prove irrefutable. After all, retributivism turns on a judgment about the moral culpability that attaches to a criminal defendant for his or her past acts. Accordingly, having an actor capable of moral judgment in the role of the judge may well be necessary. But this Comment contends that, under a more utilitarian framing, where concerns over moral culpability largely yield to more forward-facing aims, the purported moral incompetence of these tools is less problematic.

Why? This Comment suggests that forward-looking considerations are more easily standardized and aggregated than backward-looking concerns about moral culpability. Accordingly, the role of a judge need not be occupied by an actor capable of moral judgment for utilitarian purposes as it must in order to serve retributive aims. In other words, there is a sliding scale relationship of sorts between our primary theory of punishment and the degree to which the alleged moral incompetence of these tools is problematic: the more retributive our framing, the more problematic the tools. But the more utilitarian and forward-facing our interests, the more innocuous AI's moral deficit becomes.

Incidentally, this Comment suggests that our focus is indeed becoming less and less retributive, will continue to do so (especially as these tools proliferate), and *should* continue to do so. In other words, the theoretical underpinnings of our criminal justice system are shifting. In place of a retributive framing, a more utilitarian and rehabilitative approach is slowly taking primacy.[9]

algorithmic tool violated the defendant's right to due process because of the lack of transparency behind the tool's decision-making process).

[6] *See infra* Part II.

[7] *See, e.g.*, Joshua P. Davis, *Artificial Wisdom? A Potential Limit on AI in Law (and Elsewhere)*, 72 OKLA. L. REV. 51 (2019); Tania Sourdin, *Judge v. Robot? Artificial Intelligence and Judicial Decision-Making*, 41(4) UNSW L. J. 1114, 1128–30 (2018).

[8] *See supra* note 7.

[9] Although utilitarianism is a broad umbrella term that captures a variety of nuanced strains, this Comment will use it as a shorthand for those philosophies concerned with maximizing aggregate welfare, as contrasted with retributive theories concerned with moral culpability. For a more robust discussion of

An emphasis on predictive technologies in criminal justice both evinces this declining interest in retributivism and perpetuates it, fostering instead a results-oriented approach centered around utility and rehabilitation. This is hardly surprising given that prediction is forward-looking and, accordingly, more responsive to (and encouraging of) forward-facing theories like utilitarianism and rehabilitation rather than backward-facing theories like retributivism. In light of this trending paradigm shift, this Comment argues that the concerns around AI/ML's suspected moral handicap lose much of their salience. Far from being dreadful harbingers of a dystopian future, AI/ML could in fact play an invaluable role in ushering in a new era of criminal justice reform, with a renewed focus on forward-looking, results-oriented theories, not to mention significant gains in efficiency, cost, and objectivity.

Because risk-assessment tools long predate their AI/ML iterations, Part I will begin by outlining necessary and useful background on the use and development of risk-assessment tools in the criminal justice system, describing the various generations these tools have undergone. It will also offer a brief primer on Artificial Intelligence and Machine Learning, detailing current applications of these technologies at the various stages of criminal justice, as well as their potential applications as judge replacements.

Part II will then unpack the controversy surrounding the current and potential uses of these technologies in criminal justice. Specifically, it will focus on the moral handicap problem of AI. In turn, Part III will make the case for why the moral handicap argument is flawed. First, it will argue that the concerns regarding AI/ML's capacity for moral reasoning are premised on chiefly retributive considerations and that the increasing use of predictive tools in criminal justice, among other factors, evince a trend away from retribution and towards rehabilitation, undermining the moral handicap concern.

Second, it will argue that normative considerations indeed call for such a trend, and that it should therefore be viewed as a welcomed departure from the heavily retributive policies that have ushered in the era of mass incarceration with all its concomitant vicissitudes. In so arguing, Part III will undertake a wide-ranging discussion. On the one hand, it will draw upon practical concerns, such as overcrowding, alarming rates of recidivism, and unsustainable costs. On the other hand, it will challenge the fundamental notions of free will and moral culpability, upon which retributivism is premised, in light of neuroscience that casts doubt on their coherence. As such, Part III will offer a descriptive, predictive, and normative assessment of criminal justice that should serve to ameliorate the concerns of those who worry about the moral handicap of AI.

Having argued for the virtues and relative superiority of utilitarianism over retributivism, Part IV will then address some possible objections. Specifically, Part IV will respond to the challenge

the various utilitarian philosophies, see JOSHUA DRESSLER & STEPHEN P. GARVEY, CRIMINAL LAW: CASES AND MATERIALS 36–41 (7th ed. 2016).

that, even if it were true that AI will bring about a normatively preferred theoretical shift, it will also operate to push aside other values that are at least as important as prediction, namely mercy, equity, and equality. After unpacking these concerns, Part IV will argue that AI may pose less of a threat to these values than it may appear at first glance, and it will consider ways whereby we may safeguard against the erosion of these values.

Finally, this Comment will conclude by surveying the various possibilities for keeping the human in the loop in the long run. Above all, however, it will underline the nuance it has offered: while we are right to scrutinize these tools due to their destabilizing and transformative power, we must first consider the foundational questions that these tools require us to revisit. For one, what is—and, more importantly, what *should* be—our principal theory of punishment? After all, at least with respect to the moral handicap concern, these tools are not equally problematic under all theories. Thus, we must first identify what we want our criminal justice system to look like before we may properly determine what role, if any, these tools should play in it.

I. <u>Background: Risk-Needs Assessment Tools, Machine Learning, and Robot Judges</u>

Before scrutinizing the adequacy of AI tools to assist or replace judges, it is necessary to understand what these tools are. To that effect, this Part will begin by briefly chronicling the evolution of RNA tools. Because the use of these tools in criminal justice long predates their AI/ML iterations, this Part will first discuss their evolution prior to the development of AI/ML and the algorithmic revolution of the past decade. To bridge the gap between these prior incarnations and their more sophisticated successors, this Part will then offer a brief primer on AI and ML. Finally, this Part will conclude by noting how AI/ML technologies are currently implemented in criminal justice in the form of modern-day RNA tools, as well as how they might come to be used in the relatively near future as AI judges.

A. A Brief History of Risk-Needs Assessment Tools

The last decade or so has seen a rapid expansion in the use of algorithmic RNA tools, aided by advances in the fields of AI and ML.[10] However, RNA tools were part of the criminal justice system long before the AI-enhanced algorithmic explosion that has elevated them to greater prominence.[11] Although "their use has evolved and shifted in response to various competing theories of criminal punishment" throughout the decades,[12] they have long been principally

[10] *See* Starr, *supra* note 5, at 809 (discussing how "this practice has rapidly expanded much more recently" and reviewing case law and legislation to elucidate that widespread expansion).

[11] DANIELLE KEHL ET AL., ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM: ASSESSING THE USE OF RISK ASSESSMENTS IN SENTENCING 3 (2017), https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf.

[12] *Id.*

relied on to predict a criminal defendant's likelihood of reoffending, showing up for trial, and/or responsiveness to rehabilitative treatment options.

The evolution of these tools is typically understood as comprising four distinct generations.[13]

1. First Generation: Professional Judgment

Given the evidence-based orientation for which these tools are typically touted, it is perhaps ironic that the first generation that emerged in the 1920s relied almost exclusively on the judgment of psychiatrists, social workers, and probation officers.[14] Based on their training and experience, the relevant actors would simply "make judgments as to who required enhanced security and supervision."[15] Thus, "[t]he assessment of risk was a matter of *professional judgment*."[16]

2. Second Generation: Static Factors

In the 1970s, the evidence-based approach for which these tools are best known today entered the picture.[17] Rather than relying only on the judgment of clinical professionals, a new generation of actuarial, evidence-based tools emerged.[18] These tools considered various factors related to the criminal history of a defendant to calculate his or her likely risk of reoffending, such as the defendant's number of prior arrests and his or her behavior during incarceration.[19] These instruments assigned quantitative scores to the various factors, the sum of which constituted the risk of recidivism posed by the defendant.[20]

Research soon showed that "these actuarial risk assessment instruments were better at predicting criminal behavior than professional judgement."[21] However, this generation of evidence-based tools considered only static factors; that is, immutable and historical factors that might inform

[13] *See* JAMES BONTA & D.A. ANDREWS, RISK-NEED-RESPONSIVITY MODEL FOR OFFENDER ASSESSMENT AND REHABILITATION 3–4 (2007); KEHL ET AL., *supra* note 11, at 8–9; Faye S. Taxman et al., *Actualizing Risk-Need-Responsivity*, *in* ENCYCLOPEDIA OF CRIMINOLOGY AND CRIMINAL JUSTICE 2–4 (G. Bruinsma & D. Weisburd eds., 2014).

[14] *See* BONTA & ANDREWS, *supra* note 13, at 3; *see also* Taxman et al., *supra* note 13, at 2.

[15] *See* BONTA & ANDREWS, *supra* note 13, at 3.

[16] *See id.* (emphasis added).

[17] *See id.* at 3–4.

[18] *See id.*

[19] *See id.*; *see also* Taxman et al., *supra* note 13, at 3.

[20] *See* BONTA & ANDREWS, *supra* note 13, at 3.

[21] *See id.*

our understanding of a defendant's likelihood of reoffending "but [did] not account for offenders changing for the better."[22]

3. Third Generation: Static & Dynamic Factors

"Recognizing the limitations of second generation risk assessment, research began to develop in the late 1970s and early 1980s on assessment instruments that included *dynamic* risk factors;"[23] that is, "offender characteristics that are amenable to change."[24] The inclusion of these dynamic risk factors enabled these tools to be "sensitive to changes in an offender's circumstances,"[25] thereby generating a more holistic, nuanced risk score. It also served to identify an offender's needs, providing correctional staff valuable insight into what treatment options might lead to the offender's rehabilitation.[26] It is for these reasons that these tools are now largely referred to as risk-*needs* assessment tools.[27]

4. Fourth Generation: Comprehensive Case Management Approach

Finally, the most recent generation of RNA tools is often regarded simply as "an extension of third-generation tools with a focus on treatment matching/case management."[28] Like third generation tools, they focus on identifying offender needs and the appropriate treatment programs that will best address those needs with the aim of reducing recidivism.[29]

In sum, these predictive tools have been developed and applied over many decades for the purpose of identifying the risks posed by criminal defendants and their needs. Rather than relying on the fallible impressions of human judges, these tools help to ensure that predictive decisions are empirically based.

[22] *See id.* at 4; *see also* Christopher Slobogin, *A Defense of Modern Risk-Based Sentencing* 7 (Vanderbilt Univ. L. Sch. Legal Stud. Rsch. Paper Series, Working Paper No. 18-52, 2018), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3269384_code55346.pdf?abstractid=3242257&mirid=1 ("These historical risk factors are called 'static' because they cannot be changed through decisions made by the offender or through treatment interventions.").
[23] BONTA & ANDREWS, *supra* note 13, at 4 (emphasis added).
[24] Taxman, *supra* note 13, at 3. For examples of dynamic factors, see Slobogin, *supra* note 22, at 7.
[25] *See* BONTA & ANDREWS, *supra* note 13, at 4.
[26] *See id.*; *see also* Chelsea Barabas et al., *Interventions Over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment*, PROC. MACH. LEARNING RES., Feb. 23–24, 2018, at 62, 66 ("Today risk assessments are used for two primary purposes . . . 'prediction-oriented' and 'reduction-oriented' approaches to assessment. Prediction-oriented assessments are used to facilitate accurate and efficient prediction of future recidivism, while reduction-oriented tools are intended to inform treatment and supervision plans." (citation omitted)).
[27] *See* Taxman, *supra* note 13, at 3.
[28] *See id.*
[29] *See id.*

But how, then, do AI and ML fit into the picture?[30]

B. A Brief Primer on Artificial Intelligence and Machine Learning

Before we can understand how AI and ML relate to these tools, we must first answer a much more modest question: what are AI and ML, anyway?

