THE RIGHT TO A GLASS BOX: RETHINKING THE USE OF ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE

Brandon L. Garrett† *& Cynthia Rudin*††

Artificial intelligence ("AI") increasingly is used to make important decisions that affect individuals and society. As governments and corporations use AI more pervasively, one of the most troubling trends is that developers so often design it to be a "black box." Designers create AI models too complex for people to understand or they conceal how AI functions. Policymakers and the public increasingly sound alarms about black box AI. A particularly pressing area of concern has been criminal cases, in which a person's life, liberty, and public safety can be at stake. In the United States and globally, despite concerns that technology may deepen pre-existing racial disparities and overreliance on incarceration, black box AI has proliferated in areas such as: DNA mixture interpretation; facial recognition; recidivism risk assessments; and predictive policing. Despite constitutional criminal procedure protections, judges have often embraced claims that AI should remain undisclosed in court.

Both champions and critics of AI, however, mistakenly assume that we inevitably face a trade-off: black box AI may be incomprehensible, but it performs more accurately. But that is not so. In this Article, we question the basis for this assumption, which has so powerfully affected judges, policymakers, and academics. We describe a mature body of computer science research showing how "glass box" AI—designed to be fully interpretable by people—can be more accurate than the black

† L. Neil Williams, Jr. Distinguished Professor of Law, Duke University School of Law and Faculty Director, Wilson Center for Science and Justice.

†† Earl D. McLean, Jr. Professor of Computer Science, Electrical and Computer Engineering, Statistical Science, Mathematics, and Biostatistics & Bioinformatics, Duke University.

Many thanks for their invaluable comments to Sara Sun Beale, Stuart Benjamin, James Boyle, Christopher Buccafusco, Walter Dellinger, Andrew Guthrie Ferguson, Lisa Griffin, Ben Grunwald, Laurence Helfer, Songman Kang, Maggie Lemos, Abraham Meltzer, Haijin Park, Jed Purdy, Arti Rai, Neil Siegel, Jeff Ward, Jonathan Weiner, the participants in an early ideas lunch at Duke Law School, a faculty workshop at Duke Law School, a conference at University of Warwick, and a workshop at Hanyan University School of Law.

box alternatives. Indeed, black box AI performs predictably worse in settings like the criminal system. After all, criminal justice data is notoriously error prone, and it may reflect pre-existing racial and socioeconomic disparities. Unless AI is interpretable, decisionmakers like lawyers and judges who must use it will not be able to detect those underlying errors, much less understand what the AI recommendation means.

Debunking the black box performance myth has implications for constitutional criminal procedure rights and legislative policy. Judges and lawmakers have been reluctant to impair the perceived effectiveness of black box AI by requiring disclosures to the defense. Absent some compelling—or even credible—government interest in keeping AI black box, and given the substantial constitutional rights and public safety interests at stake, we argue that the burden rests on the government to justify any departure from the norm that all lawyers, judges, and jurors can fully understand AI. If AI is to be used at all in settings like the criminal system—and we do not suggest that it necessarily should—the presumption should be in favor of glass box AI, absent strong evidence to the contrary. We conclude by calling for national and local regulation to safeguard, in all criminal cases, the right to glass box AI.

Introduction

The rapid growth in the use of artificial intelligence ("AI"), now a "constant presence" in our daily lives,[1] which some AI experts and leaders fear poses "societal-scale risks,"[2] has far outpaced our legal system's ability to regulate the technology and ensure that our rights are protected.[3] This global challenge has been deepened by the pervasive use of "black box" AI designed to be non-interpretable, meaning that its processes cannot be fully understood by laypeople or even by experts.[4] The growing uses of black box AI by governments and private corporations can have substantial negative consequences for people: "[h]idden algorithms can make (or ruin) reputations . . . or even devastate an entire economy."[5]

---

1 Herbert B. Dixon, Jr., *Artificial Intelligence: Benefits and Unknown Risks*, 60 Judges J. 41, 41 (2021).

2 Ctr. for AI Safety, Statement on AI Risk (2023), https://www.safe.ai/statement-on-ai-risk [https://perma.cc/4X85-ZVTN]; *see also* Off. Sci. & Tech. Pol'y, Blueprint for an AI Bill of Rights (2023), https://www.whitehouse.gov/ostp/ai-bill-of-rights [https://perma.cc/589L-TMLP] ("Among the great challenges posed to democracy today is the use of technology, data, and automated systems in ways that threaten the rights of the American public.").

3 For recent work regarding the need for new rights in the AI context, see, e.g., Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 Colum. L. Rev. 1957 (2021); Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 Emory L.J. 797, 800–01 (2021); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1, 27 (2014); *see also infra* subpart II.A.

4 For an overview, see, e.g., Stanford Univ., Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence 6–7 (2016) ("AI is already changing our daily lives."). As Frank Pasquale puts it: "black box AI" refers to any computer system "which uses data not accessible to the data subject, and/or which deploys algorithms which are either similarly inaccessible, or so complex that they cannot be reduced to a series of rules and rule applications comprehensible to the data subject." Frank Pasquale, *Normative Dimensions of Consensual Application of Black Box Artificial Intelligence in Administrative Adjudication of Benefits Claims*, 84 Law & Contemp. Probs. 35, 35–36 (2021).

5 Frank Pasquale, *About this Book: The Black Box Society*, Harv. Univ. Press, https://www.hup.harvard.edu/catalog.php?isbn=9780674970847 [https://perma.cc/9MN2-F2ZH].

One important focus of pressing legal and policy concern has been the criminal justice system, where black box AI poses risks to both public safety and to fundamental human and constitutional rights.[6] Already, police agencies have widely adopted largely unregulated and black box AI systems, often in partnership with technology corporations.[7] This is a global problem, since use of such systems in criminal cases is growing at the local and national levels.[8] Amidst broader public concern with the scale of incarceration and racial disparities, some have called for an end to the "tech to prison pipeline"[9]—the use of new technologies that magnify surveillance, detention, and discrimination.[10] Indeed, criminal defendants have launched challenges, with limited success, to the use of black box AI to analyze complex DNA mixtures;[11] conduct risk assessments used in pretrial decision-making and sentencing;[12] and as part of facial recognition systems used by law enforcement to identify

---

6 Elisa Jillson, *Aiming for Truth, Fairness, and Equity in Your Company's Use of AI*, FED. TRADE COMM'N: BUS. BLOG, (Apr. 19, 2021), https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai [https://perma.cc/T4WZ-2TA5] (providing overview of relevant legal rules). Several pieces of federal legislation would regulate algorithms, but none enacted, while several states adopted limited legislation, which we critique. *See infra* subpart III.C.

7 *See, e.g.*, Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 1975–77 (2017); John Monahan, *Risk Assessment in Sentencing*, *in* 4 REFORMING CRIMINAL JUSTICE: PUNISHMENT, INCARCERATION, AND RELEASE 77, 79 (Erik Luna ed., 2017).

8 *See, e.g.*, FAIR TRIALS, AUTOMATING INJUSTICE: THE USE OF ARTIFICIAL INTELLIGENCE & AUTOMATED DECISIONMAKING SYSTEMS IN CRIMINAL JUSTICE IN EUROPE 2 (2021) [hereinafter FAIR TRIALS], https://www.fairtrials.org/articles/publications/automating-injustice/ [https://perma.cc/MYQ6-6SH3] (describing growing use of AI tools in European criminal justice settings with "little or no safeguards").

9 Coal. for Critical Tech., *Abolish the #TechToPrisonPipeline*, MEDIUM (June 23, 2020), https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16 [https://perma.cc/373D-GXHM].

10 *See* Jessica M. Eaglin, *Technologically Distorted Conceptions of Punishment*, 97 WASH. U. L. REV. 483, 485 (2019).

