Minding the Machines
On Values and AI in the Criminal Legal Space

There was but one passing reference to "core values" over the course of a recent U.S. Senate Judiciary hearing on artificial intelligence [AI] in criminal investigations and prosecutions.[1] This is typical. Even in spaces like the criminal legal system, where the specters of racial injustice and inhumanity loom so large, the technological sublimity of AI can be awfully distracting. People have long looked to technology to duck the hard problem of values. "[W]e have tended to believe that if we just *had more information*, we could make better policy," observes University of Nevada's Lynda Walsh in *Scientists as Prophets*. "But no matter how much data we could lay hands to—even if it were LaPlace's Demon itself—values would still stand in the way."[2]

If anything is clear about advanced AI, it is that there is much we don't know and even more that we can't begin to predict. Consider that the "generative AI" we have witnessed over the past 18 months—AI which produces *autonomous* human-impersonating content—was largely unforeseen. It's now being attributed to AI's "emergent abilities."[3]

Across sectors, most observers acknowledge that AI is a game-changing technology. The Financial Industry Regulatory Authority is illustrative: using AI, it now processes "a peak volume of 600 *billion* transactions every day to detect potential abuses," making the regulator "one of the largest data processors in the world."[4] Tell-

Authors

Julian Adler, Chief Innovation and Strategy Officer, Center for Justice Innovation

Jethro Antoine, Chief Program Officer, Court Reform, Center for Justice Innovation

Laith Al-Saadoon, Principal Prototyping Architect, Amazon Web Services

ingly, many of the people closest to the leading edges of AI development are sounding the loudest alarms about its capabilities. "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war," warned the Center for AI Safety in 2023.[5]

AI has the potential to supercharge, not mitigate, the uglier sides of humanity, much like, as one journalist puts it, "a fun-house-style... mirror magnifying biases and stripping out the context from which their information comes."[6] Advanced AI is "not just another technology," contends Nick Bostrom, Director of the Future of Humanity Institute at the University of Oxford. It is not "another tool that will add incrementally to human capabilities."[7] Echoing countless dystopian projections of the future, the Center for AI Safety predicts AI systems will likely "become harder to control" than previous forms of technology; among other disquieting scenarios, these systems could "drift from their original goals" and "optimize flawed objectives."[8]

Electronic monitoring and risk assessments are cautionary tales.

The criminal legal space offers little refuge from these scenarios. "Artificial intelligence has the potential to be a permanent part of our criminal justice ecosystem," writes the National Institute of Justice's Christopher Rigano approvingly."[9] Roughly five million people in this country are under some form of mandated community supervision—probation, chiefly, but also parole—a phenomenon scholars have dubbed "mass supervision."[10] The National Institute of Justice identifies what it sees as opportunities for AI to facilitate real-time assessment of risk and need for this population, along with mobile service delivery and what the agency calls "intelligent tracking of individuals." The Institute's Eric Martin and Angela Moore offer the following biometric use case to illustrate the potential: "an AI wearable device could monitor biological data assessing an individual's stress and mood and send alerts to the community supervision officer that the person may be in a risky situation." They contend this would enable probation officers to focus their limited resources "with surgical precision at times when recidivism is most likely." Complementing the current use of GPS-based electronic monitoring devices, they tout the potential for an AI-enhanced device to "engage with the individual to mitigate the precarious situation" before an officer can respond to a notification. For example, the AI might intervene "by encouraging the person to leave a risky location or engage in programming such as cognitive-behavioral therapy."[11]

Do these AI-enabled scenarios signify progress or existential risk? It depends on who you ask. "Artificial Intelligence in Criminal Court Won't Be Precogs" reassures the Justice Innovation Lab's Director of Analytics Rory Pulvino, invoking the future-anticipating abilities of some of the characters in the Hollywood thriller, *Minority Report*. While Pulvino cautions that "without proper oversight, even...innocuous systems may perpetuate or worsen biases in the criminal justice system," the reference to popular culture is intended to temper some of the more

dystopian fears about AI.[12] In *The Atlantic*, technology and civil-rights lawyer Frederick Pinto offered the sobering observation that, in legal systems, "less humanity could lead to more fairness."[13] But Paige Collins and Matthew Guariglia of the Electronic Frontier Foundation are less sanguine. In "Cops Running DNA-Manufactured Faces Through Face Recognition Is a Tornado of Bad Ideas," they caution of "dangerous technology."[14]

Whether AI represents a qualitative, or merely quantitative, evolution from previous implementations of big data in the criminal legal space doesn't change the fatal "garbage in, garbage out" phenomenon that undermined the erstwhile promise of risk assessment algorithms.[15] Our models are only as good as the data feeding them—if that data is marred by systemic racism (in, say, policing, prosecution, or sentencing), then the outcomes generated by AI will be similarly marred.