Generally speaking, Artificial Intelligence is a subset of computer science and refers to programs capable of automating tasks that are ordinarily thought to require human intelligence, such as playing chess or driving cars.[31] However, contrary to popular belief, these systems are not exactly "intelligent" in the sense that we attribute the term to humans. AI programs are not "thinking machines;"[32] they do not undergo the kind of cognitive processes the human brain undergoes when executing the same tasks.[33] Instead, AI generally operates in one of two ways: (1) Machine Learning, or (2) on the basis of programmed rules and knowledge.[34] Of the two approaches, this Comment will focus on Machine Learning, as it is the approach that RNA tools generally rely on.

Machine Learning effectively operates heuristically. It recognizes patterns across vast swaths of data at speeds the human brain is incapable of and, on the basis of such patterns, learns to perform tasks, draw conclusions, or predict outcomes.[35] In other words, the system self-educates "without *ex ante, explicit programming*."[36] A preternatural ability to sift through vast amounts of data and recognize useful patterns lends itself quite obviously to the project of risk-prediction. Thus, it is no surprise that these technological developments have taken criminal justice by storm.

[30] *See* Glen J. Dalakian II, *Open the Jail Cell Doors, Hal: A Guarded Embrace of Pretrial Risk Assessment Instruments*, 87 FORDHAM L. REV. 325, 327 (2018) ("[M]odern risk assessment tools are more advanced and pervasive than the basic tools used to generate parole decisions in the 1920s. *Modern tools often employ machine learning in their algorithms to inform models that are based on big data sets*." (emphasis added)).

[31] *See* Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1307 (2019); STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 1 (3d ed. 2010).

[32] *See* Surden, *supra* note 31, at 1308.

[33] *See id.*

[34] *See id.* at 1310.

[35] *See id.* at 1311.

[36] Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242, 245 n.9 (2019) (emphasis added).

C. Implementing AI and ML in Criminal Justice

1. Current Applications as Risk-Needs Assessment Tools

Thus far, AI/ML has been largely used to enhance the capabilities of tools that were already in vogue in the criminal justice system, for example, as algorithmic RNA tools. But RNA tools are themselves implemented in different ways and to varying degrees at different stages of the criminal justice process.

First, RNA tools were "once used almost exclusively by probation and parole departments to help determine the best supervision and treatment strategies for offenders."[37] In this context, RNA instruments are administered to ascertain a defendant's likelihood of threat or danger, so as to inform decisions regarding whether or not to release a defendant on parole.

Second, RNA tools have been increasingly relied on at the pretrial stage.[38] Specifically, the use of these actuarial tools has been advocated for and adopted by criminal justice reformers who see the pretrial stage "as a uniquely solvable aspect" of the problem of mass incarceration.[39] These reformers believe that, by introducing more evidence-based practices into pretrial decision-making, greater numbers of criminal defendants may be released to await trial on bail rather than behind bars, thereby eroding mass incarceration.

Finally—and controversially—some jurisdictions have begun to implement RNA tools at the sentencing stage.[40] The rationale for doing so is similar to the rationale for their application at other stages. Rather than rely on the all too fallible intuitions of human judges, these tools offer an evidence-based alternative that should help to make sentencing more fair, impartial, and effective at reducing recidivism.

2. Potential Applications as Judge Replacements

Perhaps most controversial of all are the uses to which AI and ML may yet be put. Chiefly, it is not beyond the realm of possibility that AI/ML technologies could be used not merely to compliment judicial decision-making, but to supplant and replace human judges altogether, ushering in an era of AI judges.[41]

[37] PAMELA M. CASEY ET AL., NAT'L CTR. FOR STATE CTS., USING OFFENDER RISK AND NEEDS ASSESSMENT INFORMATION AT SENTENCING 1 (2011).

[38] *See* Dalakian II, *supra* note 30, at 327.

[39] *Id.* at 329.

[40] *See* CASEY ET AL., *supra* note 37, at 1 ("The use of RNA information at sentencing is somewhat more complex than for other criminal justice decisions because the sentencing decision has multiple purposes . . . only some of which are related to recidivism reduction.").

[41] *See generally* Re & Solow-Niederman, *supra* note 36; Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135 (2019).

II. **The Controversy Over the Moral Handicap of AI**

Understandably, the use of new technologies such as AI and ML in criminal justice has raised a great number of concerns in the legal community.[42] Amongst them is one which this Comment terms "the moral handicap problem of AI."

According to proponents of the moral handicap argument, technologies such as AI/ML are incapable of exercising the sort of discretionary moral judgment that humans possess.[43] Moreover, they argue that an ability to exercise discretionary moral judgment is necessary for judicial decision-making, particularly in criminal justice where judgments regarding the moral culpability of defendants are central. Accordingly, the syllogism goes as follows: because a capacity for moral judgment is a necessary feature of a judge's decision-making process, and because AI/ML tools are incapable of such moral judgment, it follows that AI/ML predictive tools are a problematic means of deciding a criminal defendant's fate and are especially ill-suited to replace judges altogether.

Various permutations of this argument have turned up recently, as the prospect of AI judges has transcended mere science fiction and become an actual possibility. For example, in "Artificial Wisdom? A Potential Limit on AI in Law (and Elsewhere),"[44] Joshua Davis argues that moral judgment is necessary for judicial practice.[45] In turn, he posits, the first-person perspective (that is, subjectivity) is necessary for moral judgment.[46] Lastly, and crucially, he claims that AI is incapable of attaining the first-person perspective.[47] Because AI is incapable of attaining the first-person perspective, Davis concludes it is incapable of exercising moral judgment. And because it is incapable of exercising moral judgment, he concludes that AI will likely never be capable of fully replacing human actors in judicial roles, as a role for humans will be preserved by the need for moral judgment.[48]

In a different piece, Davis reiterates this position, arguing that while "[i]t is not that hard to conceive of computers" taking over the interpretation of our laws "to the extent legal interpretation involves mere description or prediction[,] . . . [i]t is much harder to conceive of computers making substantive moral judgments. So, the ultimate bulwark against ceding legal

[42] *See supra* note 5. *But see* Christopher Slobogin, *Risk Assessment*, *in* THE OXFORD HANDBOOK OF SENTENCING AND CORRECTIONS 196, 203–06 (Joan Petersilia & Kevin R. Reitz eds., 2012) (arguing that the use of certain controversial factors in risk assessment tools does not violate the Equal Protection clause); Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231 (2015) (disputing the equal protection arguments against the use of risk-assessment tools at sentencing).

[43] *See* Davis, *supra* note 7, at 55.

[44] *Id.*

[45] *Id.*

[46] *Id.*

[47] *Id.*

[48] *Id.*

interpretation to computers … may be to recognize the role moral judgment plays in saying what the law is."[49] But Davis does not stop there. Far from being capable of making moral judgments, he argues, AI tools may not even be able to *predict* the moral judgments that human judges might make in order to mimic them, such that even the appearance of moral judgment may be an unreality.[50] Thus, Davis argues, "robojudges would not seem capable of displacing human judges entirely" because "the predictive approach provides only a second-best approximation of the moral judgments necessary for legal interpretation by a judge."[51]

Similarly, Tania Sourdin identifies a number of issues that arise with the development of an AI judge.[52] She argues that the inability of AI tools to exercise discretionary judgment "may result in unfair or arbitrary decisions due to the lack of individualised justice and discretion."[53] Following the same logic as Davis, Sourdin argues that this is because "[m]any judgments within the legal system involve an element of discretion" and that the rigidity of "[c]omputer programs operate[d] based on logic . . . is arguably incompatible with discretionary decisions."[54] Even a recent student note touched on the issue, cautioning against the problematic lack of individualization that results when "[i]nstead of . . . 'moral judgment,' judges [are] called to conduct 'complex quantitative calculations that convey the impression of scientific precision and objectivity.'"[55]

In addition to the scholars that have so far cautioned against the deleterious impact these tools might have on criminal justice, there is also ample literature across disciplines discussing the (im)possibility of developing AI morality, separate and apart from any possible implications their use might have on our criminal justice system.[56] This suggests that, as far as criminal

[49] Joshua P. Davis, *Law Without Mind: AI, Ethics, and Jurisprudence*, 55 CAL. W. L. REV. 165, 165–66 (2018).
[50] *See id.* at 188.
[51] *Id.* at 188–89.
[52] Sourdin, *supra* note 7, at 1126.
[53] *Id.* at 1128.
[54] *Id.*; *see also* W. Bradley Wendel, *The Promise and Limitations of Artificial Intelligence in the Practice of Law*, 72 OKLA. L. REV. 21, 26 (2019) (similarly arguing that AI technologies will never be able to fully replace lawyers "because legal reasoning necessarily involves the types of normative judgments that are impossible for AI").
[55] Michael E. Donohue, Note, *A Replacement for Justitia's Scales?: Machine Learning's Role in Sentencing*, 32 HARV. J.L. & TECH. 657, 670 (2019) (internal citations omitted).
[56] *See* Colin Allen, *The Future of Moral Machines*, N.Y. TIMES (Dec. 25, 2011, 5:30 PM), https://opinionator.blogs.nytimes.com/2011/12/25/the-future-of-moral-machines/ (arguing that while "[f]ully human-level moral agency, and all the responsibilities that come with it, requires developments in artificial intelligence or artificial life that remain, for now, in the domain of science fiction," it is not without the realm of possibility and, in fact, may be in our own interest to develop); *see generally* WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG

justice is concerned, we have not likely heard the end of the argument. As these tools proliferate, they are likely to come under greater and greater scrutiny. Specifically, when the possibility of AI judges more deeply penetrates the national consciousness, greater controversy will arise surrounding the moral competency of these tools.

It is for this reason—the potential of the moral handicap concern to become increasingly salient over time—that it is of great interest to this Comment.

III. **Challenging the Premise: Why the "Moral Handicap" Argument is Flawed**

As stated, the moral handicap argument is predicated on the dual premises that (1) discretionary moral judgment is a necessary feature of a judge's role, and (2) that AI/ML tools are incapable of exercising such judgment (and probably always will be). The latter premise alone is in and of itself somewhat controversial. Indeed, there are those who posit that, while AI/ML tools may not yet be able to engage in moral reasoning, it is not necessarily a foregone conclusion that they will never be able to do so.[57] Needless to say, if AI/ML tools were one day capable of exercising such judgment in a manner comparable to humans, the moral handicap concern should be largely quelled.

But this Comment does not purport to know the answer to such an unsettled question, nor does it intend to speculate. Instead, this Comment challenges the major premise on which the moral handicap syllogism rests; that is, that a capacity for discretionary moral judgment is a necessary feature of a judge's decision-making process. Specifically, this Comment argues that the degree to which that proposition holds true is highly dependent on the presiding theoretical mode of justice. Accordingly, the tools in question are less problematic under some theories than others. Thus, as the truth value of the syllogism's major premise falters, the suitability of these tools increases. Naturally, the reverse is also true: the truer the premise, the less suitable the tools.

Indeed, under a retributive framing, proponents of the moral handicap argument may well be right. After all, a retributive understanding of criminal justice hinges entirely on the notion that one who commits a crime is morally culpable for that crime and, in turn, his or her moral culpability merits (and, indeed, requires) punishment.[58] Before deciding what punishment, if any, is appropriate for a criminal defendant's past acts, a preceding discretionary judgment must be made as to the moral culpability that attaches to those acts. Therefore, moral judgment is inescapably necessary to effectuate retributive goals. It follows that, to the extent that a criminal justice system is retributive, the introduction of tools that are incapable of moral judgment would seem awfully problematic, as they would be fundamentally inconsistent with the theoretical

(1st ed. 2008); CAMBRIDGE UNIV. PRESS, MACHINE ETHICS (Michael Anderson & Susan Leigh Anderson eds., 2011).

[57] *See supra* note 56.

[58] *See* JOSHUA DRESSLER, UNDERSTANDING CRIMINAL LAW 18 (8th ed. 2018).

underpinnings of criminal justice. But to call this the end of the inquiry, thereby dismissing these tools as unhelpful or worse, is to take a myopic view of criminal justice. After all, what if the problematic variable were not the moral handicap of AI, but rather the chosen theoretical mode of justice?