11 *See, e.g.*, State v. Pickett, 246 A.3d 279, 284 (N.J. Super. Ct. App. Div. 2021) (ruling TrueAllelle source code regarding DNA analysis must be disclosed to defense); *see also* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFF. OF THE PRESIDENT, FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURECOMPARISON METHODS 8 (2016) [hereinafter PCAST Report] (discussing scientific limitations of probabilistic genotyping software when used to examine complex mixtures).

12 For the most prominent ruling, see State v. Loomis, 881 N.W.2d 749, 763–64 (Wis. 2016) (ruling risk assessment information used in sentencing need not be disclosed to the defense). *See also* Case Comment, *Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*: State v. Loomis, 130 HARV. L. REV. 1530, 1530 (2017).

suspects.[13] Yet, "[o]ne of the major obstacles to challenging potential civil rights abuses via algorithm is the opacity of such 'black box' technology."[14] As it stands, judges have largely permitted police and prosecutors to use black box AI.[15]

For example, when a federal judge took the unusual step of ordering that the Office of the Chief Medical Examiner in New York City disclose the source code for its probabilistic genotyping software, used to analyze mixtures of DNA, a series of concerns regarding accuracy came to light, and the software was subsequently discontinued.[16] In a 2019 ruling, the trial judge found that it was an error to rely on such evidence and suggested that any convictions that resulted from use of the software should be reviewed.[17] The judge emphasized that the software was a "black box'" which no independent expert could examine.[18] This was particularly concerning, the judge noted, where "[e]stimates as to the likelihood of an incorrect conclusion where there actually are four or more contributors [to the DNA sample] run to over 50%."[19]

Many other judges have instead emphasized the need to permit the use of black box AI, even in criminal cases. Thus, for the same type of AI used to examine DNA mixtures, a Pennsylvania appellate court rejected a defense challenge, denying the request for review by independent scientists of the underlying "proprietary" software.[20] The court emphasized "it would not be possible to market" the software "if it were

---

13 *See* Jack Karp, *Facial Recognition Software Sparks Transparency Battle*, Law360 (Nov. 3, 2019), https://www.law360.com/articles/1215786/facial-recognition-software-sparks-transparency-battle [https://perma.cc/G3RP-WASX].

14 *See, e.g.*, Michelle Chen, *Defund the Police Algorithms*, The Nation, (Aug. 25, 2022), https://www.thenation.com/article/society/police-algorithms-artificial-intelligence/ [https://perma.cc/H76E-M6ZB].

15 *See infra* subpart III.C. Regarding inadequate legislative solutions, see Chen, *supra* note 14 (describing that New York City police repeatedly failed to make disclosures under the New York City regulations requiring disclosures of surveillance technology).

16 *See* Order at 1, United States v. Johnson, No. 1:15cr00565 (S.D.N.Y. June 7, 2016) (order granting disclosure of FST source code); Lauren Kirchner, *New York City Moves to Create Accountability for Algorithms*, ProPublica (Dec. 18, 2017), https://www.propublica.org/article/new-york-city-moves-to-create-accountability-for-algorithms [https://perma.cc/9USL-XSRD]. *See* Lauren Kirchner, *Forensic Statistical Tool Source Code*, GitHub (Oct. 20, 2017), https://github.com/propublica/nyc-dna-software [https://perma.cc/365S-G8Q3].

17 People v. Thompson, No. 4346/15, slip op. at 1 (N.Y. Sup. Ct. Sept. 25, 2019).

18 *Id.* at 6.

19 *Id.*

20 Commonwealth v. Foley, 38 A.3d 882, 889 (Pa. Super. 2012).

available for free."[21] Developing a market for a product that serves the public interest could be a laudable goal. However, when one opens the black box, one quickly realizes that the underlying AI may not be worth paying for, and instead, it can pose substantial costs to both fairness and public safety. Other courts, like the New York Court of Appeals, tolerate similar proprietary use of AI in criminal cases by concluding it is reliable, based on studies done by the corporate provider, and placing the burden on the defense to show a "particularized" need for access.[22] Such rulings too readily assume that black box AI systems have been demonstrated accurate.

We write to counter the widely-held myth that the use of such black box systems are a necessary evil, because they have a supposed performance advantage over simpler or interpretable AI systems.[23] In academic and policy debates, both champions and critics of black box AI argue—mistakenly—that we face a catch-22: while black box AI is not understandable, they assume that it achieves far greater predictive accuracy.[24] More insidiously, some corporate and judicial proponents claim these systems represent innovation and higher performance, and contend that government should support private markets for the creation of such black box technologies—even if they eviscerate the constitutional rights of criminal defendants.[25]

Many of the most trenchant critics of black box AI similarly emphasize how AI supposedly derives its efficiency and effectiveness from its "inherently uninterpretable" associations and processes.[26] One called it as difficult to understand black

---

21 *Id.*

22 People v. Wakefield, 195 N.E.3d 19, 28–29 (N.Y. 2022), *cert. denied*, 143 S. Ct. 451 (2022) ("Defendant and the concurrence raise the legitimate concern that the technology at issue is proprietary and the developer of the software is involved in many of the validation studies. This skepticism, however, must be tempered by the import of the empirical evidence of reliability demonstrated here and the acceptance of the methodology by the relevant scientific community.")

23 *See, e.g.*, Ričards Marcinkevičs & Julia E. Vogt, *Interpretability and Explainability: A Machine Learning Zoo Mini-tour*, ARXIV, Dec. 3, 2020, at 2 (describing "a widespread belief that there exists a tradeoff between accuracy and interpretability"). *See infra* subpart II.C.

24 *See, e.g.*, Davide Castelvecchi, *Can We Open the Black Box of AI?*, 538 NATURE 20, 21 (2016) (calling it "exponentially harder" today and "more urgent" to decipher "the black box" of AI).

25 For an extensive discussion of such claims, see Natalie Ram, *Innovating Criminal Justice*, 112 NW. U. L. REV. 659 (2018). *See also* Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1520 ("Mandating interpretability might render the process less complex and therefore less accurate.").

26 *See* Arun Rai, *Explainable AI: From Black Box to Glass Box*, 48 J. ACAD. MARK. SCI. 137, 138 (2020) ("[D]eep learning algorithms are a class of ML algorithms

box AI as it is to "understand the networks inside" the human brain.[27] Another stated that since "it may not be possible to truly understand how a trained AI program is arriving at its decisions or predictions," we are faced with a choice whether to embrace or reject the black box.[28] Thus the claim that we face such a trade-off lies at the heart of efforts to both critique and retain the black box and often private control over AI technology.

This false dilemma appears to leave society in a bind. There is a need to improve on biased and fallible human decision-making, which has contributed to record levels of incarceration in the United States.[29] People cannot run database searches or regressions in their heads when making important decisions and instead can fall prey to biases. Yet one cannot even assess whether AI provides real benefits without first having interpretability—so that one can know how the AI works, how well it works, and how it is used in practice. Moreover, not only are the benefits of black box AI unclear, but the black box also obscures the costs. Black box AI can magnify racial biases in existing systems, such as criminal justice,[30] and early uses of AI in criminal justice have realized many critics' worst fears regarding errors, racial bias, punitiveness, non-transparency, and privacy invasions.[31]

In this Article, we argue that AI secrecy in the criminal system is far from necessary or inevitable. The black box problem does not involve a "tragic choice" which must be made in difficult

---

which sacrifice transparency and interpretability for prediction accuracy."); *see also* Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 886 (2016) ("Interpretability comes at a cost, however, as an interpretable model is necessarily simpler—and thus often less accurate—than a black box model.").

27 *See* Castelvecchi, *supra* note 24.

28 Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 890, 892 (2018).

29 NAT'L RSCH. COUNCIL, THE GROWTH OF INCARCERATION IN THE UNITED STATES: EXPLORING CAUSES AND CONSEQUENCES 2 (2014); Dorothy E. Roberts, *The Social and Moral Cost of Mass Incarceration in African American Communities*, 56 STAN. L. REV. 1271, 1272 (2004).