> AI could supercharge, not mitigate, the uglier sides of humanity.

Reflecting a pragmatic position in a piece for the American Bar Association, Judge Herbert B. Dixon Jr. (Ret.) observes that "AI is here to stay and will play an increasing role in our personal lives and the criminal justice system." He aims for a middle ground: "We cannot look away from the potential it offers in improving law enforcement and our courts. It may be necessary, however, to step back and take stock of AI's implications."[16]

One thing is for certain: criminal legal reformers need to be actively and mindfully engaged in the proliferation of AI. "Technology is not neutral," headlines the Smithsonian in its AI Values Statement.[17] Translation: there is no perch above the fray. By way of a framework for engagement, we offer three preliminary recommendations, which we hope will be further refined and expanded in the days ahead through cross-sector dialogue, reflection, and application on the ground.

Prioritize values over technology. In 2012, inspired by Michael Lewis's book on the application of big data to baseball,[18] the criminal legal reform field became enamored by the prospect of "Moneyballing Criminal Justice" through the use of predictive risk assessment algorithms.[19] Yet there was scant discussion about the values that should guide and constrain the use of so-called "evidence-based, neutral information" to inform decisions that would affect people's liberty interests.[20] With a 2018 op-ed in *The New York Times* titled "The Newest Jim Crow," Michele Alexander helped to spur that discussion, leveling the critique that risk assessment algorithms "appear colorblind on the surface but they are based on factors that are not only highly correlated with race and class, but are also significantly influenced by pervasive bias in the criminal justice system."[21] Alexander was in some respects echoing warnings sounded by *ProPublica* in a pioneering analysis of racial bias in risk assessment tools in 2016,[22] and yet many in the field were brought up short. Indeed, that belated awakening was something of a tell: the practice of algorithmic risk assessment remains widespread today[23] and efforts to foreground values appear to have come too little, too late.[24]

The result is a practice that remains fraught and polarizing.[25]

> AI can only be tamed by the stubborn application of human values.

Or consider a seemingly benign technology like the global positioning system [GPS], the product of military and civilian partnerships in the early 1970s, which in 1983 would find its way onto the ankle of the first electronically-monitored person in Albuquerque, New Mexico. Decades later, electronic monitoring is ubiquitous in the criminal legal system and is increasingly viewed by critics of its rise not as an alternative to mass incarceration but as an extension of it.[26] "Many reformers rightly point out that an ankle bracelet is preferable to a prison cell," argued Alexander in her 2018 piece. "Yet I find it difficult to call this progress. As I see it, digital prisons are to mass incarceration what Jim Crow was to slavery."[27] Perhaps the field should have anticipated the harms this technology would engender, but the Albuquerque Metropolitan Court judge's immediate desire to arrange for a house arrest in 1983—and many more judges and prosecutors to follow—eclipsed any consideration of values. Efforts to implement policies to mitigate the harms of monitoring are still playing catch-up.[28]

The implementation of AI is well underway in criminal legal systems, including ever more complex risk assessment algorithms,[29] but there is still an opportunity—and an urgent need—to advocate for a values framework that prioritizes transparency, fairness, and the wellbeing of individuals and communities.[30] To avoid the problems described above, only a deep front-end commitment to values will suffice.

Stay active, curious, and informed about technology. AI is unlike any technology humankind has ever encountered. Its black boxes contain more than the comparatively simple codes, formulae, and algorithms of our past; they metabolize vast repositories of our interpretations, decisions, preferences, and biases—and they *generate* content.[31] This poses a challenge to criminal legal reformers who, when it comes to technology, tend toward passivity and indifference. This may be because reformers, who are rarely technologists or data scientists by training, often use technologies developed for other purposes—they are rarely at the product design tables contributing to how the technology can best serve those who will use it or be impacted by its use.