Retributivism is not the only game in town. Under a more utilitarian framing, these tools appear far less problematic. To understand why, it is first important to understand the differences between retributivism and utilitarianism.[59] Broadly speaking, retributivism is backward-looking, and utilitarianism is forward-looking. As previously explained, a retributivist evaluates the past to determine what measure of moral culpability should attach to a criminal defendant for his or her acts so as to determine what kind (or degree) of punishment they deserve. What future impact such punishment might have on society at large is largely immaterial to the retributivist; punishment is inflicted because it is deserved, not because it may engender some future good.[60]

On the other hand, a utilitarian is less concerned with the defendant's past moral culpability and more interested in how society's response to such acts will impact the aggregate welfare. Indeed, utilitarians "care about the past only to the extent that it helps *predict* the future."[61] Because passing moral judgment on defendants is not the utilitarian's concern, it follows that these tools are not as problematic under a utilitarian conception of criminal justice, even if they are incapable of such judgment. But there is more: what utilitarianism *is* awfully concerned with is *prediction*, which is precisely what these tools are especially gifted at. Thus, under a utilitarian framework, not only are these tools less problematic, they are exceedingly useful and far more capable than humans at effectuating the relevant goals of predicting defendant risk and needs. Accordingly, rather than be viewed as a source of dread and apprehension, these tools might be a cause for optimism.[62]

This is not to say that moral judgment is altogether irrelevant under a utilitarian conception of justice. Of course, both utilitarianism and retributivism are moral theories. Accordingly, both involve—and indeed require—moral evaluation. In fact, a utilitarian makes a moral judgment by the sheer fact of holding a utilitarian worldview. After all, it is morally charged to suggest that aims such as deterrence, incapacitation, or rehabilitation are "good" or, for that matter, "better"

[59] For a more in-depth discussion of the different theories of punishment, see Andrea Avila, Note, *Consideration of Rehabilitative Factors for Sentencing in Federal Courts:* Tapia v. United States, 131 S.Ct. 2382 (2011), 92 NEB. L. REV. 404, 406–07 (2013).

[60] *See* DRESSLER, *supra* note 58 ("Retributivists believe that punishment is justified when it is deserved. It is deserved when the wrongdoer freely chooses to violate society's rules. To an uncompromising retributivist, the wrongdoer should be punished, *whether or not it will result in a reduction in crime*." (emphasis added)).

[61] *Id.*

[62] For a discussion of the possible impact these tools might have on values other than prediction—such as mercy, equity, or dignity—see *infra* Part IV.

than the retributive alternative. But the normative exercise of assigning weights to interests under any theory of punishment is carried out by a number of different actors, including actors outside of the courtroom, such as legislators, correctional staff, or parole officers, to name only a few. Thus, the question is not whether moral weighing is necessary under a utilitarian system as a general matter. Of course, it is. Rather, the question is whether moral judgment is required of a *judge* at the moment of sentencing, such that AI judges would constitute inadequate replacements.

This Comment argues that the normative aims of utilitarianism need not actually require a moral judgment from the judge at the moment of sentencing, so long as those aims have been predetermined by human actors and the AI judges have been trained on data sets that will produce outcomes conducive to meeting those aims. However, admittedly, this argument presupposes a purely utilitarian system. Thus, this Comment next argues that, even under a mixed system—that is, one where multiple theories are in play at once—these tools remain the better of the alternatives so long as the *primary* motivating theory is utilitarianism.

To address the first contention, assume for a moment a purely or strictly utilitarian system. Such a system would be concerned with maximizing aggregate welfare. To be sure, the decision to prioritize aggregate welfare is a normative choice in and of itself. Moreover, defining what maximizes aggregate welfare also requires a series of further normative judgments. For example, one of those normative judgments might be that aggregate welfare is maximized when criminal defendants are rehabilitated. In such a case, rehabilitation would be the normative aim that a criminal sentence should seek to effectuate. But in order to effectuate the aim of rehabilitation at sentencing, the judge would not have to make the initial normative judgment that the aim is desirable, which of course requires moral judgment. For that matter, the judge need not make any subsequent normative judgments either.

Instead, a judge's role in such a system would be to predict what sentence or treatment program is most likely to result in the rehabilitation of the defendant, such that the predetermined normative aims would be met. To do so, the judge would not have to know or define rehabilitation normatively because rehabilitation would be captured in the training data as a reduction in recidivism. In other words, if an AI judge were trained on data from which it could pull to predict what outcome would most likely result in a reduction of recidivism, then the aims of utilitarianism would be met without it ever having had to make a moral judgment of its own. Accordingly, we can imagine a rules-based system where legislators and policymakers make all of the relevant normative judgments and devise rules to reflect those judgments.[63] An AI judge in such a system would then apply those normatively-charged rules in any given case and, based

[63] The term "rules-based" should be understood as contrasted with a *standards*-based, heavily discretionary approach.

on the data on which it was trained, produce the result (that is, the judgment) most likely to serve our utilitarian aims.

Of course, it may be argued that the same could be true under a retributive framing: so long as the relevant retributive considerations were defined ex ante by legislators, policymakers, and other relevant actors, an AI judge would merely need to predict the punishment most likely to serve said retributive aims. But this argument too hastily assumes the soundness of its premise. Such an arrangement is conceivable under a utilitarian framing precisely because the aggregative nature of utilitarian objectives allows for standardization in a way that the individualized nature of retributive concerns does not. In other words, the distinction lies in the fact that retributive concerns are less amenable to definition ex ante than utilitarian concerns. After all, the core normative aim of retributivism is to issue a punishment that is proportionate to the defendant's moral culpability. Such a determination could never be adequately captured ex ante in the way that rehabilitation, for example, might be captured as a reduction in recidivism.[64] Accordingly, it would seem that a retributive framing, with its emphasis on individualized judgments centered around moral culpability, reserves a role for discretionary moral judgment at sentencing in a way that utilitarianism, by its more aggregative and pragmatic nature, does not.[65]

[64] After all, the question of whether rehabilitation occurred—if framed, for example, as whether or not a reduction in recidivism occurred—is an easier, more empirical question to ask and answer. A reduction in recidivism either occurred or it did not. On the other hand, a question like whether a judgment proportionately punished a defendant is not as comfortably captured in data or even framed empirically, for that matter. Professor Christopher Slobogin conceives of this contrast as a failure of "outcome measures." *See* Slobogin, *supra* note 22, at 17–18. While it is conceivable to measure the degree to which an outcome met its utilitarian aims, the same cannot be said for retributive outcomes, given that the criteria for moral culpability is not as clear or readily quantified as the criteria for welfare maximization. *Id.*

[65] Some might also challenge this theoretical contrast by contending that the predictive capabilities of AI might not be limited to predicting "means." In other words, while incapable of *making* moral judgments, perhaps an AI judge could *predict* the moral judgments that a human judge would make. Such an AI could be said to predict not only the *means* by which to meet predetermined ends, but the *ends* themselves. It is not at all clear that such a development is possible; technical hurdles beyond the scope of this Comment may well impede it. *See* Davis, *supra* note 49, at 187–89 (sketching some of the reasons why such a predictive model may be unsuccessful). But assuming it were possible, and assuming no meaningful distinction exists between making substantive moral judgments and predicting them, then such an AI would surely obliterate the theoretical contrast advanced by this Comment. After all, the moral capabilities of AI would be unproblematic regardless of the underlying justification for punishment. This may seem at first glance like a sound challenge. But, in fact, it in no way undermines the thrust of the present argument. To posit that AI is capable of such prediction *and* that such a prediction may adequately stand in for moral judgment is tantamount to saying that there is no moral handicap problem of AI to begin with, or at least that it has been solved. But that it is not the premise on which the present

To address the second argument, some may point out that, while it may be possible to isolate each of these theories for purposes of a theoretical discussion, in practice there will never be such a thing as a purely or strictly utilitarian system. In other words, while at different times different moral theories may predominate, one never monopolizes the system at the expense and absolute exclusion of all others. Instead, in any given case, human beings will be influenced by different theories at different times or even by multiple theories all at once. Many would argue that not only is this the most accurate description of reality now and for any foreseeable future, it may indeed be desirable.[66] However, even under such an understanding of criminal justice, AI tools remain the better of the alternatives under a predominantly, though not exclusively, utilitarian system.

To illustrate why, assume first a predominantly retributive system. According to the view described above, even a predominantly retributive system will be influenced at times by utilitarian and rehabilitative considerations. In other words, while a judge under such a system may be primarily concerned with weighing a defendant's culpability to determine the punishment the individual deserves, that is not to say that the judge will not also be concerned, albeit perhaps to a lesser extent, with other considerations that may deviate from a purely retributive approach. For example, the judge may simultaneously be concerned with achieving the best result for the common future good of society. That is to say, the judge may be attempting to predict the future to some extent, though it may not be his or her chief focus. It is well-known that human beings are quite unexceptional at predicting the future, to put it charitably.[67] Thus, under the arrangement just described, the human judge may be quite good at judging a defendant's culpability, whereas AI tools would not be, but mediocre at predicting the future, whereas AI tools would be far better.

In other words, a human judge's performance under such a hybrid system would by no means be perfect. Nevertheless, we would prefer it to the alternative; that is, a predominantly retributive system led by AI judges rather than human judges. Why? Because despite a human judge's inability to exercise all relevant skills perfectly, (s)he is nevertheless better than an AI tool at effectuating the primary purpose of the system at hand, even if (s)he might not be so great at

argument rests. If there is no moral handicap, there is no AI problem. But so long as the moral handicap problem exists, so does the theoretical contrast between utilitarianism and retributivism.

[66] *See* Donohue, *supra* note 55, at 665–66 (discussing the problems with anchoring to a single philosophy of punishment).

[67] *See, e.g.*, Alexandra Ossola, *Why Are Humans So Bad at Predicting the Future?*, QUARTZ (Nov. 20, 2019), https://qz.com/1752106/why-are-humans-so-bad-at-predicting-the-future/; Maria Temming, *AI Can Predict Which Criminals May Break Laws Again Better Than Humans*, SCIENCENEWS (Feb. 14, 2020, 2:00 PM), https://www.sciencenews.org/article/ai-can-predict-criminals-repeat-offenders-better-than-humans; Caroline Beaton, *Humans Are Bad at Predicting Futures That Don't Benefit Them*, ATLANTIC (Nov. 2, 2017), https://www.theatlantic.com/science/archive/2017/11/humans-are-bad-at-predicting-futures-that-dont-benefit-them/544709/.

effectuating secondary or tertiary aims. The point being that our criteria for evaluating the functionality and desirability of any particular arrangement is not (and could not reasonably be) absolute perfection.[68] Rather, our criteria should be focused on whether or not the primary aims of our preferred system are being served, even if some secondary or tertiary aims may suffer as a result.

In light of this, let us now explore what a predominantly utilitarian alternative would look like. Under such a system, criminal justice would be primarily concerned with maximizing the overall future wellbeing of society. As a secondary matter, criminal justice might be concerned with assigning weight to a defendant's culpability to determine the degree of punishment deserved, due to its retributive influence. Under such a system, AI tools would be excellent at effectuating the primary purpose of criminal justice—that is, predicting the result most likely to maximize the aggregate good—whereas human agents would be mediocre at it. On the other hand, these tools would be much worse than human agents at effectuating the secondary purpose of criminal justice, that is, weighing the defendant's moral culpability. Put differently, under a predominantly but not exclusively utilitarian framework, these tools would not be capable of exercising all the relevant skills perfectly. However, as before, perfection cannot possibly be the correct criteria, for no system of justice is ever likely to be perfect and, in any case, no such system exists now. Thus, all we may reasonably strive for at present is the best of all possible alternatives. If indeed these tools are significantly better at prediction than humans, then it follows that, so long as our chosen and primary theory of punishment is utilitarianism, these tools may indeed be the best of all possible alternatives, notwithstanding the fact that they may not be good at effectuating some secondary or tertiary aims of a hybrid system.

As suggested, this conclusion depends entirely on utilitarianism being our chosen mode of theoretical justification. Accordingly, the following Subpart will explain that, indeed, the current state of our criminal justice system evinces a growing preference for utilitarianism, demonstrated in part by increasing reliance on these tools, which in turn promise to reinforce the trend as they continue to proliferate.[69]

[68] For a related discussion of the merits of a comparative analysis, see generally Volokh, *supra* note 41. *See also* Dalakian II, *supra* note 30, at 346 ("Regardless of whether bail is set using risk assessment instruments or judicial discretion, an absolute lack of false positives and false negatives is impossible. The real goal should be determining which alternative, or combination of methods, is most beneficial to society, victims, and accused individuals. It is critical to seek real and deliverable reform rather than merely avoiding flawed models or frameworks without an alternative to the profound injustices in modern pretrial detention.").

