AI and Criminal Procedure Rights:

A Response to the National Institute of Justice Request for Input

Brandon L. Garrett,[*] Alicia Carriquiry,[†] Karen Kafadar,[‡] Robin Mejia,[§] Cynthia Rudin,[**] Nicholas Scurich,[††] Hal Stern[‡‡]

May 28, 2024

Introduction

Today, as artificial intelligence (AI) has been implemented across a wide range of human activities, new warnings have been issued from a wide range of sources, academic, public policy, and government, regarding the dangers posed by artificial intelligence to society, democracy, and individual rights. In 2023, the White House issued an "AI Bill of Rights," and next, an executive order on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" asking all federal agencies to account for how they use AI systems.[1] That latter order tasks the Attorney General with submitting to the President a report regarding "the Federal Government's fundamental obligation to ensure fair and impartial justice for all, with respect to the use of AI in the criminal justice system." [2]

The National Institute of Justice (NIJ) seeks written input from the public relevant to section 7.1(b) of Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[3] We write to express our own views as scholars who study law, statistics, and constitutional criminal procedure, in response to the NIJ request for input. We note that several of us have long been affiliated with the Center for Statistics and Applications in Forensic Evidence (CSAFE), with the mission to bring improved statistical methods into the use of forensic evidence in criminal cases, to improve the quality of justice. We write to reflect our

[*] L. Neil Williams, Jr. Professor of Law, Duke University School of Law and Faculty Director, Wilson Center for Science and Justice.
[†] Distinguished Professor and President's Chair Director of the Center for Statistics and Applications in Forensic Evidence.
[‡] Commonwealth Professor, Department of Statistics, University of Virginia.
[§] Director of the Statistics and Human Rights Program and Special Faculty, Center for Human Rights Science, Carnegie Mellon University.
[**] Earl D. McLean, Jr. Professor of Computer Science, Electrical and Computer Engineering, Statistical Science, Mathematics, and Biostatistics & Bioinformatics, Duke University.
[††] Chair and Professor of Psychological Science, University of California, Irvine.
[‡‡] Provost and Executive Vice Chancellor, Chancellor's Professor, Department of Statistics, University of California, Irvine.
[1] *See* Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023, at https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[2] *See id*. at Section 7(B).
[3] Federal Registry, Request for Input From the Public on Section 7.1(b) of Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence", at https://www.federalregister.gov/documents/2024/04/25/2024-08818/request-for-input-from-the-public-on-section-71b-of-executive-order-14110-safe-secure-and.

own views, and not those of our respective institutions, as researchers in law, scientific evidence, statistics, artificial intelligence, machine learning, and computer science.

We write to emphasize two basic points, focusing on predictive uses of AI, not on large language models.

First, in high-risk settings like the criminal justice system, AI models and underlying data must be adequately tested, including independently. Relatedly, sound statistics requires disclosing defined sources of data and information regarding variability and errors in measurement.

Second, artificial intelligence (AI) must not be black box in high-risk settings such as criminal investigations, in which it affects criminal procedure rights. A mature body of computer science research shows that nothing will be lost in performance by requiring such transparency through regulation. In short, AI must be transparent, tested, and interpretable.

To accomplish both goals, far more can and should be done to apply and robustly protect the existing Bill of Rights in the U.S. Constitution as it should apply to uses by government of AI in the criminal system, particularly when AI is used to provide evidence in investigations and trials.

Further, we also highlight the need to promptly comply with the Office of Management and Budget federal procurement guidelines, released in March 2024, to implement the 2023 Executive Order. Together, these measures provide for a range of federal agency reviews and oversight, inventories of AI systems, managing risks, and most importantly—auditing of AI systems, by testing how they perform in its "intended real-world context."[4] Requiring AI vetting, review, and disclosure provides a sound model. Federal agencies must implement minimum practices for risk management of safety and rights-impacting AI by December 1, 2024.[5] Importantly, these regulations set out the fundamental need for independent testing of AI systems:

> Through test results, agencies should demonstrate that the AI will achieve its expected benefits and that associated risks will be sufficiently mitigated, or else the agency should not use the AI. [6]

Our view is that these procurement rules should be carefully followed in all contexts where due process and other rights are affected by the use of AI systems by law enforcement in criminal investigations and cases. These rules address our first point, regarding testing and transparency. They do not address our second point, regarding interpretability of AI systems in high risk settings.