30 *See, e.g.*, Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1257 (2020); Kimberly A. Houser, *Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decisionmaking*, 22 STAN. TECH. L. REV. 290, 290 (2019); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 671 (2016); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1023 (2017).

31 *See, e.g.*, Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109, 1109 (2017).

circumstances of scarcity.[32] Rather, secrecy is an avoidable poor policy choice. In the criminal system, both fairness and public safety benefit from glass box AI, and therefore, judges and lawmakers should firmly recognize a right to glass box AI in criminal cases.

In Part I of this Article, we begin by introducing what AI is, how AI systems are developed, and we review several types of AI systems. We discuss three basic challenges confronting AI systems: (1) the problems of training and input data; (2) validation; and (3) interpretation and explanation. We then describe how AI has been used in criminal settings in the areas of: (1) recidivism risk assessments; (2) facial recognition; (3) predictive policing; (4) crime series detection; and (5) forensic evidence.

In Part II, we explore the advantages of "glass box" AI, by first dispelling technological and legal misperceptions about AI systems.[33] We describe how a range of commentators and scholars claim, erroneously, a black box performance advantage. We counter that when one examines the growing body of computer science research, one discovers that black box systems consistently underperform and disguise errors. In contrast, with glass box AI, not only are results made understandable in more simple ways, but certain models can be unpacked so that the relevant factors are understood in relationship to individual decisions (the concept of interpretability).[34] Effectiveness is not lost by requiring such transparency.[35] The three basic challenges that face all AI systems pose particular challenges in criminal cases. First, regarding data, criminal justice data is often noisy, highly

---

32 Guido Calabresi and Philip Bobbit famously use the phrase to refer to difficult policy choices regarding allocating scarce resources, that societies therefore regard as tragic. GUIDO CALABRESI & PHILIP BOBBIT, TRAGIC CHOICES 131–46 (1978).

33 *See* Rai, supra note 26 (describing "inherently interpretable" AI models).

34 By "interpretable" AI, we refer to models that are inherently interpretable, while by "explainable" we refer to efforts to provide post hoc explanations for models, which could be black box models. Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NAT. MACH. INTEL. 206, 206 (2019); *see also* Marcinkevičs & Vogt, *supra* note 23.

35 We earlier submitted a short statement responding to the White House Office of Science and Technology Policy ("OSTP") call for input on an AI Bill of Rights. Eric Lander & Alondra Nelson, *Americans Need a Bill of Rights for an AIpowered World*, WIRED (Oct. 8, 2021), https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/ [https://perma.cc/TWG2-2VAM]; Notice of Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies, 86 Fed. Reg. 56,300 (Oct. 8, 2021).

selected and incomplete, and full of errors. Second, using glass box AI, we can validate the system and detect and correct errors. Third, interpretability is particularly important in legal settings, where human users of AI, such as police, lawyers, judges, and jurors, cannot fairly and accurately use what they cannot understand.[36]

In Part III, we examine judicial rulings regarding AI in criminal cases and argue that interpretability should be understood as constitutionally required in most criminal settings. In criminal cases, judges have often deferentially approved black box AI systems, assuming that they offer greater reliability, even if they threaten constitutional rights.[37] These have not conducted a careful analysis informed by law and data science.[38] But due process and equal protection claims, as well as *Daubert* and Rule 702 standards for expert evidence, each place distinct burdens of justification on the government—and unfortunately, judges have often not insisted on a searching review of forensic evidence used in criminal cases.[39] The burden on the government to justify black box uses of AI in court should be high, given commitments to defense discovery rights, nondiscrimination, and reliability of evidence.[40] Additionally, we suggest that burden will rarely be met, given the lack of a strong performance justification for not making AI open for inspection, vetting, and explanation. Further, companies lack any clear innovation interest in concealing the effectiveness and accuracy of AI products used in criminal settings.[41] Thus, particularly in criminal cases with liberty at stake, there should be a strong legal, evidentiary, and constitutional right to glass box AI.

---

36 For a discussion of the problem of judicial reliance on risk assessment information, see Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439, 444 (2020).

37 *See infra* subpart II.A.

38 For a discussion of the lack of legal basis for trade secret protection for such uses of AI, see Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1343 (2018).

39 *See* Garrett & Monahan, *supra* note 36; *see also* Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245, 1300–01 (2016) (discussing due process and Confrontation Clause concerns with AI evidence in criminal cases); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1083–101 (2019) (discussing equal protection challenges to uses of AI in criminal justice).

40 For a discussion of how similar commitments flow from international human rights treaties and obligations, see FAIR TRIALS, *supra* note 8, at 31–32.

41 For an overview, see Brandon L. Garrett & M. Chris Fabricant, *The Myth of the Reliability Test*, 86 FORDHAM L. REV. 1559–99 (2018). *See also* BRANDON L. GARRETT, AUTOPSY OF A CRIME LAB: EXPOSING THE FLAWS IN FORENSICS 122–38 (2021).

In Part IV, we call for a glass box legislative and regulatory agenda, with a high burden of justification required for any black box use of AI in criminal settings. To date, no legislative enactments or proposals in the United States have required open or glass box AI. In contrast, a 2016 revision to the European Union's Law Enforcement Directive ("LED") limited the use of AI in criminal cases,[42] and the AI Act in Europe will provide more substantial regulation of AI systems in high stakes settings.[43] We argue that absent interpretability requirements, however, any efforts to regulate AI in the criminal justice system will fail to adequately safeguard rights.[44] We will describe how high-profile examples, like the federal First Step Act, illustrate what can go wrong when AI is used in the criminal system without a glass box approach.[45] Without a performance justification, the burden shifts dramatically to the government to explain why it keeps AI secret. Thus, legislation should aim to safeguard a right to glass box AI in criminal cases.

We conclude by emphasizing that there is no necessary tradeoff between any benefits of AI and the need to resort to black box systems. If we are to use AI in criminal cases, glass box AI can far better achieve public safety goals while protecting crucial, constitutionally guaranteed rights.

---

42 Directive 2016/680, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data by Competent Authorities for the Purposes of the Prevention, Investigation, Detection or Prosecution of Criminal Offences or the Execution of Criminal Penalties, and on the Free Movement of Such Data, and Repealing Council Framework Decision 2008/977/JHA, 2016 O.J. (L 119) 89 [hereinafter LED].

43 The European Parliament passed the act, with revisions, on June 14, 2023, and talks will now begin with the European Commission, the Council of the European Union, and the Parliament, regarding the final text of the law. *EU AI Act: First Regulation on Artificial Intelligence*, EUR. PARL. (June 14, 2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [https://perma.cc/S8ES-5WU8]. For current text of the Act, see European Parliament, Draft European Parliament Legislative Resolution on the Proposal for a Regulation of the European Parliament and of the Council on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, EUR. PARL. DOC. (COM 206) (2023) [hereinafter Artificial Intelligence Act], https://www.europarl.europa.eu/doceo/document/A-9-2023-0188_EN.html#_section1 [https://perma.cc/WBJ5-ZLQU]. For prior text of the Act, see *Proposal for a Regulation of the European Parliament and of the Council, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 final (June 14, 2023).

44 *See infra* Part III.

45 *See infra* subpart IV.B.

I
AI IN CRIMINAL JUSTICE

In this Part, we provide a primer on artificial intelligence, focusing on describing how AI is developed, step by step, and defining key concepts and terms. Second, we turn to the criminal justice system and the sources of data in that system that raise special challenges for uses of black box AI. We discuss three challenges to AI development in the criminal justice context: (1) training and input data; (2) validation; and (3) interpretation and explanation. Third, we describe how AI has been used in criminal settings in the areas of: (1) recidivism risk assessments; (2) facial recognition; (3) predictive policing; (4) crime series detection; and (5) forensic evidence.

A. An AI Primer

This section introduces and defines key terms, such as "artificial intelligence," "deep learning," "explainable," "evaluation," "interpretable," "predictive model," and "transparent." We believe that it is important to be precise about definitions of AI concepts. Whether the same terminology is used, both the theoretical computer science and the legal communities need to be more consistent with the use of these concepts.