[69] *See infra* Subpart III.A.3.

A. A Descriptive and Predictive Assessment: Criminal Justice Trends Away from Retributivism

"Does nothing work?," queried New York criminologist Robert Martinson in 1974, concluding that rehabilitation was ineffective at reducing recidivism rates.[70] In doing so, Martinson helped precipitate a seismic paradigm shift in criminal justice.[71] The predominantly rehabilitative model that had dominated the arena for much of the 1960s and 1970s was effectively abandoned in favor of a more retributive model.[72] Hence, the "get tough" on crime movement was born.[73]

The new, tougher paradigm has reigned supreme ever since. Today, however, it is unclear whether it has worked any better than the rehabilitative model that failed according to Martinson.[74] Bogged down by unrelenting recidivism rates, suffocating costs, and staggering numbers of its citizens behind bars, the United States long ago entered the now-infamous era of mass incarceration.[75] In doing so, it has set itself apart on the world stage quite poignantly as the most punitive country in the world.[76] However, this Comment posits that the pendulum is swinging back, likely as a response to this undeniable crisis.

[70] Robert Martinson, *What Works? Questions and Answers About Prison Reform*, 35 PUB. INT. 22, 48 (1974); *accord* DOUGLAS S. LIPTON ET AL., THE EFFECTIVENESS OF CORRECTIONAL TREATMENT: A SURVEY OF TREATMENT EVALUATION STUDIES (1975).

[71] *See* Jerome G. Miller, *The Debate on Rehabilitating Criminals: Is It True That Nothing Works?*, WASH. POST (Mar. 1989), https://www.prisonpolicy.org/scans/rehab.html (discussing the impact of Martinson's views and the "nothing works" movement).

[72] *See* BONTA & ANDREWS, *supra* note 13, at 9; *see also* Slobogin, *supra* note 22, at 2 ("Beginning in the 1970s, a sentencing revolution took hold . . . based predominantly, and occasionally entirely, on a desert philosophy.").

[73] *See* BONTA & ANDREWS, *supra* note 13, at 9.

[74] *See id.* (discussing how "after 30 years of experimentation with getting tough," prison and probation populations have skyrocketed and recidivism has either remained the same or even increased).

[75] *See* Dalakian II, *supra* note 30, at 336 ("The United States has the largest rate of imprisonment in the world: 655 incarcerated for every 100,000 people. Representing merely 5 percent of the world's population, the United States has over 20 percent of the global incarcerated population in its jails and prisons.").

[76] Mirko Bagaric et al., *Introducing Disruptive Technology to Criminal Sanctions: Punishment by Computer Monitoring to Enhance Sentencing Fairness and Efficiency*, 84 BROOK. L. REV. 1227, 1227 (2019); *see also* James Q. Whitman, *A Plea Against Retributivism*, 7 BUFF. CRIM. L. REV. 85, 85 (2003) ("[T]hirty years of harsh justice have made for an epochal shift in American law, opening a large divide between the United States and the other countries of the western world. American criminal punishment is now staggeringly harsher . . . . There is an American world, tough and unforgiving, and a Euro-Japanese world, mild in ways that have come to seem wholly impossible in the American climate.").

1. Increasing Use of Predictive Tools in Criminal Justice Evinces a Declining Interest in Retributivism and a Growing Preference for Utilitarianism

The increasing use of predictive tools in the criminal justice system is arguably in and of itself evidence of a growing preference for utilitarian goals. After all, a criminal justice system increasingly characterized by forward-looking, predictive tools would seem to be trending towards a utilitarian approach and away from a retributive one, given the forward-looking character of the former and the backward-facing nature of the latter.[77] And there is no denying that our system is in fact increasingly touched by the use of these tools. "Predictive technologies are spreading through the criminal justice system like wildfire."[78] As Sonja Starr puts it, "[i]t is an understatement to refer to risk assessment as a criminal justice *trend* . . . [r]ather, we are already in the risk assessment era."[79] If it is true that we are already in the risk assessment era rather than merely trending towards it, then our criminal justice system has entered an era chiefly concerned with prediction. This, in turn, means that we have entered an era characterized predominantly by utilitarian goals and interests.

While this Comment is less confident that we have already entered such an era, that we are trending towards it is hardly disputable. Many states, such as New York and Tennessee, have been using algorithmic tools for some time now to facilitate decision-making in the parole context on the basis of risk scores.[80] The rapid proliferation of these algorithmic predictive tools in the parole setting is significant because it suggests greater weight assigned to forward-looking considerations, even in a context that has historically focused on a greater number of backward-looking factors. Indeed, in evaluating whether or not to let a prisoner out on parole, a parole board considers the nature and severity of the offense committed, the length of the sentence already served, the inmate's degree of remorse for the offense, and the prisoner's behavior during incarceration—all of which are backward-facing factors. To be sure, a parole board will also consider an offender's risk of recidivism, but this is only one of many otherwise backward-facing considerations weighed when making parole eligibility determinations. Accordingly, increasing reliance on RNA tools in the parole context might suggest an increase in the value assigned to the forward-looking variable(s). In other words, whereas parole could turn largely on a backward-oriented determination (that is, "how much does this prisoner *deserve* to be let out

[77] *See* Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583, 592 (2018) ("Risk assessments are orthogonal to culpability assessments, both conceptually (the first is forward-looking, the second backward-looking), and practically (for instance, a single prior robbery conviction might call for more enhancement on desert grounds than on risk grounds).").

[78] Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 61 n.1 (2017).

[79] Sonja B. Starr, *The Risk Assessment Era: An Overdue Debate*, 27 FED. SENT'G REP. 205, 205 (2015).

[80] *See, e.g.*, Tenn. Code Ann. § 41-1-412 (LEXIS through the 2020 Regular Session); N.Y. State Corr. & Cmty. Supervision Directive, COMPAS Assessments/Case Plan 1 (2019), https://doccs.ny.gov/system/files/documents/2020/02/8500.pdf.

early?"), increased reliance on predictive tools suggest a turn towards a more utilitarian framing (that is, "what would be the impact on aggregate welfare if this prisoner were let out early?").

These tools have also proliferated at the pretrial stage, where they are often relied on by judges when deciding whether to release criminal defendants on bail or keep them behind bars until trial.[81] In fact, now more than ever, with the aim of chipping away at the problem of mass incarceration, many proponents of criminal justice reform have advocated for the use of RNA tools at the pretrial stage as a promising means of reducing the pretrial jail population.[82]

Concededly, an increased use of these tools in the pretrial context is arguably less probative of a trend towards utilitarianism because pretrial decision-making has always been inherently predictive in nature, regardless of what the dominating theory of punishment may be. After all, the decision of whether to keep a defendant behind bars or release them on bail amounts to little other than a judgment regarding the likelihood of a defendant to evade trial, that is, a flight-risk prediction. Nevertheless, the degree to which these tools have been embraced by numerous states suggests that it may be the result of a broader paradigm shift. For example, some states, like New Jersey, weigh the predictions of RNA tools so heavily that judges may often consider little else besides those risk scores in deciding whether or not to let an offender out on bail.[83] Moreover, while the use of these tools at the pretrial stage may not in isolation speak clearly to a broader trend, it is at minimum consistent with such a trend, and when combined with other evidence, suggestive of it.

But of greatest significance here is the recent and growing adoption of these tools at the sentencing stage. In the last ten years, roughly twenty states have formally adopted the use of RNA tools at sentencing, either by statute, administrative policy, or judicial policy.[84] Some states have even gone as far as to adopt statutes that actually mandate the use of RNA tools at sentencing.[85] However, the most notable stamp of approval for the use of these tools at the sentencing stage comes from the American Law Institute (ALI), which in 2017 revised the Model Penal Code (MPC) for Sentencing for the first time since 1962.[86] In this revision, the ALI

[81] *See* Dalakian II, *supra* note 30.

[82] *See id.*

[83] *See id.* at 341.

[84] *See* PAMELA CASEY ET AL., CTR. FOR SENT'G INITIATIVES, USE OF RISK AND NEEDS ASSESSMENT INFORMATION IN STATE SENTENCING PROCEEDINGS 3 n.10 (2017), https://cdm16501.contentdm.oclc.org/digital/collection/criminal/id/296; Starr, *supra* note 5, at 809 n.11 (reviewing case law, sentencing commission websites, and legislation, all of which evince that as of 2014 at least twenty states took into account actuarial tools upon sentencing).

[85] *See* KY. REV. STAT. ANN. § 532.007; (LEXIS through Ch.128 of the 2020 Regular Session); OHIO REV. CODE ANN. § 5120.114; (LEXIS through File 48 of the 133rd (2019-2020) General Assembly); *see also* Starr, *supra* note 5, at 809 n.11 (detailing a comprehensive review of like statutes).

[86] *See* Starr, *supra* note 5, at 815.

expressly endorsed actuarial sentencing.[87] Not only is such a development significant because it "reflects an emerging academic consensus,"[88] it is also significant because it portends even greater acceptance and widespread use of these tools given the MPC's tremendous influence. After all, "[t]he original MPC was 'one of the most successful law reform projects in American history,' producing 'modernized penal codes in a substantial majority of the states'"[89] and becoming "the document through which most American lawyers come to understand criminal law."[90] The ALI's adoption and endorsement of the use of these tools at sentencing, therefore, reflects a growing preference for forward-looking considerations.

Actuarial sentencing has also been embraced by the National Center for State Courts (NCSC), as well as by the Conference of Chief Justices (CCJ) and the Conference of State Court Administrators (COSCA).[91] In fact, Judge Roger Warren, the president emeritus of the NCSC, credits a 2007 joint report by the NCSC and others, as well as a formal resolution of the CCJ and COSCA of the same year, for precipitating the growing embrace of actuarial sentencing.[92]

In short, this growing trend is undeniable and, especially in light of the ALI's recent endorsement of it, only likely to gain further traction. For present purposes, however, what is most important about this trend is not so much that it is taking place, but rather what it represents. In light of the fact that an ever-growing number of states are now adopting and relying on these predictive tools not only at the pretrial and parole stages but, significantly, at the sentencing stage—and in light of the fact that the ALI, NCSC, and others endorse them—this Comment argues that a preference for a predominantly utilitarian approach to sentencing is well underway.

2. Factors Other Than the Use of These Tools Also Evince a Declining Interest in Retributivism

Some may refute the conclusion that the increasing use of RNA tools is the result of a growing philosophical shift, arguing instead that it is merely the result of pragmatic and seductive

[87] MODEL PENAL CODE: SENT'G § 6B.09 (AM. L. INST., Proposed Final Draft 2017).

[88] *See* Starr, *supra* note 5, at 815.

[89] *See id.*

[90] Gerard E. Lynch, *Revising the Model Penal Code: Keeping It Real*, 1 OHIO ST. J. CRIM. L. 219, 220 (2003).

[91] *See* CASEY ET AL., *supra* note 37, at 2–3; CONF. OF CHIEF JUSTS. & CON. OF STATE CT. ADM'RS, RESOLUTION 7: IN SUPPORT OF THE GUIDING PRINCIPLES ON USING RISK AND NEEDS ASSESSMENT INFORMATION IN THE SENTENCING PROCESS (2011).