Compounding matters, some technologists behind AI maintain there is a trade-off between AI's accuracy and its comprehensibility or "explainability," i.e., AI becomes "more optimal" the more its complexity eludes our understanding.[32] Yet as Duke University's Brandon L. Garrett and Cynthia Rudin point out: "black box AI performs predictably worse in settings like the criminal system." They make the case for "glass box" AI systems "designed to be fully interpretable by people" such as judges and attorneys.[33] Yet even inside a glass box, it will take sustained effort for criminal legal reform writ large to understand AI well enough to influence its trajectory. And it will require a similar effort for practitioners and advocates to develop an understanding of how AI tools can and will

inevitably fail, and how the resulting harms can be mitigated through policy and regulation.[34]

Resist the siren song of efficiency long enough to weigh unintended consequences. Criminal legal systems are notoriously under-resourced and overwhelmed. AI purports to create efficiencies in operations without the appreciable budgetary requirements of human staffing. Some of these efficiencies would be routine—analogous to the financial regulator's newfound capacity for 600 billion daily transactions. Others could prove expansive, with the potential to keep more people from experiencing the traumas of incarceration.

With respect to bias identification and mitigation, for example, AI could be used to analyze the text of laws, sentencing guidelines, court decisions, transcripts, and other relevant documents to identify linguistic patterns, wordings, or criteria that may encode or enable racial biases in charging and sentencing. By surfacing these, AI could point toward alternative language and help prompt changes to make the system more equitable. Relatedly, AI has the potential to highlight racial disparities at scale—analyzing datasets to identify where people of color are receiving significantly harsher sentences compared to their white counterparts for similar crimes. AI could uncover many specific cases of racially biased extreme sentencing and accelerate advocacy for new sentencing guidelines. And for incarcerated people, AI could expedite petitions for compassionate release and clemency, as well as second-look initiatives. AI could help analyze and summarize large volumes of applicant case files, quickly identifying strong candidates for a second chance and generating summaries that facilitate faster reviews by pardon boards, judges, or executives with the authority to modify sentences. This could help get more people out of prison more quickly.

In the pretrial space, AI could be used to finally realize the potential of ability-to-pay determinations and transform how lengthy rap sheets are reviewed and analyzed in bail hearings to surface evidence of such ills as bad police stops, disparate policing practices, the criminalization of mental illness, anti-trans enforcement… The list could go on.

> Unintended consequences follow any introduction of new technology.

These are all welcome applications of AI to the criminal legal space, ones where the technology would have no direct negative bearing on people's liberty interests, to the contrary. Yet we must heed the lessons of algorithmic risk assessment and electronic monitoring alluded to above—the consequences, often unanticipated and unintended, that flow from any introduction of new technology to the justice space. Consequences such as the exacerbation of racial and ethnic disparities in jails and prisons and a marked extension of the carceral system beyond its brick-and-mortar facilities.

The field must proceed warily before jumping headlong into AI's mind-bending functionality. Although there are countless regulatory schemes to be devised, we reiterate that a shared values framework—what does a jurisdiction want to use AI *for*—is our best bet for

guiding the responsible development and use of AI in the criminal legal context.

Moving Forward

Innovations in data science and technology have generally proved polarizing in the criminal legal reform space. While many systems rush to overreliance, reformers and advocates for the incarcerated often cry foul, foreseeing unintended consequences and old carceral wine in new Orwellian bottles. Students of the best twentieth-century writing on technology will be quick to point out that technology has *never* been merely a "tool"—a neutral apparatus for, in this instance, the administration of justice.[35] But even viewed through that capacious lens, AI is a different beast, one that will not be tamed via additional technology, but only by the stubborn effort to harness it through the application of human values. "Achieving technical design that soundly incorporates values requires not only competence in the technical arts and sciences," Dartmouth College professor Mary Flanagan and colleagues conclude. It demands "a reflective understanding of the relevant values and how these values function in the lives of people."[36] *This* is the only ground from which we can attempt to maximize AI's potential for good and minimize its potential for harm. It is here that we need to double-click.