Artificial Intelligence

"Artificial intelligence" simply refers to machines that perform tasks that are typically performed by humans and that normally require human intelligence.[46] Those tasks can include

---

46 *See, e.g.*, B.J. Copeland, *Artificial Intelligence*, ENCYC. BRITANNICA, https://www.britannica.com/technology/artificial-intelligence [https://perma.cc/DW48-WFDR] ("[T]he ability of a digital computer or computercontrolled robot to perform tasks commonly associated with intelligent beings."). For a wonderful overview of the definition of artificial intelligence and its history, see John McCarthy, *What is AI? / Basic Questions*, PROF. JOHN MCCARTHY (2004), http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html [https://perma.cc/H83L-PQAY] ("It is the science and engineering of making intelligent machines, especially intelligent computer programs . . . ."). Alan Turing influentially defined intelligence in the context of computing, focusing on systems that can indistinguishably think and act like humans. A.M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433, 433 (1950). More recently, Stuart Russell and Peter Norvig have defined artificial intelligence to both include human approaches to problems, focusing on acting and thinking like humans, and ideal approaches, that think and act rationally, but not necessarily based on existing human approaches. *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (4th ed. 2021). There is no single definition of artificial intelligence in legal usage in the United States. The John S. McCain National Defense Authorization Act for Fiscal Year 2019 included the first definition of AI in federal statute in the United States. Pub. L. No. 115–232, 132 Stat. 1636 (2018). It included a series of definitions,

speech recognition and generation, visual perception, decision-making, or translation between languages. In general, the goal of AI is to solve problems. Doing so often involves probabilistic reasoning, that is, making decisions based on prior knowledge, and quantifying uncertainty when one does so.[47]

Machine Learning

"Machine learning" is a subfield of AI, and it heavily overlaps with predictive statistics.[48] We should think of machine learning as a kind of pattern-mining, where algorithms are looking for useful patterns in data.[49] The data is supplied to the machine, which relies on past patterns to develop methods for making recommendations for what to do next.

Deep Learning

"Deep learning" refers to neural networks, which are a specific type of machine learning model that is particularly useful for image analysis, sound wave analysis, text generation, and other types of complex signals.[50] Neural networks use compositions of functions (i.e., a function of a function of a function, etc.) which makes their calculations particularly difficult for a human to understand, but also gives these models powerful predictive capacity.[51]

For instance, when predicting whether someone might be at a risk of suffering a drug overdose, patterns in their medical record and social media feeds, as well as those of others, might help a machine learning method predict the likelihood of that unfortunate outcome. This information can help human decisionmakers because no human can calculate patterns from large databases in their heads. Moreover, individual people may be biased or place undue weight on information that is not particularly predictive. If we want humans to make better

---

including, "[a]n artificial system designed to think or act like a human, including cognitive architectures and neural networks." *Id.* at 1697. The National Artificial Intelligence Initiative Act of 2020, established federal priorities for the use of AI, as part of the National Defense Authorization Act for Fiscal Year 2021, contained this definition: "The term 'artificial intelligence' means a machinebased system that can, for a given set of humandefined objectives, make predictions, recommendations or decisions influencing real or virtual environments." Pub. L. No. 116–283, 134 Stat. 3388, 4524 (2021). That same definition was used in the CHIPS and Science Act of 2022. Pub. L. No. 117–167, 136 Stat. 1366, 1405 (2022) (providing support for development of safe, secure, and trustworthy AI systems).

47 *See* RUSSELL & NORVIG, *supra* note 46, at 208.

48 *See id.* at 651.

49 *See id.*

50 *See id.* at 750.

51 *Id.*

data-driven decisions, machine learning can help with that. Simply put, machine learning methods can extract patterns from large databases that humans cannot. However, humans have a broader systems-level way of thinking about problems that is absent in AI.

It is also important to distinguish several key terms relevant to understanding an AI system: "algorithm," "predictive model" (or just "model"), and "evaluation procedure." We will use terminology specific to computer science and use the task of recidivism prediction as an example to illustrate how to define each key term.

Algorithm

An algorithm is a set of instructions to be followed when making a calculation. An algorithm need not be created by machine learning or a form of AI. Many algorithms can and have been created by humans, and they can range from quite simple to complex.

Predictive Model

A "predictive model" is a formula that takes a new observation (represented by a set of features, such as statistics of a person's criminal history, age, prison misconduct history, and education) and produces a prediction (e.g., there is a 14% chance of re-arrest within 2 years of release). Predictive models can become black box models when their formulas are too complicated for humans to comprehend (e.g., a sum of exponentiated weighted distances between the new observation and each of the previous individuals in the database). Conversely, predictive models are glass box models, or "interpretable" models, when the formula is understandable by humans.

Interpretable

By "interpretable" AI, we refer to predictive models whose calculations are inherently understandable. For an interpretable AI system, a person can see how the AI system works and what information it relies upon in a particular instance. The predictive model is disclosed to the users. It provides information regarding the model, the factors used to provide a result, and how those factors were in fact combined to provide a result.

Explainable

By "explainable," we refer to a system that provides a post hoc explanation for its model, which could be a black box model. In effect, this approach uses proxies to explain what the AI may have done. In the next Part, we further discuss the distinction between interpretable and explainable AI.

Machine Learning Algorithm

These predictive models can be created by machine learning algorithms. A machine learning algorithm uses a database of past cases to create the model in a way that it is accurate for the past cases and, hopefully, predictive of future cases. The complexity or simplicity of the algorithm may vary, depending on the task. We emphasize that the complexity of the predictive model matters in practice, since it provides the information regarding how a prediction is made in a particular instance. The algorithm, which is used to create that predictive model, does so by choosing which features to use in the model and how to combine them based on the historical data, called the "training set."[52]

We also underscore that some modern machine learning methods are extremely complicated, yet they produce very simple predictive models. Some can be printed on an index card, as if they were created by a person. They could, for instance, appear similar to the Public Safety Assessment ("PSA"), which is used for pretrial risk assessment and uses just nine factors, which a person can score on a short worksheet, based on standard information such as a person's age, pending charge, and criminal record.[53] To provide another example, researchers found that a simple model relying on age, gender, and prior criminal record was just as predictable as the COMPAS algorithm, which is a proprietary black box model, and claims to rely on up to 137 inputs.[54] This was the entire model and explanation:

> [I]f the person has either >3 prior crimes, or is 18–20 years old and male, or is 21–23 years old and has two or three prior crimes, they are predicted to be rearrested within two years from their evaluation, and otherwise not.[55]

---

52 Warren E. Agin, *A Simple Guide to Machine Learning*, 2017 BUS. L. TODAY 4 (stating that to build a prediction model, one selects cases at random to use as a "training set," using the remainder as a "test set.").

53 For examples, see *Public Safety Assessment*, ADVANCING PRETRIAL POL'Y & RSCH., https://advancingpretrial.org/psa/factors/ [https://perma.cc/2TZR-LYX2] (last visited Feb. 8, 2024).

54 *See* Elaine Angelino, Nicholas LarusStone, Daniel Alabi, Margo Seltzer & Cynthia Rudin, *Learning Certifiably Optimal Rule Lists for Categorical Data*, 18 J. MACH. LEARNING RSCH., 2018 at 1; Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, SCI. ADVANCES, Jan 17, 2018, at 1.

55 *Id.*; Cynthia Rudin & Joanna Radin, *Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition*, HARV. DATA SCI. REV., Fall 2019, at 5.

This model was created by a complex machine learning algorithm that looked at many different factors and chose among them, combining them in a specific way to as to yield high accuracy. Indeed, researchers of recidivism risk prediction have long found that a small number of simple factors are predictive: largely age, gender, and prior criminal activity.[56]

Evaluation

The evaluation procedure is designed to assess how accurate the model's predictions are. After all, it is extremely important to know whether these predictions are reliable. Such an evaluation uses a new dataset, called a "test set," which must be separate from the one used by the algorithm to train the model. The evaluation involves determining whether the predictions made on the test set are accurate and lead to better decisions. While some types of evaluations narrowly evaluate the prediction quality of a model on a test set, evaluations can more broadly consider how a human-in-the-loop performs when working with the algorithm to make decisions.