[92] *See* Starr, *supra* note 5, at 811; *see also* Roger K. Warren, *Evidence-Based Sentencing: Are We Up to the Task?*, 23 Fed. Sent'g Rep. 153, 153 (2010); Conf. of Chief Justs. & Conf. of State Ct. Amd'rs, Resolution 12: In Support of Sentencing Practices That Promote Public Safety and Reduce Recidivism (2007); Roger K. Warren, Nat'l Ctr. for State Cts., Evidence-Based Practice to Reduce Recidivism: Implications for State Judiciaries (2007), https://static.prisonpolicy.org/scans/nic/023358.pdf.

considerations such as inexpensiveness, speed, and efficiency. While it is no doubt true that such pragmatic considerations play a role and will continue to do so, broader signs exist of a growing rejection of retributivism in favor of a more liberal approach focused on rehabilitation. For example, a public survey conducted by Princeton University and sponsored by the National Center for State Courts (NCSC) suggested that, as early as 2006, the public feared that our criminal justice system had become exceedingly punitive and was in dire need of reform.[93] Specifically, survey respondents seemed to favor a return to more rehabilitation-centric practices. Indeed, 75 percent of those who participated in the survey thought sentencing practices need major changes; 79 percent thought many offenders could be rehabilitated; and 59 percent thought prisons were unsuccessful at rehabilitating offenders.[94]

This growing preference for rehabilitation in the United States is also likely to be bolstered by developments abroad. Countries like Norway, Germany, the Netherlands, and Sweden have all implemented policies with an eye towards rehabilitation that have proven enormously effective.[95] Norway has one of the lowest recidivism rates in the world;[96] Germany has an incarceration rate one-tenth that of the United States;[97] and the prison crisis in the Netherlands consists of a "shortage of prisoners."[98] Meanwhile, Sweden has seen such a decline in its number of prisoners and recidivism rates that the country has been closing jails and prisons.[99] Though he concedes

[93] *See* Casey et al., *supra* note 37, at 2 (summarizing and discussing the results of the Princeton survey).

[94] *See id.*

[95] *See* Jeff Rosen, *Germany: Low Crime, Clean Prisons, Lessons for America*, YOUTUBE (Jan. 30, 2017), https://www.youtube.com/watch?v=wtV5ev6813I (discussing the greater emphasis on rehabilitation and societal re-entry of Germany's corrections system and contrasting it from the "historically unprecedented and internationally unique" system of the United States); Nicholas Turner & Jeremy Travis, Opinion, *What We Learned From German Prisons*, N.Y. TIMES (Aug. 6, 2015), https://www.nytimes.com/2015/08/07/opinion/what-we-learned-from-german-prisons.html ("[T]ruly transformative change in the United States will require us to fundamentally rethink *values*. How do we move from a system whose core value is retribution to one that prioritizes accountability and rehabilitation? In Germany we saw a potential model: a system that is premised on the protection of human dignity and the idea that the aim of incarceration is to prepare prisoners to lead socially responsible lives, free of crime, upon release."); Erwin James, *The Society Interview: Prisons and Probation*, GUARDIAN (Nov. 26, 2014, 3:00 AM), https://www.theguardian.com/society/2014/nov/26/prison-sweden-not-punishment-nils-oberg; Christina Sterbenz, *Why Norway's Prison System Is So Successful*, BUS. INSIDER (Dec. 11, 2014, 10:31 AM), https://www.businessinsider.com/why-norways-prison-system-is-so-successful-2014-12; Senay Boztas, *Why Are There So Few Prisoners in the Netherlands?*, GUARDIAN (Dec. 12, 2019, 2:00 AM), https://www.theguardian.com/world/2019/dec/12/why-are-there-so-few-prisoners-in-the-netherlands.

[96] Sterbenz, *supra* note 95.

[97] Turner & Travis, *supra* note 95.

[98] Lucy Ash, *The Dutch Prison Crisis: A Shortage of Prisoners*, BBC (Nov. 10, 2016), https://www.bbc.com/news/magazine-37904263.

[99] *See* James, *supra* note 95.

that this is difficult to explain empirically, Nils Oberg, the director general of Sweden's prison and probation service, believes that it is the country's emphasis on rehabilitation that has led to this transformation.[100]

Examples such as these suggest a changing zeitgeist on a global level in favor of rehabilitating prisoners rather than punishing them for retribution's sake. Such continuing developments abroad will apply pressure to the United States to enact similar reforms. In turn, a shift in focus from punishment to rehabilitation will further perpetuate the use of RNA tools. Moreover, even if it were true that the increase in the use of these tools is due to their ease and inexpensiveness rather than to a growing societal predilection for utilitarian principles, the increased use of these tools will, in and of itself, reshape the criminal justice system and refashion our underlying justifications for punishment.

3. This Declining Interest is Likely to Self-Generate at the Hands of Predictive Tools

This Comment argues in the first instance that the criminal justice system is undergoing a shift in philosophical preference.[101] If it is not, then at a minimum the increasing use of RNA tools will itself bring about said philosophical shift. Most likely, however, criminal justice is caught in a self-reinforcing cycle, whereby a shifting philosophical preference may be causing the proliferation of these tools, while the increasing use of these tools will itself bolster, perpetuate, and entrench that philosophical shift.[102] In other words, there is a two-way causal relationship between the burgeoning theoretical shift described in this Comment and the increased use of these tools in criminal justice: a declining interest in retributivism can both cause, and be caused by, an increasing reliance on predictive tools in criminal justice.

In their joint paper, Richard Re and Alicia Solow-Niederman argue that the proliferation of AI adjudication will both foster and benefit from a shift in adjudicatory values.[103] In doing so, Re and Solow-Niederman juxtapose two models of adjudication.[104] On the one hand, they refer to equitable justice, which places a high premium on judicial discretion.[105] On the other hand, they discuss codified justice, which heavily favors standardization over discretion and "aspires to establish the total set of legally relevant variables *in advance*."[106] While they reject the notion

[100] *Id.* ("Prison is not for punishment in Sweden. We get people into better shape.").

[101] *See supra* Subparts III.A.1, III.A.2.

[102] For the prediction that forms the basis for this contention, see Re & Solow-Niederman, *supra* note 36, at 246.

[103] *See id.*

[104] *See id.* at 252–55.

[105] *Id.* at 252.

[106] *Id.* at 253–54 (emphasis added).

that AI is flatly incompatible with equitable justice,[107] Re and Solow-Niederman argue that "AI adjudication is likely to generate a shift in attitudes and practices that will alter the values underlying the judicial system . . . [and] will tend to strengthen codified justice at the expense of equitable justice."[108] Moreover, they contend that this "shift in values will in turn facilitate [even] greater use of AI adjudication, creating a self-reinforcing cycle."[109] In short, because "[t]he main strengths of AI adjudication are two hallmarks of codified justice"—that is, efficiency and uniformity—AI adjudication will tend to promote codified justice, which in turn will tend to promote AI adjudication.[110]

While Re and Solow-Niederman do not expressly frame their prediction around theories of punishment, this Comment argues that the same phenomenon they predict with respect to adjudicatory styles will take place in criminal justice with respect to underlying theories of punishment. Indeed, because "these tools function as an overly persuasive input," they hold the potential to "anchor [their] users on the chosen philosophy of punishment, to the exclusion of the others."[111] Thus, the "value updating" that Re and Solow-Niederman refer to will also take the form of theory updating in criminal justice.[112] Further still, a similarly self-reinforcing cycle is likely to unfold: AI adjudication will tend to promote utilitarianism at the expense of retributivism, which in turn will tend to promote AI adjudication.

Naturally, it may be argued that even if this argument is correct, it does not follow that we should simply accept this transformation as inevitable, throw our hands up, and embrace these tools. In other words, whether or not this value updating is occurring or will occur, it may nevertheless be true that it is a sort of value updating that is socially undesirable. However, the next Subpart will argue that, not only is criminal justice trending towards utilitarianism—it *should* be.

B. Normative Considerations: Why Criminal Justice *Should* Trend Away from Retributivism

Of course, the mere observation that a self-reinforcing cycle is taking place that will fundamentally transform our criminal justice theory says nothing about whether such a transformation is actually desirable. But this Comment suggests that, not only is this theoretical transformation probably inevitable, it is probably desirable as well. While the primary focus of this Comment is predictive rather than normative, this Part provides two reasons why we might be optimistic about this transformation. First, it turns to findings in the field of neuroscience to call into question the coherence of the notion of moral culpability, thereby challenging the

[107] For a discussion of how "AI adjudication could – counterintuitively – preserve or even foster equitable justice," see *id.* at 258.
[108] *Id.* at 247.
[109] *Id.* at 246.
[110] *See id.* at 255.
[111] Donohue, *supra* note 55, at 666.
[112] *See* Re & Solow-Niederman, *supra* note 36, at 250.

appropriateness of a retributive framing of criminal justice. Second, it draws from far more prosaic considerations to argue that retributivism has failed us and that the current state of affairs itself recommends a shift towards utilitarian aims, chiefly recidivism reduction and rehabilitation.

1. The Free Will Hurdle: Are the Moral Reasoning Concerns Even Coherent?

"[W]e are responsible for wrongs we freely choose to do, and not responsible for wrongs we lacked the freedom . . . to avoid doing."[113] Do such wrongs exist? That is, are there really wrongs that we "freely choose to do"?[114] The notion that human agency or free will exists, such that we are responsible for our actions and choices in some meaningful way, has been called into question in recent decades due to findings in the field of neuroscience.[115] While it is by no means a settled question, it is worth addressing here, as its implications bear directly on the adequacy of AI judges and, more generally, the justifiability of our current criminal justice system as a whole. After all, if humans lack free will, such that in no meaningful way can they be credited or faulted for their actions, would the notion of a system of punishment not be fundamentally undermined? It certainly would, at least if its foundations are retributive in character; perhaps less so if they are utilitarian.[116]

In the early 1980s, the concept of free will was dealt its first serious epistemological blow by the field of neuroscience.[117] Famously, Benjamin Libet and colleagues conducted a series of

[113] MICHAEL S. MOORE, PLACING BLAME: A GENERAL THEORY OF THE CRIMINAL LAW 548 (1997).

[114] *Id.*

[115] See, e.g., Benjamin Libet et al., Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act, 106 BRAIN 623 (1983); see also Editorial, Free to Choose? Modern Neuroscience Is Eroding the Idea of Free Will, ECONOMIST (Dec. 19, 2006), https://www.economist.com/leaders/2006/12/19/free-to-choose.

[116] See Joshua Greene & Jonathan Cohen, For the Law, Neuroscience Changes Nothing and Everything, 359 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON B 1775, 1776 (2004) (recognizing the incompatibility between retributivism and a finding that free will is illusory); see generally, SAM HARRIS, FREE WILL (2012) (arguing that the illusoriness of free will recommends an overhaul of the criminal justice system that rejects retributivism as an unfounded and indefensible theory of punishment). But see Henry T. Greely, Neuroscience, Artificial Intelligence, CRISPR — and Dogs and Cats, 51 U.C. DAVIS L. REV. 2303, 2305 (2018) (rejecting the notions "that 'free will' is necessary, legally or morally, for accountability" or that "the inferred absence of free will makes accountability either legally or morally impossible"); Stephen J. Morse, Neuroscience and the Future of Personhood and Responsibility, in CONSTITUTION 3.0: FREEDOM AND TECHNOLOGICAL CHANGE 113, 119, 122 (Jeffrey Rosen & Benjamin Wittes eds., 2011) (arguing that "free will plays no doctrinal role in criminal law and is not genuinely foundational for criminal responsibility").

[117] See Libet et al., supra note 115. To be clear, these experiments did not deal the concept of free will its first blow as a general matter. The philosophical debate regarding the soundness of the free will claim

experiments designed to measure "the timing of our intentions to act using electrodes on the scalp to measure brain activity."[118] During these experiments, subjects were instructed to press a button whenever they felt inclined to do so. Consistently, researchers recorded brain activity that signaled an imminent press of the button "[r]oughly 350 to 800 milliseconds *before* subjects were consciously aware that they intended to press the button."[119] Naturally, such results tend to undermine the notion that these subjects exercised free will in any meaningful sense. Instead, they appear to have acted mechanistically in response to (and in accordance with) preceding brain activity over which they had no conscious control. Accordingly, the notion of criminal responsibility is left wanting in the way of support.

Unsurprisingly, vigorous debate stemmed from these findings and other related experiments,[120] as the coherence of a construct that had long been viewed as intrinsically human came under increased scrutiny. Across disciplines, a firefight ensued.[121] Notably, a 2003 paper by Joshua Greene and Jonathan Cohen has served as the bedrock for ongoing debate in legal academia.[122] In their piece, Greene and Cohen discuss whether free will is illusory and what impact, if any, such a finding should have on our understanding of criminal justice.[123]

has raged on for centuries. Rather, these experiments dealt free will its first blow from a neuroscientific perspective, rather than a philosophical one. As the stamp of scientific approval carries with it a certain amount of clout that tends to reverberate, these experiments were significant in emboldening and (arguably) substantiating the longstanding philosophical position of free will skepticism.

