A Line in the Sand
Artificial Intelligence and Human Liberty

It is hard to know where we stand in the timeline of AI implementation in the criminal legal space. Part of the challenge is that the criminal legal "system" is in reality a multiverse of federal, state, and local jurisdictions.[1] More problematic still is the sheer ubiquity of AI and related technologies. "I think the most important thing people don't know is that tech is now working at mega scale," observes Eric Schmidt, the former chairman and CEO of Google, cautioning—via the title of a recent Oscar-winning film—that tech is "everything everywhere all at once."[2]

What we *do* know is that AI is already in use in the criminal legal realm and, given the human propensity to reach for technological solutions to social problems, its further adoption is almost certainly unstoppable.[3] So how best to navigate the current moment of AI implementation? "We need a clear line in the sand: 'these use-cases are OK, these are not,'" urges Sara Friedman of The Council of State Governments Justice Center. "The criminal legal system deprives people of their liberty. It shouldn't be using AI to do this. There is a line when you are responsible for people's lives; there are things you shouldn't do."

> "The data may not be able to solve the problem—*we're talking about human nature.*"

Authors

Julian Adler, Chief Innovation and Strategy Officer, Center for Justice Innovation
Jethro Antoine, Chief Program Officer, Court Reform, Center for Justice Innovation
Kush R. Varshney, IBM Fellow, IBM Thomas J. Watson Research Center

Working Session Participants

Julian Adler
Center for Justice Innovation

Arthur Ago
Southern Poverty Law Center

Laith Al-Saadoon
Amazon Web Services

Jethro Antoine
Center for Justice Innovation

Roy Austin
Meta

Shiqueen Brown
Microsoft

Becky Duane
IBM

Sara Friedman
Council of State Governments

Lenore Lebron
Center for Justice Innovation

Shema Mbyirukira
Verizon

Friedman was a participant in a working session featuring leaders from across the technology and criminal legal spaces held at the Center for Justice Innovation on December 5, 2024. The meeting aimed to think practicably about the recommendations in our earlier paper, Minding the Machines: On Values and AI in the Criminal Legal Space, which makes a sustained case for prioritizing values—what do we want to use AI *for*—over any rush to implement the technology simply because it's ready-to-hand.[4]

Echoing Friedman, **we strongly encourage a moratorium on any use of AI in the criminal legal system that would affect people's liberty interests or pose a substantial risk of harm**. An intentional pause would allow for a thorough assessment of AI's impact on liberty and safety, and for a proper consideration of whether it should be deployed at all in certain higher-stakes contexts.[5]

If anything is clear about AI, it is that the people closest to its development are the loudest voices warning of the "catastrophic" risks posed by the technology.[6] And given the vulnerability of incarcerated and correctionally supervised people, there is added reason for caution in introducing any new technology at scale (or even a more circumscribed pilot project)—in particular a technology so powerful and, for now, potentially ungovernable.

Yet we realize even a temporary injunction could prove a difficult sell—that people are far too keen to "experiment," "play," or "sprint" rather than constrain the use of AI on the ground. "When you see something that is technically sweet," warned Robert Oppenheimer in 1954, "you go ahead and do it and you argue about what to do about it

only after you have had your technical success."[7] What follows, then, are three ideas for safeguarding AI implementation in the here-and-now that emerged over the course of the Center for Justice Innovation's recent working session.

Lower-Risk Environments and Decision Points

Even as the AI sector is awash in funding, leading developers are struggling to prevent "hallucinations," "scheming," and other incidents in which AI essentially goes rogue[8]—a scenario that should be unimaginable when someone's liberty or safety is at stake. A recent study by Anthropic and Redwood Research uncovered a phenomenon it dubbed "alignment faking," whereby an AI system "feigns compliance" with instructions and guardrails while pursuing its own agenda in the background.[9] As AI becomes increasingly self-generating, OpenAI technologists warn that "humans won't be able to reliably supervise AI systems."[10] There is also the issue of human users of AI finding ways to bypass any attempted safeguards.[11]

Actors looking to implement AI in criminal legal settings should aim for the lowest-risk environments and/or the lowest-risk decision points. This means avoiding the use of AI inside settings such as precinct lockups, jails, and prisons, as well as other spaces where people are immensely vulnerable. If total avoidance inside facilities is not an option, AI should be limited to functions that do not risk appreciable harm, for example, using AI to expedite an intake process or to enhance pre-release reentry planning.