In criminal cases in which liberty is at stake, there should be a strong presumption that fully interpretable AI be used, when it is directed towards providing evidence against criminal defendants. The burden to justify "black box" uses of AI in court should be a high one, given our

[4] Executive Office of the President, Office of Management and Budget, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, 18-19 (March 28, 2024), at https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf.
[5] *Id*. at 14.
[6] *Id*. at 18.

commitments to due process, public judicial proceedings, and defense rights of access. There is no evidence that performance and efficiency depend on keeping the operation of AI secret from the public and unintelligible to users. That fundamental point, that AI can and should be open, for inspection, vetting, and explanation, is a simple one and it can be more forcefully insisted on at the federal level.

Finally, we do not disagree that existing rights need to be at times reinterpreted for the AI era. However, most important is a strong commitment to enforce existing constitutional criminal procedure rights, particularly given how difficult it is to amend the U.S. Constitution, but also the unfortunate reality that those constitutional rights have been unevenly enforced in criminal cases, where largely indigent defendants face challenges in obtaining adequate discovery and the pressures to plead guilty and waive trial rights.[7]

The National Institute of Justice, the federal law enforcement agencies, and the Department of Justice, should lead by example, in the use of AI technologies to vigorously adhere to statistical and scientific standards, and to protect constitutional rights of criminal defendants.

I. What is AI?

"Artificial intelligence" simply means that machines perform tasks that are typically performed by humans.

Machine learning is a subfield of AI, and it heavily overlaps with predictive statistics. We can view machine learning as a kind of pattern-mining, where algorithms are looking for patterns in data that can be useful. The data is supplied to the machine, which relies on past patterns to develop methods for making recommendations for what to do next. For instance, when predicting whether someone might have a drug overdose, patterns in their medical record and twitter feed, as well as those of others, might help us predict that outcome. These patterns can help human decision makers because no human can calculate patterns from large databases in their heads. Individual people may in fact be biased or place undue weight on information that is not particularly predictive. If we want humans to make better data-driven decisions, machine learning can help with that.

An "interpretable" AI model allows people to interpret what the AI system's formula actually relied upon when it made a prediction. The underlying models, or algorithms, used by the AI may be extremely complex. The factors that the model relies upon, however, may be simple. There are many examples of simple, fully interpretable AI models used in criminal justice settings.

A less desirable approach, "explainable AI," provides an account of what the algorithm may have done, by sharing general explanations for what it might have done. These post-hoc explanations may not be correct, especially in hard cases.[8]

---

[7] For a detailed discussion of these issues, *see* Brandon L. Garrett & Cynthia Rudin, *The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice*, 113 Cornell L. Rev. 561 (2024).
[8] *See* Neil Savage, *Breaking into the Black Box of Artificial Intelligence*, Nature, Mar. 29, 2022.

In the context of AI, "transparency," as we define it, refers to sharing adequate information in order to evaluate the accuracy of the predictive model. Simply put, transparency permits and requires testing AI systems.

Transparency for an AI system requires sharing several types of information. First, it requires sharing the underlying formula for the model. This can permit an independent researcher to conduct an evaluation and assess the accuracy of the model.

It will also be necessary to share test set and training data to replicate the evaluations done in the past on a model. For many models, the performance of AI is sensitive to test data used to develop the model. If the data used to develop the model is unrepresentative or biased, which is not unusual, then the model's predictions will similarly be biased.[9]

One final definitional distinction may be helpful. An "open" AI system, or at least an "open source" system, makes the source code available for free, and includes a license permitting derivative use of the software. An open-source AI system by definition makes its source code available; however, it may still not be fully transparent if training and test data is not shared to permit testing of the model. We take no position here on when and whether it is appropriate or of public benefit for AI systems to be open and shared publicly. There may be social and national security concerns with sharing code that could be adopted for harmful uses and concerns with sharing data regarding criminal justice outcomes. However, transparency, and testing by independent researchers can and should occur even if AI is not made "open."

II. Black Box Models Are *Not* More Accurate Than Interpretable "Glass Box" Models

Having set out these definitions, a black box predictive model not interpretable. It uses a formula that is too complicated for any human to understand or it is deemed by the designer to be proprietary, which means no one can understand its inner workings, because those inner workings are not shared or are not designed to be share-able. These models can cause problems for high stakes decisions like criminal risk scoring, where someone could get denied parole and they and their defense lawyer, the parole officers, and the public, are not able to figure out why the person did or did not get a high-risk score.