[118] See Adam J. Kolber, Will There Be a Neurolaw Revolution?, 89 IND. L.J. 807, 814 (2014); see also Libet et al., supra note 115.

[119] Kolber, supra note 118.

[120] *See* John A. Osmundsen, *'Matador' With a Radio Stops Wired Bull*, N.Y. TIMES, May 17, 1965, at A1 (recounting an experiment by a neuroscientist who was able to cause a charging bull to stop moments before impact by pressing a button on a radio transmitter that activated a device inserted in the bull's brain, thereby allowing the scientist to control the animal's actions).

[121] *See, e.g.*, Rosalind English, *Guilty, but Not Responsible?*, GUARDIAN: UK HUM. RTS. BLOG (May 29, 2012, 11:01 AM), https://www.theguardian.com/law/2012/may/29/will-neuroscience-change-criminal-justice (discussing the implications of these findings on criminal justice). *But see* Christopher Slobogin, *Neuroscience Nuance: Dissecting the Relevance of Neuroscience in Adjudicating Criminal Culpability*, 4 J.L. & BIOSCIENCE 577, 578 (2017) (arguing that, while "[n]euroscience does have something to offer court determinations of criminal liability and punishment, . . . it is far from upending the criminal law's basic premise that most choices to commit crime are blameworthy").

[122] *See* Greene & Cohen, *supra* note 116.

[123] *Id.* Of course, such a finding would be pertinent not only to criminal law but to the law in general, as many doctrines in the areas of tort law or even contract law, for instance, are at least partially predicated on the notion of responsibility (i.e., liability). However, such implications are outside of the scope of this Comment, which is focused on criminal law alone.

Answering the first question in the affirmative, Green and Cohen vigorously endorse the view that free will is an illusion.[124] And indeed, their position is not without support. Prominent scholars, philosophers, and scientists, such as Judea Pearl, Sam Harris, Jerry Coyne, and Gregg D. Caruso, have repeatedly voiced their conviction that free will is an incoherent notion.[125] If this is true, fashioning a system of punishment around retributive notions of moral culpability would be utterly indefensible, not to mention completely illogical. After all, if a criminal defendant acted mechanistically, he or she made no choice when acting and cannot be said to deserve punishment.[126] Thus, in order to punish, a different justification would be needed.[127]

Greene and Cohen recognize the fundamental incompatibility between the illusoriness of free will and a retributive understanding of criminal justice.[128] In fact, fundamentally, their piece is a prediction that "neuroscience will change the law . . . by transforming people's moral intuitions about free will and responsibility."[129] Specifically, they predict "a shift away from punishment aimed at retribution in favour of a more progressive, consequentialist approach to the criminal law."[130] This Comment is not persuaded by their predictive claim that the illusoriness of free will alone holds the potential to catalyze such a shift.[131] However, this Comment does endorse

[124] Greene & Cohen, *supra* note 116, at 1783.

[125] *See generally* HARRIS, *supra* note 116; EXPLORING THE ILLUSION OF FREE WILL AND MORAL RESPONSIBILITY (Gregg D. Caruso ed., 2013); RICHARD OERTON, THE NONSENSE OF FREE WILL: FACING UP TO A FALSE BELIEF (2012); Lex Fridman, *Judea Pearl: Causal Reasoning, Counterfactuals, and the Path to AGI: Lex Fridman Podcast #56*, YOUTUBE, at 10:47 (Dec. 11, 2019), https://www.youtube.com/watch?v=pEBI0vF45ic; Jerry A. Coyne, *You Don't Have Free Will*, CHRON. HIGHER EDUC. (Mar. 18, 2012), https://www.chronicle.com/article/Jerry-A-Coyne-You-Dont-Have/131165.

[126] *See* English, *supra* note 121 ("Clearly we need to lock up dangerous people. But there is no sense to the idea that they somehow deserve it. Retributive justice is like requiring us to hate, as well as shoot, a wild animal who escapes from the zoo.").

[127] *See* DAVID M. EAGLEMAN, INCOGNITO: THE SECRET LIVES OF THE BRAIN 151 (2011) (arguing, on the basis of neuroscientific findings, that "blameworthiness is the wrong question"); *see generally* Elizabeth Bennet, *Neuroscience and Criminal Law: Have We Been Getting it Wrong for Centuries and Where Do We Go From Here?*, 85 FORDHAM L. REV. 437 (2016) (discussing the possible impact of neuroscientific advances on the notion of criminal responsibility).

[128] Greene & Cohen, *supra* note 116, at 1777–78.

[129] *Id*. at 1775; *see also* Kolber, *supra* note 118, at 810–11 (noting that, despite the fact that their free will skepticism has received the most attention, Greene and Cohen recognize that their position is not new and instead spend the bulk of their paper discussing their prediction in light of that position).

[130] Greene & Cohen, *supra* note 116, at 1775.

[131] Indeed, the mere fact that the shift predicted by Greene and Cohen had not *already* happened by the time they predicted it itself suggests that it was never likely to happen. As Adam J. Kolber points out, "neuroscientists have made remarkable progress over the last several decades in understanding the brain" and "several experiments . . . seem to vividly remind us that we are mechanisms . . . [y]et these vivid

the underlying normative implication—that an understanding of free will as illusory *should* entail a rejection of retributive principles in favor of consequentialist aims—that forms the basis of their prediction. After all, if free will is illusory, retributivism would be neither morally viable nor intellectually honest.[132] Thus, the shift towards utilitarianism predicted in this Comment (for reasons different from those offered by Greene and Cohen)[133] may not only be imminent but also normatively preferred.

Of course, some might resist this normative claim. For instance, it could be argued that, even if it were true that free will is an illusion, it is such a powerful illusion that we could not possibly fashion a criminal justice system around any other notion if we wanted to. After all, illusory or not, it is inescapably apparent to each of us that we are in charge of the choices we make, and it is impossible to operate except in accordance with that experience. Thus, for all intents and purposes, free will is real. Accordingly, while it may be intellectually interesting to contemplate its illusoriness, it should not be (and, arguably, quite literally could not be) of any real practical consequence in the world.

Indeed, it is impossible to act as though we do not have free will, even if we were to accept the notion that we do not. When asked whether he believed he had free will, Christopher Hitchens cleverly retorted: "I have no choice" but to believe it.[134] There is undeniable wisdom to this framing. There are many truths about the world around which we could not practically organize a society. For instance, we could not possibly structure our lives as though time is anything but linear, even though time as a linear construct may be illusory,[135] because we cannot experience it any other way. The same goes for free will. Nevertheless, while we ourselves may not be able to act as though we lack free will, we can certainly treat others accordingly. In other words, we

displays have yet to change the legal system in any obvious ways." Kolber, *supra* note 118, at 814. Even before Greene and Cohen published their paper, the Libet experiments had already taken place, as well as experiments whereby neuroscientist Jose Delgado stopped a bull in its tracks by pressing a button on a radio transmitter which triggered a device inserted in the bull's brain. *Id.*; Osmundsen, *supra* note 120. If these experiments were not sufficiently vivid so as to refashion the public's conception of free will, it is difficult to imagine what would. It may be that the intuition of free will is simply so salient that no demonstration of its folly will prove vivid or persuasive enough to sway a critical mass.

[132] Indeed, "retribution is based on the dual premises that humans possess free will and that punishment is justified when it is deserved." DRESSLER, *supra* note 58, at 19. Thus, if free will is illusory, then retributivism is unjustifiable.

[133] *See supra* Part III.A.

[134] Jerry Coyne, *Why Do Intellectuals Avoid Discussing Free Will and Determinism?*, WHY EVOLUTION IS TRUE (Jan. 22, 2018, 9:30 AM), https://whyevolutionistrue.com/2018/01/22/why-do-intellectuals-avoid-discussing-free-will-and-determinism/ (discussing Christopher Hitchens' response, among others, to the vexing question of free will).

[135] *See generally* CARLO ROVELLI, THE ORDER OF TIME (Erica Segre & Simon Carnell trans., 2018) (discussing the disconnect between our perception of time and its objective reality).

can fashion our criminal justice system around principles we conceptually hold as true, even if they are not subjectively salient or even available to us moment to moment.[136]

In fact, the law already adapts in this manner in various contexts where the question of free will (or lack thereof) is similarly implicated.[137] Mitigating factors—such as juvenile status, mental incompetence, or overriding emotional reactivity (such as acting in the "heat of passion")—frequently tilt our understanding of criminal behavior and inject our response to it with greater mercy and empathy.[138] The underlying rationale in these cases is that the presence of such factors undermines the notion of criminal responsibility and, by extension, the notion that punishment is deserved. In such cases, we dial down our retributive impulses because our moral intuitions suggest that it would be unfair not to. Although a far more radical claim, a finding that free will is illusory should likewise compel us to dial down our retributive impulses because, well, it would be unfair not to.

But others might resist this argument on the basis of a different normative claim. Chiefly, it could be argued that, if free will is illusory, a shift towards a utilitarian system of punishment would be no more normatively justified than a retributive system because *any* system of punishment would be unjustifiable. After all, if we are not responsible for our actions, then punishment is impossible to justify as a general matter. The only normatively appropriate response, then, would be the total abolition of punishment, an arguably untenable and anarchic alternative.

However, this concern is easily quelled. Living in an organized society entails certain costs and compromises. Even in the face of free will skepticism, a utilitarian system of punishment could still be justified on the basis that incapacitation or incarceration constitute the price some of us must pay when the aggregate welfare would be maximized by us doing so. After all, the price for any given transgression must be borne by some actor if welfare is to be maximized and

[136] Joshua Davis' framing is most useful here. *See* Davis, *supra* note 7, at 74–81. Davis draws a contrast between an understanding of free will from a third-person perspective on the one hand and our experience of it from our first-person perspective on the other. *Id.* From the first-person perspective, it may well be impossible for us to appreciate the illusoriness of free will. But that does not mean that we are incapable of grasping it from the third-person perspective. After all, we understand objectively that the Earth is round, but we certainly cannot appreciate it subjectively, at least not from sea level. Thus, our myopic subjective experience does not preclude us from operating in accordance with the truth. We rely on the Earth's roundness when we make travel arrangements, for example, all the while being unable to subjectively experience the world as round. The same could well be true of free will. We could well structure a criminal justice system that honors the illusoriness of free will and that treats criminal defendants accordingly, even if our subjective experience precludes us from experiencing its illusoriness first-hand.

[137] For a discussion of various mitigating factors like insanity, diminished capacity, or infancy, see DRESSLER, *supra* note 58, at 317, 343.

[138] *Id.*

balance restored. Whether or not a perpetrator can be said to be morally responsible for their offense, (s)he is the only party on whom it would make sense to impose the burden of the offense. After all, it is (s)he who must be rehabilitated, incapacitated, or otherwise treated in order to achieve welfare maximization.

It may be rather unsettling to contemplate the idea of incarcerating or otherwise punishing those who cannot be said to be at fault for their actions. Justifying such measures on purely utilitarian or consequentialist grounds may seem cold, detached, and uncaring.[139] But this Comment posits that the opposite is true. Leaving aside the obvious fact that if free will is truly illusory then retributive punishment does not serve its purpose, this Comment contends that such an understanding of the concept of free will could actually make our justice system much more empathetic, understanding, and thoughtful. Indeed, a systemic transformation could follow that emulates our approach to cases where mitigating factors are present: the illusoriness of free will might well mitigate the harshness of our punishment, shifting our focus from retributivism to more utilitarian aims like rehabilitation.[140]

In sum, this Subpart has posited that the shift towards utilitarianism predicted herein may be normatively preferable. One salient reason to prefer it lies in the arguable illusoriness of free will which—if true—renders a retributive system indefensible while lending support to an increasingly utilitarian (and, specifically, rehabilitative) approach, one that mirrors and amplifies our current approach where mitigating factors are implicated. But there are also far more prosaic reasons to prefer such a trend. These are the subject of the next Subpart.