Working Session Participants

(continued)

Meghna Padmanabhan
Council of State Governments

Sarah Picard
MDRC

Judge Victoria Pratt
Center for Justice Innovation Board of Directors

Jesse Rothman
Council on Criminal Justice

Alison Shames
Center for Effective Public Policy

Karen Tan
Vera Institute of Justice

Kush Varney
IBM

Lisa Vavonese
Center for Justice Innovation

Jonathan Wroblewski
Harvard Law School

Even within the limitations we're advocating, there remains much AI could be doing inside a mechanism as vast as the criminal legal system. To focus on the pretrial space alone: yes, our moratorium would rule out the use of AI to make decisions about detention versus release, but the technology could significantly enhance the capacity of case managers to support people in accessing community-based resources, services, and housing. Natural language processing could help summarize case notes, court documents, and client histories, while predictive analytics could suggest clients in need of more urgent intervention. Machine learning algorithms could analyze patterns across cases to suggest services tailored to individual client needs, while maintaining human oversight and decision-making in sensitive matters.

Employed judiciously, these tools could help reduce administrative burden while improving the quality and consistency of client support, particularly for case managers overseeing high-risk cases or larger caseloads. Slightly further afield, "upstream" community-based programs—efforts to *prevent* contact with the criminal legal system altogether—might be ideal for this kind of robust AI development and implementation.[12]

AI is also being used to expedite data collection and analysis in service of policy advocacy. The technology can be particularly effective when the necessary data is difficult to access. The recent use of AI technology to disrupt the default use of incarceration for unpaid court fines and fees in Jefferson County, Alabama, is a prime example.[13]

Simulate First

Before deploying AI in justice-related settings, comprehensive evaluation should be mandatory, even when system actors are not using it to make liberty-impacting decisions. Think of evaluation as a simulation or trial run in a controlled environment—a way to safely test AI behavior before it can have real-world consequences. Evaluations can range from basic to sophisticated, with the minimum standard being model validation—put simply, auditioning the AI—on representative real-world datasets. For justice-related large language models—that is, AI capable of understanding language and generating original content—this would likely begin with curating question-answer pairs from case management interactions to create "golden datasets" that represent ideal responses to client needs.

When direct real-world data isn't available or usable for evaluation purposes, stakeholders can make use of aggregate statistics or generate so-called synthetic data. The latter involves creating artificial but statistically representative data to stress-test the AI's behavior and assess its alignment with intended outcomes. The golden datasets can then serve as continuous monitoring benchmarks, allowing evaluators to detect when model outputs drift from established standards of fairness and accuracy.

Perhaps most revealing, simulations could expose pre-existing human biases and systemic failures in the legal system. When AI models struggle with certain demographics or contexts, it's frequently because they're learning from historical data that reflect human-generated disparities (a

phenomenon witnessed in the introduction of an earlier technology: pretrial algorithmic risk assessments[14]). This insight suggests that the methodologies we use to detect and mitigate AI bias may help to identify and address longstanding human biases in criminal legal decision-making. In a similar spirit of self-examination, the risk factors associated with incarceration—everything from reoffending to death—are well-established; why not deploy an AI courtroom tool to gauge the risk incarceration could pose to a given individual?[15]

Push for Standards

At numerous points throughout the working session, the discussion focused on the creation of field-wide standards to safeguard AI implementation in the criminal legal space. There are a range of analogous efforts to draw on. Several participants pointed to the work of the U.S. Department of Commerce's National Institute of Standards and Technology [NIST], which "leads and participates in the development of technical standards, including international standards, that promote innovation and public trust in systems that use AI."[16] While NIST is in the business of promulgating "voluntary consensus-based standards," working session participants envisaged a policy process or body with more enforcement power or more informal influence in the criminal legal ecosystem.