There is a common misconception that black box AI must be more accurate than any model a human could understand. That is just not true.[10] Models that are fully interpretable to humans can perform just as well as models that are not. This has been shown to be true across fields, including computer vision.[11] And recidivism risk scoring.[12] The ways in which AI affects rights

[9] Adam M. Chekroud et al., *Illusory Generalizability of Clinical Prediction Models*, 383 Science 164-167 (2024).
[10] Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead*, Nature Machine Intelligence, 2019.
[11] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Barnett, Jonathan Su, Cynthia Rudin, *This Looks Like That: Deep Learning for Interpretable Image Recognition*, NeurIPS, 2019.
[12] Jiaming Zeng, Berk Ustun, and Cynthia Rudin, *Interpretable Classification Models for Recidivism Prediction*, Journal of the Royal Statistical Society, 2017; Caroline Wang, Bin Han, Bhrij Patel, Feroze Mohideen, and Cynthia Rudin, *In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction*, Journal of Quantitative Criminology, 2022, at 10.1007/s10940-022-09545-w.

and interests need not be hidden or secret. AI need not be a black box to attain the accuracy of a black box.

There are strong reasons to prefer interpretable AI in criminal justice settings. Black Box AI tends to lead to less accurate decision-making, because such models are harder to troubleshoot and use in practice. Typographical errors in the input to black box recidivism prediction models has led to catastrophic errors in decision making, deeply affecting people's lives.[13] This type of poor decision-making is a direct result of unnecessary secrecy, weighted in favor of companies that sell black box models to the justice system, rather than weighted towards those individuals in the justice system subjected to the decisions made from these models.

III. Transparency, Testing, and Statistics for AI

Turning from interpretability to transparency and testing, we not only need to know what the AI system based its predictions on in a particular case, but we need assurance that the AI system has been tested and is sufficiently reliable to be used in a high-risk setting like the criminal justice system. Testing requires transparency, or some degree of adequate disclosure of code and data to independent researchers. Further, once that testing is done, the statistical limitations of the AI system's predictions must be disclosed.

1. Transparency to Permit Testing of AI

In an important National Academy of Sciences consensus report released in 2019 on the need for reproducibility in the sciences, for purposes of computation, reproducibility is defined as: "obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis."[14]

As the NAS definition makes clear, testing for reproducibility regarding AI systems requires the ability to access a system, in order to examine whether it produces consistent results, and whether another researcher obtains the same basic results. Thus, the NAS highlights that "[r]eproducibility is strongly associated with transparency; a study's data and code have to be available in order for others to reproduce and confirm results."[15] In addition, it may be necessary to provide detailed information about the computing environment in which the system was executed.[16]

While reproducibility is a key component of the scientific method and of assuring the reliability of any results or system, as we will describe, in many criminal justice contexts, it has not been assessed. We do not know whether the AI model works as claimed, based on an independent person testing its operation.

---

[13] Cynthia Rudin, Caroline Wang and Beau Coker, *The Age of Secrecy and Unfairness in Recidivism Prediction*, Harvard Data Science Review, 2020.
[14] *See* National Academies of Sciences, *Consensus Study Report: Reproducibility and Replicability in Science* 3 (National Academies Press, 2019).
[15] *Id*. at 2.
[16] *Id*. at 7-8.

In addition to reproducibility, in which prior results are confirmed, we need to know whether an AI system is repeatable, or whether others would reach the same results using that system. The testing of repeatability is particularly important where the operation of many AI systems is not designed to be static, but rather to change over time, as the system adapts its predictive model to new data.

Third, replicability, or the ability to obtain the same results in different circumstances, is also crucial to the use of AI systems. The National Science Foundation defines replicability as follows: "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."[17] Or, consistent with that definition, as the NAS Report put it, "Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data."[18]

To take one high-profile examine, the Federal Bureau of Investigation, and many state and local law enforcement agencies use facial recognition technology (FRT) systems to identify potential suspects from photos and video. As the National Academy of Sciences developed in an important 2024 report, what we do know suggests that "accuracy varies widely across the industry," and recommending industry-wide standards for testing and evaluation of the performance of such systems. [19]

If AI is being used, its users need to know how reliable it is. Reliability may be present in general, but it will often also be important how reliable a system is when applied to more specific populations or tasks. We discuss bias more below, but the NAS highlighted how FRT systems "yield consistently higher false positive match rates when applied to racial minorities, including among populations that are Black, Native American, Asian-American, and Pacific Islanders."[20] The reliability of an AI system may not be self-apparent and nor may its differential accuracy regarding particular populations or uses. And unfortunately, many developers do not share information about how reliable the AI system is, or how well it has been validated with the government, much less with the affected public.