[139] Indeed, "[r]etributivists criticize [utilitarianism] on the ground that it justifies using persons solely as a means to an end. To the utilitarian, the punished individual is an instrument for the improvement of society. This system ignores the dignity and human rights of the wrongdoer." DRESSLER, *supra* note 58, at 21. But "[u]tilitarians respond that . . . [t]he right each member of society possesses is the right to have the law used for the benefit of the whole community." *Id.* Moreover, "because the wrong-doer is a member of society, he benefits from his own punishment." *Id.*

[140] Some might contend that the illusoriness of free will would work against utilitarianism just as much as retributivism. If there is no free will, does it not follow that the future is already written, in which the case the project of criminal justice would be undermined regardless of its underlying theory? While perhaps intuitive at first glance, this contention conflates determinism and fatalism. *See* HARRIS, *supra* note 116 (explaining why the illusoriness of free will entails determinism but not fatalism, and why that distinction means the future is not already written). But the more relevant point harkens back to the distinction between the first-person and third-person perspectives. *See supra* note 136. The illusoriness of free will changes nothing from a subjective point of view. It only changes our objective understanding of the behavior of others. From a subjective point of view, we have no choice but to carry on making choices; it is impossible not to. *See id.* Thus, the only impact the illusoriness of free will might have on our lives is to inform (and perhaps change) how we think of others and their behavior, such that we might treat them more compassionately and mercifully.

2. Clearing the Free Will Hurdle, Practical Persuasions Remain

Beyond questioning the soundness of the free will claim, there are numerous other philosophical grounds for challenging retributivism and preferring utilitarianism.[141] Indeed, a centuries-old debate has been raging between those who favor retributivism and those who prefer some flavor of utilitarianism or consequentialism.[142] However, this Comment does not presume to resolve a debate that has vexed philosophers for millennia. Instead, this Comment grounds its second normative claim in far more prosaic facts, namely the checkered track record of retributivism and the current state of our criminal justice system.

Regardless of which theory could claim philosophical superiority in the abstract, the fact of the matter is that our current criminal justice system, which for decades has been largely characterized by retributive leanings,[143] is in a shameful state of disarray.[144] As of year-end 2016, the most recent year for which statistics from the Bureau of Justice Statistics are available as of this writing, a staggering 6,613,000 people were subject to corrections systems in some capacity (either as probationers, parolees, or prisoners).[145] Of those roughly 6.5 million people, a whopping 2,162,400 were incarcerated, either in prisons or jails.[146] Indeed, "[t]he imprisonment rate [in the United States] is more than five times the average incarceration level of other

[141] *See generally* Whitman, *supra* note 76; Mirko Bagaric & Kumar Amaraskekara, *The Errors of Retributivism*, 24 MELB. U. L. REV. 124 (2000); Russell L. Christopher, *Deterring Retributivism: The Injustice of "Just" Punishment*, 96 NW. U. L. REV. 843 (2002); Robert Weisberg, *Reality-Challenged Philosophies of Punishment*, 95 MARQ. L. REV. 1203 (2012); Chad Flanders, *Retribution and Reform*, 70 MD. L. REV. 87 (2010).

[142] *See* sources cited *supra* note 141; *see also* Michael Tonry, *Introduction* to WHY PUNISH? HOW MUCH?: A READER ON PUNISHMENT 3, 3 (Michael Tonry ed., 2011) (describing the debate).

[143] *See* Whitman, *supra* note 76, at 87 ("The old belief in rehabilitation . . . has been widely abandoned. In its place has come the triumph of American neo-retributivism. Thirty years ago, a new generation of philosophers demanded a criminal law founded on blame—on unembarrassed condemnation where condemnation is warranted. . . . Indeed, we have had nothing less than a *renaissance of retributivist punishment philosophy* . . . ." (emphasis added)); *see also* Slobogin, *supra* note 22, at 2 ("Beginning in the 1970s, a sentencing revolution took hold . . . based predominantly, and occasionally entirely, on a desert philosophy.").

[144] *See* Flanders, *supra* note 141, at 87 ("The last twenty years have seen a . . . 'retributivist revival' . . . [b]ut, those same twenty years have also seen increases in the length of criminal sentences, in the amount of activity subject to criminal sanction, and in the sheer number of people behind bars.").

[145] DANIELLE KAEBLE & MARY COWHIG, U.S. DEP'T OF JUST., NCJ 251211, CORRECTIONAL POPULATIONS IN THE UNITED STATES, 2016, at 1 (2018), https://www.bjs.gov/content/pub/pdf/ppus16.pdf.

[146] *Id.* at 1–2.

developed countries,"[147] and while "'[t]he United States has 4% of the world's population,'" it has "'21% of the world's prisoners.'"[148] In short, the United States has "the most punitive criminal justice system on Earth."[149]

Perhaps most relevant for present purposes is the fact that recidivism rates in the United States are difficult to overstate. "[O]f the approximately 95 percent of imprisoned offenders who are ultimately released back into free society, most reoffend."[150] In fact, "[n]early three-quarters of released prisoners reoffend and are arrested within five years of release and 60 percent of them are reconvicted."[151] One's philosophical predilections aside, something must be done to remedy this crisis. This Comment subscribes to the notion that one of the primary ways in which the current crisis may be ameliorated is by curbing the problem of recidivism.[152] If this is true, then a shift towards utilitarianism—with a focus on rehabilitation—comes normatively recommended by the practical reality of our criminal justice system. After all, the reduction of recidivism is an inherently forward-looking aim, which is inextricably tied to the maximization of aggregate welfare and indifferent to the moral culpability that attaches to past acts.

[147] Mirko Bagaric et al., *Mitigating America's Mass Incarceration Crisis Without Compromising Community Protection: Expanding the Role of Rehabilitation in Sentencing*, 22 LEWIS & CLARK L. REV. 1, 3 (2018) (citing MELISSA S. KEARNEY ET AL., THE HAMILTON PROJECT, TEN ECONOMIC FACTS ABOUT CRIME AND INCARCERATION IN THE UNITED STATES 10 (2014)).
[148] *Id.* (quoting TASKFORCE ON MASS INCARCERATION, N.Y. CITY BAR ASS'N, MASS INCARCERATION: WHERE DO WE GO FROM HERE? 1–2 (2017)).
[149] Bagaric, *supra* note 76, at 1227; *see also* Adam Liptak, *Inmate Count in U.S. Dwarfs Other Nations'*, N.Y. TIMES, (Apr. 23, 2008), https://www.nytimes.com/2008/04/23/us/23prison.html ("[T]he United States leads the world in producing prisoners . . . ."); Christopher Slobogin, *How Changes in American Culture Triggered Hyper-Incarceration: Variations on the Tazian View*, 58 HOW. L.J. 305, 307 (2015) (noting that the United States incarcerates its citizens at a rate roughly six times that of European countries).
[150] Bagaric et al., *supra* note 147, at 4 (emphasis added).
[151] *Id.* at 5 n.10 (citing NATHAN JAMES, CONG. RESEARCH SERV., RL34287, OFFENDER REENTRY: CORRECTIONAL STATISTICS, REINTEGRATION INTO THE COMMUNITY, AND RECIDIVISM 6–7 (2015)); *see also* James Gilligan, *Punishment Fails. Rehabilitation Works*, N.Y. TIMES (Dec. 19, 2012, 11:43 AM), https://www.nytimes.com/roomfordebate/2012/12/18/prison-could-be-productive/punishment-fails-rehabilitation-works ("Two-thirds of prisoners reoffend within three years of leaving prison, often with a more serious and violent offense.").
[152] *See* Bagaric, et al., *supra* note 147, at 5–6; Slobogin, *supra* note 22, at 5 ("Overall, the impact of energetically incorporating risk assessment into sentencing might be a significant reduction in the prison population which, in the United States at least, has burgeoned to alarming proportions. To the extent prison is criminogenic, such policies might also reduce the overall crime rate, by exposing fewer offenders to prison's ill effects and by facilitating identification of causal risk factors that can be the focus of rehabilitation efforts (many of which can and should take place outside of prison).").

Put differently, to the extent that the practical reality of our justice system calls for a reduction of recidivism, it calls for utilitarianism rather than retributivism. After all, in light of the fact that our prisons and jails are unsustainably overcrowded—and our offenders extremely likely to reoffend and return to those overcrowded facilities—is it reasonable for our focus to be on determining how long an individual deserves to be incarcerated for his or her offense? Should our focus not be chiefly placed, instead, on how to get offenders out of prisons and jails and back into society with a lower likelihood of reoffending?[153] This Comment contends that the latter position is the more sensible alternative, as the former perpetuates a system that currently threatens to crumble under the pressure of its own weight.

In 2007, the CCJ and COSCA were already advocating for an approach focused on the reduction of recidivism.[154] Moreover, they endorsed RNA tools as particularly effective means for achieving such reductions and effectuating rehabilitative aims.[155] In a 2018 paper, Mirko Bagaric, Gabrielle Wolf, and William Rininger likewise stressed the importance of a shift towards rehabilitative practices, urging greater reliance on empirically based tools to effectuate such aims.[156] In fact, the authors recommend that:

> [C]ourts should be required to use empirically-tested means of predicting an individual offender's likelihood of rehabilitation and recidivism . . . . Specifically, we suggest that courts make greater use of "risk assessment" and "risk and needs assessment" tools, which evaluate the likelihood of offenders reoffending and the measures that could best reduce those individuals' risk of recidivism.[157]

While it may seem counterintuitive to entrust the determination of a human defendant's likelihood of rehabilitation to an artificially intelligent agent—and even more radical to require it—the rationale is inescapable in light of the sheer lack of rigor and reliability inherent to the alternative. Indeed, "[a]t present, an offender's prospects of rehabilitation are almost totally determined by a sentencing judge's impressions" and "raw intuition."[158] It should not be difficult to appreciate why such a method is unreliable, undesirable, and—frankly—indefensible.

In sum, the practical realities of our criminal justice system urge a shift away from retributivism and toward utilitarianism. Specifically, an increased focus on rehabilitation and recidivism

[153] *See* Slobogin, *supra* note 22, at 9 ("Prison is not the only way, and certainly not the most effective way, of preventing or reducing reoffending. Risk management alternatives involving treatment, counselling, job training, and surveillance can curtail recidivism.").
[154] *See* CASEY ET AL., *supra* note 37, at 3.
[155] *Id.*
[156] Bagaric et al., *supra* note 147, at 7–8.
[157] *Id.* (emphasis added).
[158] *Id.*

reduction is critical at this juncture, from both an economic and a moral point of view.[159] Furthermore, utilitarian efficacy is supported by empirical studies.[160] Because RNA tools arguably offer the best means for effectuating such aims,[161] they come as equally recommended by the practical reality of our criminal justice system as they do by the more esoteric notion that moral accountability may be an incoherent concept.

IV. A Shift in Underlying Theory, But at What Cost? Possible Objections and Responses Regarding Endangered "Soft" Values

Even if a shift towards utilitarianism were indeed normatively preferred, some nonetheless caution that increased standardization will come at the expense of certain values which are at least as important as prediction. For example, Andrea Roth worries that values such as mercy, equity, and dignity—which she terms "soft" values—may be pushed aside as criminal justice becomes increasingly mechanized.[162] The threat to these soft values, Roth argues, stems from the fact that their virtue is difficult to define and quantify.[163] As a result, "they might be inadvertently set aside" in a heavily mechanized criminal justice system.[164] But, she cautions, these are virtues with "a pedigreed history in American criminal justice" and must be preserved.[165] Commentators discussing other values, such as racial or gender equality, echo Roth's concerns.[166]

It is hardly contestable that these values are important, and it seems quite sensible to worry that increased automation will threaten them. To address this thorny problem, the framework provided by Christopher Slobogin proves a useful starting point.[167] In his work, Slobogin proposes three principles that should govern our evaluation of modern risk assessment tools in

[159] *See* Slobogin, *supra* note 22, at 6 ("Preliminary research in Virginia suggests that use of RAIs can substantially reduce the proportion of non-violent prisoners in prison, while minimizing re-conviction rates. The Justice Research Institute has estimated that the evidence-based sentencing programs now in existence, focused on risk rather than desert, will reap about 4.6 billion dollars in savings in the next ten years.").

[160] KAREN HESELTINE ET AL., AUSTL. INST. OF CRIMINOLOGY, PRISON-BASED CORRECTIONAL OFFENDER REHABILITATION PROGRAMS: THE 2009 NATIONAL PICTURE IN AUSTRALIA 14 (2011), https://aic.gov.au/publications/rpp/rpp112.