To summarize, an algorithm produces a predictive model, which is then evaluated on a test set, consisting of separate data. We advocate for both the model to be interpretable and for the evaluation procedure to be transparent and reproducible. Thus, for a risk assessment instrument, not only should a check-sheet given to a judicial officer be simple and understandable, but the underlying evaluations should have been conducted on appropriate and separate data and shared publicly, so that others can reproduce the evaluation.[57]

Transparent

Model "transparency" is different than interpretability: transparency refers to 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 may be necessary to share test set and training data as well, to replicate the evaluations done in the past on a model. We view evaluation, or validation of the accuracy of a model, as extremely important. However, our focus in this Article is on interpretability.

In general, interpretability will often come with transparency. We note, though, that it is possible for a model to be interpretable

---

56 *See, e.g.*, John Monahan & Jennifer L. Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CLINICAL PSYCH. 489, 500–01 (2016).

57 *See, e.g.*, OPERATIONS & PROGRAMS DIV., JUD. COUNCIL OF CAL., PRETRIAL RISK ASSESSMENT TOOL VALIDATION (June 2021), https://www.courts.ca.gov/documents/Pretrial-Risk-Assessment-Tool-Validation_June-2021_FinalPosted.pdf [https://perma.cc/U8YX-KDR8].

but not transparent, in the sense that the reasoning process behind an individual prediction is shared, but one cannot validate the model on a test set because one does not have access to the full model. It is also possible for a model to be transparent but not interpretable, which is the case for most public models whose formulas are too complicated to understand.

B. Uses of AI in Criminal Justice

The use of AI is increasingly pervasive in large and small law enforcement agencies and jurisdictions. Increasing the use of AI in criminal settings has been a priority of government funders; for example, the National Institute of Justice, in explaining its priorities, summarizes: "Artificial intelligence has the potential to be a permanent part of our criminal justice ecosystem, providing investigative assistance and allowing criminal justice professionals to better maintain public safety."[58] Similarly, the European Union has prioritized the use of AI in criminal justice, although also calling for risk assessments regarding legal, ethical and fundamental rights implications.[59] Use of AI is pervasive in ways that raise grave human rights concerns, such as in China, where police engage in mass surveillance of electronic data,[60] and where the State Counsel has declared intent to use AI in a range of judicial decision-making, including sentencing.[61] There is a surveillance industry, and leading private technology companies market law enforcement-related products around the world.[62] Already, a range of AI systems have been used by criminal justice actors, including: (1) risk assessments; (2) facial recognition; (3) predictive policing; (4) crime series detection;

---

58 *See* Christopher Rigano, *Using Artificial Intelligence to Address Criminal Justice Needs*, NAT'L INST. JUST. J., no. 280, Jan. 2019, at 1, 37, 38 (noting that the National Institute of Justice is "committed to realizing the full potential of artificial intelligence to promote public safety and reduce crime").

59 *See* EULISA AND EUROJUST, ARTIFICIAL INTELLIGENCE SUPPORTING CROSSBORDER COOPERATION IN CRIMINAL JUSTICE 6 (June 2022), https://www.eulisa.europa.eu/Publications/Reports/AI%20in%20Justice%20-%20Report.pdf [https://perma.cc/79QL-CF32].

60 *See* Paul Mozur, Muyi Xiao & John Liu, '*An Invisible Cage': How China is Policing the Future*, N.Y. TIMES (June 25, 2022), https://www.nytimes.com/2022/06/25/technology/china-surveillance-police.html [https://perma.cc/V3MC-BHY9].

61 *See* Jiahui Shi, *Artificial Intelligence, Algorithms and Sentencing in Chinese Criminal Justice: Problems and Solutions*, 33 CRIM. L.F. 121, 121 (2022).

62 Steven Feldstein, *The Global Expansion of AI Surveillance*, CARNEGIE ENDOWMENT INT'L PEACE (Sept. 17, 2019), https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847 [https://perma.cc/4PX5-FYA7].

and (5) forensic evidence. To provide an introduction to AI in criminal justice, we briefly describe each of these uses and how they have commonly involved black box approaches but rarely use the glass box approaches we recommend.

1. *Risk Assessments*

Risk assessments play a role in many stages of the criminal justice system in the United States, informing decisions regarding pretrial detention, sentencing, corrections, and reentry.[63] Risk assessment is the use of factors that can estimate the likelihood of an outcome occurring in a population.[64] In criminal justice, risk assessment tools typically seek to predict the likelihood (or probability) of recidivism, for example, whether a person will be re-arrested or convicted for new criminal charges.[65] Two types of errors can occur when making such predictions: "False positive" predictions are when a judicial officer releases a person based on a prediction of low risk and the person then commits a new crime. "False negative" predictions are when a person predicted as low-risk, who would not likely have committed a new crime, is nevertheless jailed. [66]

Risk assessment algorithms date back to at least 1928, but before risk assessment algorithms were commonly used, instead, individual judicial officers made these types of predictions on their own.[67] Beginning in the 1970s, the focus of pretrial decision-making (e.g., whether to release a person on their own recognizance, set bail of some amount, or detain a defendant without bail) was a broadly defined definition of "dangerousness" in which judges had discretion to decide whether a person arrested for a crime should be jailed pretrial, if the person was deemed "dangerous" or likely to fail to appear for future court dates.[68] That discretion was not informed by data regarding whether a person actually did pose a risk of re-arrest or non-appearance.

In sentencing, risk assessments disappeared from judicial practice for a different reason: beginning in the 1970s, legislators

---

63 For an overview, see Garrett & Monahan, *supra* note 36, at 440–41.

64 Helena C. Kraemer et al., *Coming to Terms with the Terms of Risk*, 54 ARCHIVES GEN. PSYCHIATRY 337, 340 (1997).

65 *See* Garrett & Monahan, *supra* note 36, at 449.

66 *Id.* at 450.

67 *See* Monahan & Skeem, *supra* note 56, at 490.

68 *See* Shima Baradaran & Frank L. McIntyre, *Predicting Violence*, 90 TEX. L. REV. 497, 506–07 (2012).

focused on retributivism and enacted harsher mandatory sentencing laws. This meant that more judges imposed sentences based on a defendant's past acts and criminal history, but without the ability to be forward-looking.[69] More recently, lawmakers have shifted to providing decisionmakers, like judges, with better empirical data concerning risk, rather than asking them to predict "dangerousness" themselves or to impose purely backward-looking sentences on persons.[70] There are real concerns with the racially disparate and arguably punitive decisions that judges make, using their discretion, at bail hearings and at sentencing. At the same time, however, others have raised concerns that risk assessments may not sufficiently improve the system, and they may introduce new harms.[71]

There are many hundreds of risk assessment tools in use, but most involve the same basic factors.[72] An algorithm or basic statistical analysis is used to examine outcomes using criminal justice data and to identify factors that can usefully predict recidivism outcomes. Research has shown quantitative assessments can be more reliable than the decisions that individuals make based on their intuitions and experience.[73] These risk assessments may also separate different types of risk (e.g., differentiating risk of any re-arrest, including for minor violations, or from the risk of arrest for a serious or violent offense), and some have been validated and designed for particular jurisdictions.[74]

Since most of these models are developed and used by government agencies, the vast majority are not black boxes, and the predictive models clearly set out how factors are scored and, as a result, why a person is labeled high or low risk.[75] Such transparency is beneficial because when these risk assessment instruments are used in bail contexts, or sentencing, or by probation officers, the predictive model is designed to inform, but the government officials still retain discretion to follow or ignore

---

69 *See* John Monahan & Jennifer L. Skeem, *Risk Redux: The Resurgence of Risk Assessment in Criminal Sanctioning*, 26 FED. SENT'G REP. 158, 158–59 (2014).