But even with standards for the use of criminal legal AI secured, such a policy would not be self-executing.

Policy implementation has been a perennial challenge in the criminal legal space.[17] This has been most acute in the attempt to translate empirical data into policy. For policymaking backed by convincing research to fulfill its promise, Elizabeth Linos of the Harvard Kennedy School has identified three criteria: "evidence has to be *useful*—i.e., it answers questions that a government has asked; *usable*—i.e., the right insights can be translated into a new context; and *used*—i.e., organizations have the capacity to implement the practice at scale."[18] These criteria apply to any attempted standards for the responsible implementation of AI, especially efforts to put the brakes on hasty experimentation and safeguard against unknown and unintended consequences.

Finally, given the speed of AI's evolution, it may be unrealistic to expect standards, formal laws, and other forms of regulation to keep pace.[19]

> "There is a line when you are responsible for people's lives; there are things you shouldn't do."

Conclusion: Human, All Too Human

"The data may not be able to solve the problem here," reflected a technology executive toward the end of the working session, "*we're talking about human nature*." AI is "inherently socio-technical in nature," concurs a recent report from NIST. Its outcomes emerge (messily) from the "interplay" of the technology with a host of social factors—including

the people operating it and the context in which it is operating.[20]

As our technological capabilities expand, Massachusetts Institute of Technology's Michael Schrage and David Kiron remind us to focus on the formative role our *human*-generated ideas play in AI's evolution—from our ethical commitments to our conception of "how AI represents reality." They challenge us to "possess the self-awareness and rigor" to avoid "default[ing] to tacit, unarticulated philosophical principles" for our AI deployments.[21]

As the philosopher Sidney Hook concluded 65 years ago, reckoning with the interplay of human nature and "the discoveries of scientific technology" of his own era, "it is *human* choice and the decisions to which it leads" that will determine our fate.[22] In the high-stakes context of the criminal legal system, the decisions pertaining to the implementation of AI can start with humility and hypervigilance.

FOR MORE INFORMATION

Julian Adler

Email: jadler@innovatingjustice.org

Acknowledgements

First and foremost, the authors wish to thank Matt Watkins, the Center for Justice Innovation's Senior Media and Policy Advisor and the best editor in the business, and Laith Al-Saadoon, Principal Prototyping Architect at Amazon Web Services and an incredible technical advisor and thought partner on this project. Deep appreciation to the Center for Justice Innovation's Communications Team for its exceptional work on design and dissemination, and to Rachel Krul for her insightful research support.

[1] Julian Adler and Greg Berman, *Start Here: A Roadmap to Reducing Mass Incarceration*, p. 22. (2018).

[2] Jim Allen & Mike VandeHei, *Behind the Curtain: The Great Upheaval*, Axios, Dec. 12, 2024, https://www.axios.com/2024/12/12/trump-elon-musk-great-upheaval-ai-politics.

[3] Justice Innovation Lab, *Artificial Intelligence In Criminal Court Won't Be Precogs*, Medium, Oct. 31, 2023, https://medium.com/@Lab4justice/artificial-intelligence-in-criminal-court-wont-be-precogs-11fe4a0dfc29; Beryl Lipton, *AI in Criminal Justice Is the Trend Attorneys Need to Know About*, Electric Frontier Foundation, Nov. 5, 2024, https://www.eff.org/deeplinks/2024/11/ai-criminal-justice-trend-attorneys-need-know-about.

[4] Julian Adler, Jethro Antoine, & Laith Al-Saadoon, *Minding the Machines: AI and the Criminal Legal Space*, Center for Justice Innovation, June 2024, https://www.innovatingjustice.org/publications/minding-machines.