[161] Hamilton, *supra* note 42, at 277 ("Retributive . . . orientations are less amenable to evidence-based practices while utilitarian and rehabilitative foci would embrace them.").

[162] Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245, 1282 (2016).

[163] *Id.*

[164] *Id.*

[165] *See id.*

[166] *See generally* Starr, *supra* note 5; Michael Gorelik, *Descending Back Into Plato's Cave: The Use of Artificial Intelligence in Criminal Sentencing*, 9 L.J. SOC. JUST. 150, 163 (2018); Angwin et al., *supra* note 5.

[167] *See* Slobogin, *supra* note 77, at 586–92.

criminal justice: the fit principle, the validity principle, and the fairness principle.[168] For purposes of this discussion, only the third principle—fairness—is relevant. Intended to safeguard against the erosion of "soft" values, the fairness principle calls for "balancing the incremental validity of each risk and protective factor against the extent to which it undermines the autonomy and dignity values that undergird the criminal justice system."[169] In other words, the validity of any given factor must be weighed against the fairness concerns it might raise. But what do these fairness concerns consist of?

Building on Slobogin's framework, this Comment construes "fairness" as comprising two distinct buckets, each circumscribing a distinct set of endangered "soft" values: the Dignity bucket and the Disparity (or Discrimination) bucket.[170] Let us, therefore, address each in turn.

A. The Dignity Claim

The Dignity bucket encompasses values like mercy and equity. Needless to say, these are meaningful values within criminal justice. Thus, to the extent that the use of AI in criminal justice is likely to erode them, mechanisms for their protection must be proposed and debated. But before delving into what these mechanisms may look like, we must first define the substance of these values. In other words, what exactly is meant by "fairness" with respect, specifically, to the Dignity variety?

Slobogin defines fairness as relating to "the traditional assumption that criminal justice dispositions should be related to blameworthy conduct."[171] Similarly, Roth summarizes her concern around these soft values by arguing that "we should not allow [mechanization] to eliminate moral condemnation from the equation," which in her view should "retain its rarity and gravity and signaling effect."[172] In other words, these soft values, as construed by scholars like Slobogin and Roth, would seem to be rooted in the idea that punishment should be predicated on the defendant's moral culpability. A sentence is unfair, then, if it does not track how much the defendant is to be blamed for what they did. Soft values, in turn, are those values like mercy and equity that are necessary to protect the defendant against such an outcome—that is, to ensure that a sentence does not overpunish by disproportionately tracking the defendant's moral culpability.[173]

[168] *Id.*

[169] *See* Slobogin, *supra* note 22, at 8.

[170] *See id.* at 18–20; Slobogin, *supra* note 77, at 590.

[171] Slobogin, *supra* note 77, at 589; *see also* Slobogin, *supra* note 22, at 10 ("[T]he fairness principle is meant to address only the concern that risk assessment is insufficiently cognizant of the traditional tenet that criminal justice dispositions be based on blameworthy conduct.").

[172] Roth, *supra* note 162, at 1304.

[173] *See id.* at 1285–86.

But, of course, this Comment has challenged both the normative propriety and the logical coherence of a system of punishment premised on the notion of moral culpability. Moreover, normative claims aside, this Comment has predicted a trend towards utilitarianism at the expense of retributivism. In an increasingly utilitarian system, an understanding of fairness premised on moral culpability may not be the most apt. Does that mean that fairness considerations are just not relevant under a utilitarian framing? Of course not. But it does mean that the concept of fairness might be reframed so as to more neatly fit the aims of utilitarianism. Namely, it must be untethered from the notion of moral culpability.

One possibility is to evaluate the fairness of a sentence not on the basis of its proportionality to the defendant's moral desert, but on the basis of how tailored it is to the relevant utilitarian outcome it purportedly strives to achieve. In other words, a sentence is fair if it is no more (and no less) harsh than is necessary to maximize aggregate welfare. Of course, no utilitarian sentence will ever be perfectly tailored to the precise level of harshness that will maximize wellbeing, but neither would a retributive sentence ever be perfectly proportionate to a defendant's moral blameworthiness. The relevant point is simply that the criteria for fairness has changed: instead of asking that a sentence track the moral culpability of a defendant, we ask that it track the maximization of aggregate welfare.

If fairness is so defined—and if AI is properly trained on data that allows it to predict the sentence or treatment most likely to achieve welfare maximization by the least restrictive means possible—then overpunishment becomes much less likely and the need for soft values (or, rather, the frequency with which that need arises) might decrease. Moreover, fewer perverse incentives to overpunish might exist because overpunishment may in fact impede the maximization of aggregate welfare—if disproportionately punished individuals are more likely to recidivate, for example. In that sense, the interests of the defendant, which soft values like mercy and equity are meant to protect, would be more aligned with the interests of society as a whole under a utilitarian framing than a retributive framing: whereas a retributive society may be overrun by emotion in the face of an especially heinous crime, resulting in overpunishment, a utilitarian society would be incentivized to fashion a more narrowly tailored judgment. Put succinctly, then, a utilitarian framing may well threaten dignity values less than a retributive framing.

Moreover, these values may be further safeguarded by adhering to the very fairness principle that Slobogin prescribes: "[t]o minimize further any affront to dignity associated with [risk assessment instruments], risk assessment should be based as much as possible on dynamic or 'causal risk factors[]' . . . . [R]isk factors that can be changed through intervention and thus focus on traits that the person can do something about."[174] In a system that emphasizes dynamic factors over static factors such that defendants are given the opportunity to reform, automation may pose less of a risk to dignity values. After all, automated or not, a system which at its core

[174] Slobogin, *supra* note 77, at 593 (emphasis added).

is concerned with identifying the best means of rehabilitating defendants for reentry into society is arguably a system inherently more merciful, equitable, and dignified than one which is not.

B. The Disparity Claim

The Disparity subset of fairness concerns implicates values of equality. For example, while RNA tools do not explicitly consider race as a factor, they do consider factors that could be reasonably construed as proxies for race, such as demographic or socioeconomic factors.[175] They also often consider other (arguably) immutable characteristics, such as gender and age.[176] Because they make distinctions based on such broad categories, critics worry that these tools unfairly discriminate.[177]

One way to mitigate these concerns about equal treatment may also harken back to the demands of Slobogin's fairness principle. As discussed, Slobogin argues that, to safeguard against the erosion of fairness, the usefulness of any given risk factor must be weighed against the fairness concerns it might raise before deciding on its inclusion.[178] For example, because race is a poor predictor of risk on the one hand and an equitably fraught concept on the other, it should be excluded from RNA instruments because its potential costs far outweigh its potential benefits.[179] However, if factors such as gender or age tend to improve predictive accuracy, as they appear to, then perhaps the balance weighs in favor of their inclusion if they are more helpful than fraught. In the words of Slobogin, "a normative judgment must be made about when a level of correlation is so low it requires a factor's exclusion."[180]

Even in a system where AI replaced human judges altogether, unfair disparate treatment may at least be mitigated so long as these normative judgments are taken seriously, such that factors likely to animate the Disparity concerns never make it to sentencing in the first place. To be sure, this is at best a start. Ample thought and deliberation will have to go into how we might guard against the costs of automation as we embrace its benefits. However, this Comment insists that we need not shun automation altogether, so long as we anticipate the threat it poses to certain values, such that we may act decisively and preemptively to preserve them.

Conclusion

The use of AI in criminal justice will no doubt have a transformative impact on criminal justice. But this Comment has argued that, at least with respect to their moral handicap, the use of AI

[175] *See* Hamilton, *supra* note 42, at 240–41, 261–62.
[176] *See* Slobogin, *supra* note 77, at 590.
[177] *See generally* Starr, *supra* note 5.
[178] *See* Slobogin, *supra* note 22, at 12.
[179] *See id.*
[180] Slobogin, *supra* note 77, at 593.

predictive tools in criminal justice is less problematic under a utilitarian conception than a retributive framing. Specifically, the moral handicap of AI tools may be of less consequence to judicial decision-making when a judge's focus is on utilitarian interests such as maximizing aggregate welfare, rather than on retributive interests such as fashioning punishment proportionate to a defendant's moral culpability; that is, the punishment that is deserved. After all, while the latter by definition requires a moral judgment, it is less clear that such a judgment is truly necessary for a judge to properly effectuate utilitarian aims at sentencing.

For example, if aggregate welfare would be maximized by rehabilitating a criminal defendant, said rehabilitation could be achieved by predicting what punishment or treatment is most likely to result in rehabilitation, defined and captured as a reduction in that defendant's recidivism. Such a prediction, in turn, could be made on the basis of data mined from past cases involving comparable defendants, offenses, and circumstances where a reduction in recidivism was achieved. AI tools could sift through such data, identifying useful patterns upon which to predict the measures most likely to result in the desired reduction of recidivism in any given case. Because humans are dwarfed by AI in their data processing and pattern recognition capabilities, it follows that—to the extent that criminal justice is concerned with predicting and reducing recidivism rather than retribution—AI tools may indeed be preferable to the human alternative, notwithstanding their alleged moral handicap.

So, what next? Should humans simply step aside and let AI take over the courtroom, displacing human judges altogether? It is still too early to know the answer. That said, a number of possibilities emerge. For one, we could stick with the status quo. That is, AI tools could continue to be used in criminal justice exclusively as they are now: to support or assist—but never to replace—human judges. At the other extreme, AI might replace human judges altogether, leaving no role for the human in judicial decision-making. But there may be a middle ground. AI tools could replace human judges—with some exceptions. In other words, a role could be preserved for human discretion in cases where a particular result runs afoul of the fairness values discussed in this Comment or, in some other manner, offends our human sensibilities. For example, a human could always be present in the room at the moment of sentencing, acting as a check on the AI judge, whose judgment will remain undisturbed so long as it aligns with our utilitarian aims and does not implicate exceptional circumstances. Under this approach, the human judge-assistant would have the ability to override the judgment in real time so long as, again, certain exceptional circumstances are implicated.

Alternatively, a human panel could exist to which decisions of the lower AI courts could be appealed. That is to say, we might limit the replacement of human judges by AI judges to the lower state and federal trial courts, reserving the role of humans within Courts of Appeals.[181] An AI trial court could hear the facts, apply the law to those facts, and—based on the data on which it has been trained and on the particular defendant's criminogenic needs and risks—issue the judgment it predicts is most likely to maximize the aggregate welfare. The defendant, however, would retain the right to appeal the judgment to a human Court of Appeals.

This Comment endorses and recommends the latter approach, at least at the start. Ultimately, however, the chief focus of this Comment is not to recommend a particular course of action moving forward. This is an incredibly complex transition implicating concerns that stem from a wide variety of disciplines and perspectives. Any pithy recommendation made in the span of a few concluding paragraphs would fail to do justice to the complexity of the issues. Thus, the chief focus of this Comment is instead to shed light on an important nuance that should inform the conversation around the issues raised by the use of AI in criminal justice, whether as assistants to human judges or as judges themselves. This Comment posits that it is not enough to simply point to the moral handicap of these tools and, on that basis, dismiss their value or categorically reject the adequacy of AI judges. Instead, we must recognize that these tools take on a different light depending on the theoretical lens through which they are viewed.

Thus, while it is undeniably important to scrutinize these tools and our increased reliance on them, we must first confront the realities of our present system and think deeply on what kind of criminal justice system we want to have. Only after we have carefully defined the principles by which we want to be guided can we fairly evaluate what role, if any, these tools should play in our criminal justice system.

[181] *See* Sourdin, *supra* note 7, at 1124 (noting that humans "may play an appellate or review function only" as AI comes to replace them in adjudicatory decision-making). For a similar discussion, see Volokh, *supra* note 41, at 1190.

> Maybe, though, there are some decisions that . . . so depend on debates about our most important values [] that we as humans won't want to delegate them to an AI, no matter how high quality that AI's decisionmaking might seem. . . . If we really want such decisions to be made by humans, we can easily construct rules that allow it. For instance, there could be a procedure for discretionary review of the AI Supreme Court's decisions by an all-human Highest Constitutional Council.

Id.