70 *See* Garrett & Monahan, *supra* note 36, at 452–53.

71 *See id.*

72 For a metaanalysis, see Sarah L. Desmarais, Samantha A. Zottola, Sarah E. Duhart Clarke & Evan M. Lowder, *Predictive Validity of Pretrial Risk Assessments: A Systematic Review of the Literature*, 48 CRIM. JUST. & BEHAV. 398 (2021).

73 *See* Christopher Slobogin, *A Jurisprudence of Dangerousness*, 98 NW. U. L. REV. 1, 1–2 (2003).

74 *See* Garrett & Monahan, *supra* note 36, at 452–53.

75 *See* Desmarais, Zottola, Duhart Clarke & Lowder, *supra* note 72.

that information, and they can understand the basis for the risk prediction.[76] Indeed, while the move to adopt risk assessments has been often motivated by a desire to reduce overreliance on incarceration, there is much research now on the degree to which judges commonly do not follow risk assessment recommendations, which can undermine their potential benefits.[77]

The use of risk assessments has become controversial, including, as we will discuss, due to concerns they can reproduce racially biased or otherwise unfair outcomes. For example, then-Attorney General Eric Holder questioned the use of risk assessment as potentially causing "fundamental unfairness."[78] There are related concerns regarding accuracy of the instruments, including whether they are trained on poor data or adequately validated.[79] The Model Penal Code recommended use of risk assessments in sentencing, but only if the instrument is regularly evaluated.[80] There is evidence that judges in some jurisdictions have been more likely to release low-risk individuals when they rely on risk assessments but also evidence that other judges have not paid attention to these risk assessments.[81] In past work, we have criticized the use of particular risk assessment instruments but also pointed to their potential to refocus judges on alternatives to incarceration.[82] If AI is used, it should be carefully evaluated, and it must be glass box. We will further discuss these concerns in sections that follow.

2. *Facial Recognition Technology*

Across the country driver's license photos and other images are being fed into a federal face recognition system.[83]

---

76 *Id.* at 400.

77 *See id.*

78 Joshua Barajas, *Holder: Big Data is Leading to 'Fundamental Unfairness' in Drug Sentencing*, PBS NEWS HOUR (July 31, 2014), https://www.pbs.org/newshour/politics/holder-big-data-leading-fundamental-unfairness-drug-sentencing [https://perma.cc/X2X3-TY7K].

79 *See* Garrett & Monahan, *supra* note 36, at 465–66.

80 MODEL PENAL CODE: SENTENCING § 6B.09(1) (AM. LAW INST., Proposed Final Draft April 10, 2017).

81 For an overview of Virginia data concerning the high variability of judges' use of a risk assessment permitting the release of lowrisk persons, see Garrett & Monahan, *supra* note 36, at 459–62.

82 *See* Garrett & Monahan, *supra* note 36; Brandon L. Garrett & Megan Stephenson, *Open Risk Assessment*, 38 BEHAV. SCI. & L. 279 (2020).

83 *The Use of Facial Recognition Technology by Government Entities and the Need for Oversight of Government Use of this Technology Upon Civilians: Hearing*

Facial recognition technology ("FRT"), generally involves trying to identify a person by trying to match an image of a face to a database of known faces, or to a reference face.[84] The uses of these systems vary, from face verification, used to confirm a person's portrait-type photo in an identification document, to face identification from images taken in the field, to face tracking to follow a person across locations.[85] Facial recognition searches are now extremely common; the Federal Bureau of Investigations ("FBI") conducts hundreds of thousands of searches each year, often on behalf of local police, who themselves may conduct large numbers of face searchers.[86]

These FRT systems can take measurements of faces, and code distances between major landmarks on the face, like the distances between the eyes or the width of the mouth.[87] (They can also scan the images looking for specific features on a face that are required by a neural network, though these features cannot be easily described.) Algorithms are trained on large datasets of millions of images to identify features to code that can more accurately compare face images.[88] First, unless using a mugshot or portrait, the software must identify a face within an image if one is present. Next, the program engages in "feature extraction" to identify major features in the face, such as the center of the eyes, the point of the nose, and the corners of the mouth. A series of measurements are made between those features, which are coded in a "faceprint."[89] An algorithm is then used to search through a dataset of many faceprints.

Unlike risk assessments, these FRT systems are mostly black box AI systems.[90] As a result, the accuracy (or lack thereof) is not well understood. The National Institute of Standards and

---

Before the H. Comm. on Oversight & Reform, 116th Cong. 3 (2019) [hereinafter Del Greco Statement on Facial Recognition] (statement of Kimberly J. Del Greco, Crim. Just. Info. Servs. Div., Fed. Bureau of Investigations).

84 *See* Kimberly N. Brown, *Anonymity, Faceprints, and the Constitution*, 21 GEO. MASON L. REV. 409, 428 (2014).

85 *See* Andrew Guthrie Ferguson, *Facial Recognition and the Fourth Amendment*, 105 MINN. L. REV. 1105, 1112–13 (2021).

86 *Facial Recognition Technology (Part II): Ensuring Transparency in Government Use: Hearing Before the H. Comm. on Oversight and Reform*, 116th Cong. 21 (2019) (statement of Kimberly J. Del Greco, Crim. Just. Info. Servs. Div., Fed. Bureau of Investigations).

87 For a description, see Ferguson, *supra* note 85, at 1111–15.

88 *Id.* at 1112.

89 *Id.* at 1111.

90 For a comprehensive description of how FRT systems work and their limitations, and detailed recommendations for regulating their use, the National Academy of Sciences completed a report in 2024. *See generally* NAT'L ACADS. OF

Technology ("NIST") has tested facial recognition algorithms and described improvements in accuracy.[91] In less controlled settings, such as when people are walking through an airport boarding gate or a sports venue, accuracy rates range broadly.[92] Further, error rates can be greater based on demographics.[93] A 2019 NIST study found, in testing 189 different algorithms from ninety-nine developers, that misidentified Black faces more than white faces—up to one hundred times more.[94]

Not only has there been little independent scientific review of these FRT systems, but these black box systems have not undergone judicial review of their reliability either. The FBI resisted calls, including by the United States Government Accountability Office ("GAO"), to audit the accuracy and uses of FRT, replying that it provides not a "positive identification," but rather an investigative lead.[95] When treated as a lead, the AI itself is not introduced as evidence at trial, although it may support probable cause for arrest. Thus, in *People v. Reyes*, a person stealing packages from a mailroom was caught on a security camera, and the New York Police Department's Facial Identification Section ran a search and located a single "possible match" mug shot.[96] A detective compared the mug shot to a still from the video, and decided the defendant was the culprit, but the prosecution did not seek to introduce the

---

Scis, Eng'g, & Med., Facial Recognition: Current Capabilities, Future Prospects, and Governance (2024).

91 *See generally* Patrick Grother, Mei Ngan & Kayee Hanaoka, Nat'l Inst. Standards & Tech., Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification (2019), https://www.nist.gov/system/files/documents/2019/11/20/frvt_report_2019_11_19_0.pdf [https://perma.cc/C9GZ-YXH2].

92 *See* Patrick Grother, George Quinn & Mei Ngan, Nat'l Inst. Standards & Tech., Face in Video Evaluation (FIVE) Face Recognition of Noncooperative Subjects 37 (2017) (describing the range in false negative identification rates found).

93 *See generally* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proc. Machine Learning Res. 1 (2018).

94 *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*, Nat'l Inst. Standards & Tech (Dec. 19, 2019), https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software [https://perma.cc/6SE4-YP7H]. Some algorithms appear to be more accurate, mainly those that scrape the internet for labeled photographs, raising privacy concerns in the process.