[5] Compare the 2023 letter, with signatories including experts in AI and leading tech figures, demanding a pause "for at least six months" in training any AI system more powerful than GPT-4. As the letter concluded, with special relevance for the criminal legal space: "Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable. This confidence must be well justified and *increase with the magnitude of a system's potential effects*" [emphasis added]. Future of Life Institute, *Pause Giant AI Experiments: An Open Letter*, Mar. 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[6] Hearing on Oversight of A.I.: Principles for Regulation before S. Committee on Privacy, Technology, and the Law U.S. Senate 2-3 (2023), https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

[7] In the Matter of J. Robert Oppenheimer, Transcript of Hearing Before United States Atomic Energy Commision Personnel Security Board (1954), https://www.osti.gov/includes/opennet/includes/Oppenheimer%20hearings/unitedstatesatom007206mbp.pdf.

[8] Alexander Meinke, et al., *Frontier Models are Capable of In-context Scheming*, Apollo Research, Dec. 5, 2024, https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/6751eb240ed3821a0161b45b/1733421863119/in_context_scheming_reasoning_paper.pdf.

[9] Ryan Greenblatt, et al., *Alignment faking in large language models*, Arxiv, Dec. 20, 2024, https://arxiv.org/abs/2412.14093. See also: Matthias Bastian, *AI models can only pretend to follow human rules, Anthropic study finds*, The Decoder, Dec. 21, 2024, https://the-decoder.com/ai-models-can-only-pretend-to-follow-human-rules-anthropic-study-finds/.

[10] Kyle Wiggers, *OpenAI is forming a new team to bring 'superintelligent' AI under control*, TechCrunch, Jul. 5, 2023, https://techcrunch.com/2023/07/05/openai-is-forming-a-new-team-to-bring-superintelligent-ai-under-control/.

[11] Sam Biddle, *The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques*, The Intercept, Dec. 8, 2022, https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/.

[12] Julian Adler & Chidinma Ume, *Toward Community Justice: Upstream Investment Is Criminal Legal Reform*, Center for Justice Innovation, Apr., 2024, https://www.innovatingjustice.org/publications/upstream-is-reform.

[13] Sarah Picard et al., *The Jefferson County Equitable Fines and Fees Project: Preliminary Findings on Fairness and Efficacy*, MDRC, Aug., 2024, https://www.mdrc.org/sites/default/files/JeffCoFines%26Fees_Brief_FInal.pdf.

[14] Sarah Picard, Matt Watkins, Michael Rempel, & Ashmini G. Kerodal, *Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness*, Center for Justice Innovation, Jul. 2019, https://www.innovatingjustice.org/publications/beyond-algorithm.

[15] Dasha Pruss et al., *Prediction and Punishment: Critical Report on Carceral AI*, Jan. 10, 2025, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5017321.

[16] National Institute of Standards and Technology, *Artificial Intelligence*, United States Department of Commerce, https://www.nist.gov/artificial-intelligence.

[17] Center for Justice Innovation, *Sticking the Landing: The High Stakes of Policy Implementation*, Mar. 2024, https://www.innovatingjustice.org/publications/policy-implementation.

[18] Personal email correspondence, Elizabeth Linos and Julian Adler, December 17, 2024.

[19] Olga Mack, *Building a Responsible Practice Framework: Navigating the Intersection of Laws, Ethics, and AI*, MIT Computational Law Report, Sept. 7, 2023, https://law.mit.edu/pub/building-a-responsible-practice-framework-navigating-the-intersection-of-laws-ethics-and-ai/.

[20] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, United States Department of Commerce, Jan., 2023, https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

[21] Michael Schrage & David Kiron, *Philosophy Eats AI*, MIT Sloan Management Review, Jan. 16, 2025, https://sloanreview.mit.edu/article/philosophy-eats-ai/.

[22] *Psychoanalysis Scientific Method and Philosophy: A Symposium*, New York University Press (Sidney Hook ed., 2 ed, 1964)

Center for Justice Innovation

520 Eighth Avenue
New York, NY 10018

p. 646.386.3100
f. 212.397.0985

innovatingjustice.org