2. Valid Statistical Statements and AI

In an important statement in 2019, the American Statistical Association set out basic principles for reporting the results of any forensic analysis. In summarizing sound statistical statements in general, the ASA set out:

> Statistical statements should rely on: (1) a defined relevant database describing characteristics, images, observed data, or experimental results; (2) a statistical model that describes the process that gives rise to the data; and (3) information on variability and errors in measurements or in statistics or inferences derived from measurements. This

[17] Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science* (2015).
[18] NAS Report, *supra*, at 6.
[19] National Academy of Sciences, Committee on Facial Recognition, *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance* 52, 110 (National Academies Press, 2024).
[20] *Id*.

> information permits a valid statistical statement regarding the probative value of comparisons or computations (e.g., how rare is an observed positive association when two items arise from the same source and when they arise from different sources?).[21]

Second, for reporting on forensic evidence in particular, the ASA explained:

> The ASA recommends that trace, impression, or pattern evidence practitioners follow a valid and reliable process to determine the extent to which evidence supports the hypothesis of an association between a questioned sample and a sample whose source is known (such as a control sample from a person of interest). Reliability and validity should be established via scientific studies that have been subjected to independent scientific scrutiny.

This statement was not specifically geared towards AI evidence. However, the same principles apply. The data used to develop an AI system must be defined and relevant. As the National Academy of Sciences Committee on Human-AI Teaming noted, "it is well established that bias can be inadvertently introduced into an AI system due to underlying data sample selection bias."[22] And, predictive statements made using AI with those data must present information on variability and errors. In short, in high-risk settings like criminal justice, AI must be fully interpretable. Just like a human examiner that does not explain any statistical basis for their conclusions is a "black box," black box AI is not susceptible to any statistical validation and therefore makes no valid statistical statements. Such statements lacking in statistical validity have no place in criminal investigations or cases.

3. The 2024 OMB Memorandum

The 2024 OMB memorandum address the transparency of data, testing, and statistics supporting AI models. For example, agencies were instructed that: "Any data used to help develop, test, or maintain AI applications, regardless of source, should be assessed for quality, representativeness, and bias."[23] The memorandum calls for adequate infrastructure for developers to be able to adequately "develop, test, and maintain AI applications."[24] Further, the memorandum instructs that agencies shall share code, including models and model weights, on a public repository, absent a waiver.

The OMB memo requires that before using safety or rights-impacting AI, by December 1, 2024, an agency must conduct an AI impact assessment. That impact assessment requires examining and documenting the intended purpose and benefits of AI, including with quantifiable measures of the desired outcomes, and then the potential risks of using AI to achieve those outcomes.[25] Second, the agencies must examine the quality and appropriateness of the relevant data used to design, develop, train and test the AI system, given its intended goals. Even if the

[21] American Statistical Association Position on Statistical Statements for Forensic Evidence, Presented under the Guidance of the ASA Forensic Science Advisory Committee, January 2, 2019, at https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf.
[22] NAS Report, Human-AI Training, supra note xxx, at 58.
[23] Id. at 11-12.
[24] Id. at 11.
[25] Id. at 16-17.

data is held by an outside vendor, the agency must make efforts to secure "sufficient descriptive information" about it.[26]

Next, the agencies must conduct AI testing, including an independent evaluation, using testing designed to ensure that the system performs in "its intended real-world context." [27] The agency itself needs to conduct these tests, and not just rely on the developer's own testing. As with the underlying data, it is not enough to rely on work of a vendor.[28] This requirement of an independent evaluation, with documentation of real-world testing, following examining of the data, models, code, and outputs, provides an important model for requiring testing of AI in order to be able to make statistically valid statements about any predictions an AI model makes.

IV. What Constitutional Criminal Procedure Rights are Implicated?

We now have far greater appreciation for the fact that AI can affect people's lives in important ways. These include a wide and growing range of applications in our criminal system. AI is already being used in a host of criminal investigation, pretrial, and sentencing-related settings. For example, algorithms are used for risk scoring, in order to predict the risk that someone will commit a crime if they are released on bail, or given parole. Many states mandate that risk scores be used in various decisions, always to inform a judicial or other officials' discretion, to be sure (and there are real questions concerning variability with which judges and others incorporate quantitative information into their decision-making). Another high-profile example is the use of facial recognition technology as a forensic tool and for surveillance.