95 *See Law Enforcement's Use of Facial Recognition Technology: Hearing Before the H. Comm. on Oversight & Gov't Reform*, 115th Cong. 1 (2017) (statement of Kimberly J. Del Greco, Deputy Assistant Dir., Crim. Just. Info. Servs. Div., Fed. Bureau of Investigations).

96 133 N.Y.S. 3d 433, 435 (Sup. Ct. 2020).

FRT comparison into evidence.[97] The judge ultimately found: "Facial recognition analysis thus joins a growing number of scientific and near-scientific techniques that may be used as tools for identifying or eliminating suspects, but that do not produce results admissible at a trial."[98]

In an important ruling in *State v. Arteaga*, a New Jersey Appellate Court affirmed a trial court order, ruling that if the prosecutor plans 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."[99] That court also explained that such evidence is relevant to the accuracy of the human eyewitness identification, as well. The court noted that "[t]he FRT's reliability has obvious implications for the accuracy of the identification process because an array constructed around a mistaken potential match would leave the witness with no actual perpetrator to choose."[100] Further, the court noted that the reliability of the FRT system "bears direct relevance to the quality and thoroughness of the broader criminal investigation, and whether the potential matches the software returned yielded any other viable alternative suspects to establish third-party guilt."[101] Thus, properly viewed, the technology does not just supply "leads," it can affect human witnesses, and it can affect a criminal investigation and prosecution. The court concluded that the "[d]efendant must have the tools to impeach the State's case and sow reasonable doubt."[102]

The reliance on fallible human eyewitnesses raises real challenges in criminal cases, and has resulted in many wrongful convictions—a topic of substantial research.[103] The interaction between FRT technology and human witnesses will require similar research to find out how different uses and presentations of a facial recognition system affect eyewitnesses. For instance, FRT systems are now often designed to present the top five potential matches to the user in a random order, so as to force

---

97 *Id.*

98 *Id.* at 437.

99 296 A.3d 542, 557 (N.J. Super. Ct. App. Div. 2023).

100 *Id.*

101 *Id.*

102 *Id.* at 558.

103 *See* Thomas D. Albright & Brandon L. Garrett, *The Law and Science of Eyewitness Evidence*, 102 B.U. L. Rev. 511, 516 (2022).

the user to think carefully and hopefully reduce bias.[104] Access to video and search technology has the potential to improve upon traditional eyewitness identification procedures, but it could also raise reliability concerns if it is used poorly, and it also raises concerns regarding privacy, over-surveillance, and racial disparities. When the FRT is a black box, these harms cannot be easily investigated, much less addressed.

3. *Predictive Policing*

Law enforcement agencies have long made predictions regarding incidence of criminal activity, and they deploy police officers and resources based on those assessments of risk. Increasingly, AI has informed predictive decision-making by police, including by introducing social network analysis and other new tools to try to predict offending.[105] The concern, however, is that if police think that a neighborhood is higher risk, they may engage in more enforcement there, creating a "feedback loop" that does not reflect the actual public safety needs but rather can serve to justify invidious practices, like racial profiling.[106] Conversely, there is evidence that law enforcement, relying on their intuitions and not on AI, may broadly label neighborhoods as high crime, without regard for actual public safety risks.[107] As with the use of FRT, predictive policing AI generates leads or helps to prioritize deployment of police, but has not been introduced as evidence, and therefore has not been judicially reviewed. In several studies, the effects of predictive policing have been mixed, both regarding effectiveness in deploying police officers, and also in whether racial disparities may result;[108] new models are also under

---

104 William Crumpler & James A. Lewis, *How Does Facial Recognition Work? A Primer*, Ctr. Strategic & Int'l Stud., June 2021, at 3.

105 For an overview of predictive policing methods and research, see Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith & John S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations (2013).

106 *See* Ferguson, *supra* note 31, at 1148.

107 *See* Ben Grunwald & Jeffrey Fagan, *The End of Intuition-based High-crime Areas*, 107 Calif. L. Rev. 345, 369 (2019).

108 For a study finding that predictive policing had some deterrent effects, but also produced marked racial disparities, see Ranae Jabri, *Algorithmic Policing* (Nov. 2021) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4275083 [https://perma.cc/4AES-8E3R]. For a study finding that the AI program used in Los Angeles was much more accurate than human predictions regarding timing and locations of theft, see G.O. Mohler et al., *Randomized Controlled Field Trials of Predictive Policing*, 110 J. Am. Stat. Ass'n. 1399, 1399 (2015). For a study finding that the program in Los Angeles did not

development.[109] Los Angeles recently stopped the use of its predictive policing algorithms for the stated reason that they were not worth the expense given the value of information they provided.[110]

4. *Crime Series Detection*

Crime series detection is the problem of determining which crimes were committed by a single group of individuals.[111] AI can seek to identify groups of crimes that are similar in modus operandi. For instance, a crime series might consist of break-ins within a similar location, on weekdays, where the residents were not present, and where the offender entered through an unlocked door. Another crime series might be spread across locations, where the offender entered by pushing in the air conditioner and climbing through the window, on Thursdays during lunch hour, while the residents are not present. Identifying such a crime series from a database of crimes is a task that human analysts often attempt manually, using database queries for many possible modus operandi. Algorithms can speed this process up substantially. The problem of finding crime series is a clustering problem, but is not an ordinary clustering problem because the algorithm needs to find the modus operandi (the set of variables on which the crimes are similar) at the same time as finding the crimes themselves. This is called a

---

produce racial biases as compared with a control method for allocating police, see P. Jeffrey Brantingham, Matthew Valasik & George O. Mohler, *Does Predictive Policing Lead to Biased Arrests? Results from a Randomized Controlled Trial*, 5 STAT. & PUB. POL'Y, Apr. 9, 2018, at 1.

109 *See, e.g.*, Victor Rotaru, Yi Huang, Timmy Li, James Evans & Ishanu Chattopadhyay, *Eventlevel Prediction of Urban Crime Reveals a Signature of Enforcement Bias in US Cities*, 6 NATURE HUM. BEHAV. 1056 (2022).

110 Leila Miller, *LAPD Will End Controversial Program that Aimed to Predict Where Crimes Would Occur*, L.A. TIMES (Apr. 21, 2020), https://www.latimes.com/california/story/2020-04-21/lapd-ends-predictive-policing-program [https://perma.cc/SVR6-A9E5] (citing statements by the police chief that the program was not worth the cost but also that its effectiveness could not be shown); Johana Bhuiyan, *LAPD Ended Predictive Policing Programs Amid Public Outcry. A New Effort Shares Many of their Flaws*, THE GUARDIAN (Nov. 8, 2021) https://www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform [https://perma.cc/X65J-7LKG] (describing how the model used simplistically assessed where arrests had been made and sought to focus police response to those locations).

111 Regarding the definition of a crime series, see INT'L ASS'N OF CRIME ANALYSTS, CRIME PATTERN DEFINITIONS FOR CRIME ANALYSTS 3 (2021), https://www.iadlest.org/Portals/0/Files/Documents/DDACTS/Webinars/Automation/Lessons/CRIME%20PATTERN%20DEFS_IACA.pdf?ver=SWBnC2STwP-bH4HfyRVOQA%3D%3D [https://perma.cc/MQV4-G8B3].

subspace clustering problem.[112] Since 2016, New York City has been using algorithms for crime series detection.[113] These are used only to understand whether crimes are connected, and, to our knowledge, have not been adjudicated in court. The information gleaned can help investigators determine possible leads for unsolved cases. However, crime series detection AI is different from other examples discussed in that conclusions about individuals cannot be made from the results alone.