FOR MORE INFORMATION

Julian Adler: jadler@innovatingjustice.org

Jethro Antoine: jantoine@innovatingjustice.org

The authors wish to acknowledge Center for Justice Innovation Senior Media and Policy Advisor Matt Watkins for his editorial prowess, as well as Center for Justice Innovation Coordinator, New York Legal Policy Olivia Kramer for thorough research support, and Samiha Amin Meah for her impeccable document design and photo editing. Original photo by National Institute of Standards and Technology.

[1] Gilani, H. (2024, January 26). *Senators explore AI in criminal investigations and prosecutions*. Tech Policy Press. https://www.techpolicy.press/senators-explore-ai-in-criminal-investigations-and-prosecutions/

[2] Walsh, L. (2013). Interlude: Competing Ethical Models and a Catch-22. In *Scientists as Prophets: A Rhetorical Genealogy* (p. 86). essay, Oxford University Press. In an 1825 work, the French scholar Pierre-Simon de Laplace hypothesized an intellect for whom "nothing would be uncertain and the future just like the past could be present before its eyes."

[3] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. (2023). A Survey of Large Language Models. ArXiv, abs/2303.18223.

[4] FINRA. (2023, November). *Technology*. https://www.finra.org/about/technology (emphasis added)

[5] Center for AI Safety. (2023, May). *Statement on AI Safety*. https://www.safe.ai/work/statement-on-ai-risk#open-letter; See also: Roose, K. (2023, May 30). *A.I. Poses "Risk of Extinction," Industry Leaders Warn*. The New York Times. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

[6] O'Neil, L. (2023, August 16). *These Women Tried to Warn Us About AI*. Rolling Stone. https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/

[7] Bostrom, N. (2003). *Ethical Issues in Advanced Artificial Intelligence*. Nick Bostrom. https://nickbostrom.com/ethics/ai

[8] Woodside, T., Hendrycks, D., & Mazeika, M. (2023). *An Overview of Catastrophic AI Risks*. Center for AI Safety. https://www.safe.ai/ai-risk

[9] Rigano, C. (2019). Using Artificial Intelligence to Address Criminal Justice Needs. *National Institute of Justice Journal*, (280), 1–10. https:// www.nij.gov/journals/280/Pages/using-artificial- intelligence-to-address-criminal-justice-needs.aspx

[10] Phelps, M. S. (2017). Mass probation: Toward a more robust theory of state variation in punishment. *Punishment & Society*, 19(1), 53-73. https://doi.org/10.1177/1462474516649174

[11] Martin, E., & Moore, A. (2020, August 6). *Tapping into Artificial Intelligence: Advanced Technology to Prevent Crime and Support Reentry*. National Institute of Justice. https://nij.ojp.gov/topics/articles/tapping-artificial-intelligence#1-0

[12] Justice Innovation Lab. (2023, October 31). *Artificial Intelligence In Criminal Court Won't Be Precogs*. Medium. https://medium.com/@Lab4justice/artificial-intelligence-in-criminal-court-wont-be-precogs-11fe4a0dfc29

[13] Pinto, F. (2023, February 13). *Can AI Improve the Justice System?*. The Atlantic. https://www.theatlantic.com/ideas/archive/2023/02/ai-in-criminal-justice-system-courtroom-asylum/673002/

[14] Collings, P., & Guariglia, M. (2024, March 26). *Cops Running DNA-Manufactured Faces Through Face Recognition Is a Tornado of Bad Ideas*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2024/03/cops-running-dna-manufactured-faces-through-face-recognition-tornado-bad-ideas

[15] See, e.g., Minow, M., Zittrain, J., & Bowers, J. (2019, July 17). *Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns*. Berkman Klein Center for Internet and Society at Harvard University. https://cyber.harvard.edu/story/2019-07/technical-flaws-pretrial-risk-assessments-raise-grave-concerns Also: The Leadership Conference on Civil and Human Rights. (2019, February 1). *More than 100 Civil Rights, Digital Justice, and Community-Based Organizations Raise Concerns About Pretrial Risk Assessment*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[16] Dixon, H. B. (2021, January 15). *Artificial Intelligence: Benefits and Unknown Risks*. American Bar Association. https://www.americanbar.org/groups/judicial/publications/judges_journal/2021/winter/artificial-intelligence-benefits-and-unknown-risks/

[17] Smithsonian Institution. (2022). *AI Values Statement*. Smithsonian Data Science Lab. https://datascience.si.edu/ai-values-statement

[18] Lewis, M. (2013). *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton.