We emphasize that the particular use of AI can greatly alter the accuracy, privacy, and fairness interests at stake, as well as the fair trial rights involved. Thus, using AI to search for a missing person feared to have been kidnapped raises far fewer questions than using AI to identify a culprit from a surveillance video. Any use of AI that results in evidence introduced during a criminal investigation, or in court, will raise serious constitutional concerns. Our focus here is on uses of AI that can generate evidence, whether introduced in court or used in an investigation.

A range of constitutional rights apply to protect individuals against deprivations of important interest through government action, and a range of rights are focused on the rights of individuals during criminal investigations and criminal adjudication. The most expansive constitutional provision implicated by uses of AI in criminal investigations strikes at the fundamentals of government action: the Due Process Clauses of the Fifth and Fourteenth Amendments.[29] The federal government can ensure that no federal agency uses AI in a way that arbitrarily deprives persons of the rights during criminal investigations and adjudication. Simply put, we should look more closely at uses of AI that might result in evidence used to arrest a person and that might result in evidence used in court during a criminal case. We do not focus here on Fourth Amendment privacy rights relating to searches and seizures, although similar principles should apply. We focus here on due process, Sixth Amendment Confrontation, equal protection,

[26] Id. at 16-17.
[27] Id. at 18.
[28] Id. at 18.
[29] Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1, 10-16 (2014).

and related rights: on uses of AI to generate evidence that could be used in court in a criminal case, to determine bail, to convict a person, or to impose a sentence.

The due process protections in criminal cases include assurances that all material and exculpatory evidence of innocence be disclosed to criminal defendants.[30] Defendants have a right to effective assistance of counsel; defense counsel, in our view, cannot meaningfully defend a person without information about what AI evidence is bring introduced in a case.[31] The Equal Protection Clause protects against purposeful discrimination of protected groups, including based on race. The federal government must insist that AI be carefully vetted to assure against discriminatory impacts on minority groups. Further authority under civil rights legislation can assure that federal grant recipients do the same.

Further, the defense cannot meaningfully defend a person without knowing whether the AI formula was calculated without error; in the case of risk scoring, there has been much evidence of typographical errors or other types of data errors influencing the scores.[32] In some cases, it has been reported that the wrong score is being computed for all defendants. In the case of COMPAS in Broward County, FL, it was reported that the wrong scoring model had been used for years. The COMPAS parole score was used to determine pretrial risk, rather than the COMPAS pretrial score that was designed for this purpose.[33,34]

We emphasize the importance of affirmatively adopting policies to ensure that constitutional rights are protected, because in practice, many are not asserted or enforced. Discovery in criminal cases is typically quite limited, making it difficult for defendants to be aware that there is even an issue that exculpatory evidence may not have been disclosed. A criminal defendant may not be aware that AI was used to generate leads or evidence. Nor are evidentiary rights clearly defined in pretrial settings, or in sentencing proceedings in many jurisdictions.

In general, expert evidence admissibility decisions have also been quite deferential in criminal cases; the National Academy of Sciences itself has explained that scientific safeguards must be put into place by government, given the limited ability of defendants to challenge even wholly unscientific expert evidence in criminal cases.[35] That report highlighted how courts have routinely found admissible a range of forensic evidence of reliability that simply has not been established, where: "With the exception of nuclear DNA analysis, however, no forensic method

---

[30] *Brady v. Maryland,* 373 U.S. 83 (1963). Regarding questions whether machine-generated results are themselves "testimonial" under the Sixth Amendment Confrontation Clause, *see* Andrea Roth, *Machine Testimony*, 126 Yale L.J. 1972, 2039 (2017).

[31] Strickland v. Washington, 466 U.S. 668 (1984).

[32] Cynthia Rudin, Caroline Wang and Beau Coker*, The Age of Secrecy and Unfairness in Recidivism Prediction*, Harvard Data Science Review, 2020.

[33] How We Analyzed the COMPAS Recidivism Algorithm, Propublica, Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, May 23, 2016

[34] Jackson, E., & Mendoza, C. (2020). Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.1b3dadaa

[35] *See* Comm. on Identifying the Needs of the Forensic Sci. Cmty. & Nat'l Res. Council, *Strengthening Forensic Science in the United States: A Path Forward* 87 (2009) [hereinafter NAS Report]; Peter J. Neufeld, *The (Near) Irrelevance of Daubert to Criminal Justice: And Some Suggestions for Reform*, 95 AM. J. PUB. HEALTH S107, S110 (2005).

has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source."[36]

There remains a critical need to validate a range of traditional forensic methods.[37] CSAFE was founded to address this important need, and much work remains to be done to conduct those validations. Such validation is of course equally important for new methods, such as AI systems, used in criminal cases. Further, a criminal defendant, if indigent, may often be denied funds to retain an expert to examine AI technology used by a prosecution expert.[38] The defendant may have no way to independently verify the work done, using AI, by government investigators.

Federal Rule 702 was recently amended in December 2023, including to underscore the responsibility of federal judges to serve as gatekeepers regarding expert evidence used in criminal cases. That rule change makes it all the more important that federal agencies deploy only interpretable AI systems for criminal investigation and adjudication purposes, and to disclose information needed to evaluate the reliability of that AI system. The reliability of an expert cannot be evaluated if the AI system is not fully interpretable and if information about it is not disclosed.

Courts are beginning to address such issues, including discovery in criminal cases. In an important ruling in *State v. Arteaga*, a New Jersey Appellate Court affirmed a trial court order, ruling that if the prosecutor planned to use FRT, or the eyewitness who selected the defendant in a photo array, then they must provide the defense with information concerning "the identity, design, specifications, and operation of the program or programs used for analysis, and the database or databases used for comparison," as all "are relevant to FRT's reliability."[39] The court concluded, the "[d]efendant must have the tools to impeach the State's case and sow reasonable doubt."[40]

Further, regarding discovery, Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack argue that a judge cannot possible address any of the reliability questions raised by the use of AI as evidence, "unless the party offering the AI evidence is prepared to disclose underlying information concerning, for example, the training data and the development and operation of the AI system sufficient to allow the opposing party (and the judge) to evaluate it, and the party against whom the AI evidence will be offered to decide whether and how to challenge it."[41]

To be sure, in the past, judges have tended to narrowly view defense requests for discovery regarding evidentiary uses of AI, as well as forensic evidence more broadly, in criminal cases. They have tended to more expansively view discovery requests only when errors have come to

---

[36] NAS Report, supra, at 7.

[37] Regarding the use of unvalidated algorithms, *see, e.g.* Nicholas Scurich and Daniel Krauss, *A Widely Used Criminal Justice Algorithm for Assessing Child Pornography Recidivism is Flawed*, Scientific American, May 20, 2024; regarding lack of validation of traditional forensics, *see, e.g.* David Faigman, Nicholas Scurich, and Thomas Albright, *The Field of Firearms Analysis is Flawed,* Scientific American, May 25, 2022.

[38] *See* Paul C. Giannelli & Sarah Antonucci, *Forensic Experts and Ineffective Assistance of Counsel,* 48 No. 6 CRIM L. BULLETIN 8 (2012).

[39] State v. Arteaga, Docket No. A-3078-21 *33 (App. Div. 2023).

[40] *Id*. at 99.

[41] Grimm, Grossman, and Cormack, *supra*, at 90.

light and the judges have realized that there were important reasons why that evidence could have resulted in exculpatory information. Often those revelations occur many years after a conviction.[42]

Conclusion

A strong presumption of interpretability, testing, and transparency for criminal investigation and courtroom uses of AI is essential. There may also be reasons to protect certain types of AI systems from "open" disclosure outside of court, for which the presumption may be overcome. For example, if there is a strong national security justification for not making the full AI model public, at a minimum, it should be carefully vetted by independent researchers, with complete information about its strengths and limitations made available to users in law enforcement and the courts.

This is an area where the federal government needs to lead in showing that use of AI robustly protects constitutional rights.

We close by emphasizing that *in order for evidence based on the use of AI to be statistically valid, complete information about data, variability, and error, must be disclosed in court – in short, it must be transparent and interpretable AI, and outside researchers must be able to evaluate it independently, including with access to test data, training data, code, and software*.

Government secrecy should never be the norm for AI systems. The federal government should lead by example, and *because the Constitution, including the Due Process Clauses should be understood to require it*. The existing Bill of Rights provides important protections as against arbitrary action, without notice, and action that harms defendant's fair trial and defense rights, as well as against discriminatory action in violation of the Equal Protection Clause and implementing civil rights acts. However, those constitutional and statutory protections have not proved effectual as remedies in criminal cases, given limited pretrial discovery, inadequate defense resources, and a tradition of deferential gatekeeping regarding expert evidence.

We ask NIJ attend to these basic principles of interpretable, transparent, and tested AI and careful and robust adherence to existing constitutional criminal procedure rights, as it conducts the important work of complying with the 2023 Executive Order, as well as the 2024 OMB Memorandum.

[42] NAS Report, supra, at 44-45 (describing audits and quality control failures at labs around the country).