[19] Milgram, A. (2012, June 20). *Moneyballing Criminal Justice*. The Atlantic. https://www.theatlantic.com/national/archive/2012/06/moneyballing-criminal-justice/258703/

[20] Laura and John Arnold Foundation. (n.d.). *Public Safety Assessment: Risk Factors and Formula*. Laura and John Arnold Foundation. https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/PSA-Risk-Factors-and-Formula.pdf

[21] Alexander, M. (2018, November 9). *The Newest Jim Crow*. The New York Times. https://www.nytimes.com/2018/11/08/opinion/sunday/criminal-justice-reforms-race-technology.html

[22] Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[23] Advancing Pretrial Policy and Research. (2021, April). *Pretrial Assessment Tools*. Advancing Pretrial Policy and Research. https://shorturl.at/byTHW

[24] Picard, S., Picard, S., Watkins, M., Rempel, M., & Kerodal, A. G. (2019, July 1). *Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness*. Center for Justice Innovation. https://www.innovatingjustice.org/publications/beyond-algorithm

[25] See, e.g., Minow, M., Zittrain, J., & Bowers, J. (2019, July 17). *Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns*. Berkman Klein Center for Internet and Society at Harvard University. https://cyber.harvard.edu/story/2019-07/technical-flaws-pretrial-risk-assessments-raise-grave-concerns Also: The Leadership Conference on Civil and Human Rights. (2019, February 1). *More than 100 Civil Rights, Digital Justice, and Community-Based Organizations Raise Concerns About Pretrial Risk Assessment*. https://civilrights.org/2018/07/30/more-than-100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/

[26] Dholakia, N. (2024, January 30). *Electronic Monitoring Is an Extension of Mass Incarceration*. Vera Institute of Justice. https://www.vera.org/news/electronic-monitoring-is-an-extension-of-mass-incarceration

[27] Alexander, M. (2018, November 9). *The Newest Jim Crow*. The New York Times. https://www.nytimes.com/2018/11/08/opinion/sunday/criminal-justice-reforms-race-technology.html

[28] Nichols, Y., Idowu, A., Frankel, A., & Inglehart, M. (2022, September). *Rethinking Electronic Monitoring: A Harm Reduction Guide*. American Civil Liberties Union. https://www.aclu.org/wp-content/uploads/publications/2022-09-22-electronicmonitoring.pdf

[29] Barabas, C. (2020). Beyond Bias: Re-imagining the Terms of "Ethical AI" in Criminal Law. *Georgetown Journal of Law and Modern Critical Race Perspectives*, 12(2), 83–111. https://doi.org/10.2139/ssrn.3377921

[30] For a recent example of the dangers of opacity: Stelloh, T. (2024, May 3). *An AI tool used in thousands of criminal cases is facing legal challenges*. NBCNews.com. https://www.nbcnews.com/news/crime-courts/ai-tool-used-thousands-criminal-cases-facing-legal-challenges-rcna149607

[31] Tawari, T., Tawari, T., & Tawari, S. (2018). How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different? *International Journals of Advanced Research in Computer Science and Software Engineering*, 8(2), 1–9.

[32] Ryberg, J. (2024). Criminal Justice and Artificial Intelligence: How Should we Assess the Performance of Sentencing Algorithms? *Philosophy & Technology*, 37(1). https://doi.org/https://doi.org/10.1007/s13347-024-00694-3

[33] Garrett, B. L., & Rudin, C. (2023). The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice. *Duke Law School Public Law and Legal Theory Research Paper Series*. https://doi.org/http://dx.doi.org/10.2139/ssrn.4275661

[34] Kumar, R. S. S., Brien, D. O., Albert, K., Viljöen, S., & Snover, J. (2019, November 25). *Failure Modes in Machine Learning Systems*. arXiv.org. https://arxiv.org/abs/1911.11034

[35] For the axiomatic example (originally published in 1954): Heidegger, M. (1977). The Question Concerning Technology. In *The Question Concerning Technology and Other Essays* (pp. 3–35). essay, Harper & Row.

[36] Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In *Information Technology and Moral Philosophy* (pp. 322–353). essay, Cambridge University Press.

Center for Justice Innovation

520 Eighth Avenue
New York, NY 10018

p. 646.386.3100
f. 212.397.0985

innovatingjustice.org