5. *Forensic Evidence AI*

A fundamental challenge for forensic examiners is linking evidence from a crime scene to a potential culprit. When law enforcement have no leads, they may try to compare a DNA profile, fingerprint or a toolmark (such as a spent bullet or shell casing) left at a crime scene to a database.[114] In the case of DNA databases, a numeric profile reflecting a defined set of a person's genetic markers is entered in a database of DNA profiles.[115] These DNA tests use genetic markers selected to be highly variable in the population, and therefore are useful to link evidence to particular individuals.[116]

Forensic AI has been introduced in court in the context of complex DNA mixtures. For DNA mixtures of multiple and sometimes unknown numbers of contributors, algorithms have

---

112 For an overview of the subspace clustering problem, see René Vidal, *A Tutorial on Subspace Clustering* (Jan. 2010), https://www.cis.jhu.edu/~rvidal/publications/SPM-Tutorial-Final.pdf [https://perma.cc/3WXP-Y4PX].

113 Tong Wang, Cynthia Rudin, Daniel Wagner & Rich Sevieri, *Learning to Detect Patterns of Crime*, Proceedings Euro. Conf. Machine Learning & Principles & Prac. Knowledge Discovery Databases (2013) (proposing a crime series detection algorithm); *see* Alex Chohlas-Wood & E.S. Levine, *A Recommendation Engine to Aid in Identifying Crime Patterns*, 49 INFORMS J. on Applied Analytics 154, 162 (2019) (examining the Patternizr systems used in New York, which built on the algorithm developed in Wang, Rudin, Wagner & Sevieri, *supra* note 113).

114 For pattern evidence, forensic examiners may use AI to search databases of images, based on features thought to help to predict correspondence between such objects. A human examiner examines each in a list of possible corresponding fingerprints, and at a trial, that expert testifies regarding the fingerprint comparison conducted. The operation of the AI search model is not introduced in court. For an overview of these databases, see, e.g., Roben Bowen & Jessica Schneider, *Forensic Databases: Paint, Shoe Prints, and Beyond*, Nat'l Inst. Just. (Oct. 1, 2007), https://nij.ojp.gov/topics/articles/forensic-databases-paint-shoe-prints-and-beyond [https://perma.cc/S2TB-QGNG].

115 *See generally Combined DNA Index System (CODIS)*, Fed. Bureau Investigations, https://www.fbi.gov/services/laboratory/biometric-analysis/codis [https://perma.cc/PTW2-GVJ5] (last visited Feb. 8, 2024) (providing an overview of the federal DNA index).

116 *Id.*

been designed to interpret the test results, to try to determine whether a suspect might or not have contributed to a sample from the crime. The scientific community has found these approaches, called probabilistic genotyping, promising, but not yet well validated outside certain well-defined ranges.[117] However, DNA mixture results have been introduced in court, and experts have claimed that their software is proprietary, protected by trade secrets, and that it would harm their business to share it with the defense, for purposes of evaluation.[118] Despite constitutional concerns with AI secrecy in a criminal prosecution, courts have often ruled against defense requests for access.[119] As we will discuss, where independent evaluation is not possible or permitted, there are substantial due process and policy concerns with permitting such black box AI results to be used as evidence.

II
The Black Box Performance Myth

There is a common misconception that black box AI is more accurate than any model that a human could understand. Thus, scholars have argued: "Requiring every AI system to explain every decision could result in less efficient systems, forced design choices, and a bias towards explainable but suboptimal outcomes."[120] Or, as another scholar put it simply: "making an algorithm explainable may result in a decrease in its accuracy."[121] Such claims are often repeated in the computer science, policy, and law literatures,[122] but on scrutiny, we argue that they lack support.

A. Black Box Performance Assertions

Some scholars in computer science, policy, and in law, have assumed that interpretable AI simply cannot be as accurate as black box AI. As an article in Scientific American put it, "today's AI

---

117 PCAST Report, *supra* note 11, at 148 (finding probabilistic genotyping approaches were validated only for DNA mixtures of three individuals in which the minor contributor consists in twenty percent of the sample).

118 Wexler, *supra* note 38, at 1358–62.

119 *Id.*

120 *See* Finale DoshiVelez et al., *Accountability of AI Under the Law: The Role of Explanation*, arXiv, Nov. 3, 2017, at 2.

121 Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 Colum. L. Rev. 1829, 1834 (2019).

122 *See* Rudin, *supra* note 34, at 206–07.

conundrum: The most capable technologies—namely, deep neural networks—are notoriously opaque, offering few clues as to how they arrive at their conclusions."[123] A writer in the MIT Technology Review called the black box problem a "dark secret at the heart of AI."[124] These statements have persisted as the technology surrounding AI has advanced. In 2021, two scholars stated, "the interpretability of 'black box' machine learning algorithms is a challenging technical problem for which no solutions have been found."[125] Researchers have focused on the complexity of machine learning algorithms. They, for example, find geometric patterns that "humans cannot perceive" because they connect variables across hundreds of dimensions.[126] Yet this is not so.

As interpretable and explainable AI approaches have become more common, as subject of computer science scholarship as well as used in society, it is increasingly understood that there is a glass box alternative.[127] However, many persist in viewing black box AI as superior to that alternative due to a perception of its super-human (and unintelligible) performance. Thus, some argue that "instead of worrying about the black box, we should focus on the opportunity," that AI technology may provide.[128] While this type of optimistic perspective about technology may be reasonable when the stakes are low, i.e., when incorrect predictions do not matter, or when the predictions are 100% accurate, they become highly problematic when the stakes are high and the predictions are not perfect, as in the criminal justice domain.

A range of scholars have sounded concerns regarding use of AI in areas in which important rights and public interests are at stake, particularly in the criminal justice setting. Andrea Roth has described concerns with trial by machine, if defendants

---

123 Ariel Bleicher, *Demystifying the Black Box that is AI*, Sci. Am. (Aug. 9, 2017), https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/ [https://perma.cc/4P4C-4GCD].

124 Will Knight, *The Dark Secret at the Heart of AI*, MIT Tech. Rev. (Apr. 11, 2017), https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/ [https://perma.cc/833R-5F4F].

125 Jarek Gryz & Marcin Rojszczak, *Black Box Algorithms and the Rights of Individuals: No Easy Solution to the "Explainability" Problem*, 10 Internet Pol'y Rev. (2021).

126 Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 Harv. J.L. & Tech. 890, 893 (2018).

127 *See* Rudin, *supra* note 34.

128 Vijay Pande, *Artificial Intelligence's 'Black Box' is Nothing to Fear*, N.Y. Times (Jan. 25, 2018), https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html [https://perma.cc/9EAQ-L23M].

cannot cross-examine AI evidence.[129] In addressing the uses of AI by the judiciary, Frank Pasquale states: "Explainability matters because the process of reason-giving is intrinsic to juridical determinations—not simply one modular characteristic jettisoned as anachronistic once automated prediction is sufficiently advanced."[130]

Regarding tort liability, legal scholars have asserted, for example, that: "The AI's thought process may be based on patterns that we as humans cannot perceive, which means understanding the AI may be akin to understanding another highly intelligent species—one with entirely different senses and powers of perception."[131] Others have noted that leading AI systems used by government are not transparent.[132] Regarding uses of AI in health care, a scholar noted, "the algorithms themselves are often too complex for their reasoning to be understood or even stated explicitly."[133] Again, we disagree where, even if the algorithms are complex, the outputs may often rely on quite simple and easy to understand factors.

Still others have argued that we should accept and try to work around the black box problem, for a range of reasons: since other types of post-hoc vetting may be possible, the benefits of black box AI may outweigh the costs, human judgment is not always understandable, and "just because we can't completely understand something doesn't mean we shouldn't trust it."[134] In so doing, however, they champion explainable AI approaches, and not interpretable or glass box AI.

A final group of scholars fear that black box AI makes it impossible to protect important rights, where the right to contest AI decisions and the right to an explanation of that decision,

---

129 *See* Roth, *supra* note 39, at 1300.

130 Frank Pasquale, *Toward A Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 OHIO ST. L.J. 1243, 1252 (2017).

131 Bathaee, *supra* note 126, at 893.

132 *See* DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANOFLORENTINO CUÉLLAR, ADMIN. CONF. U.S., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 7 (2020), https://law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf [https://perma.cc/2WVT-D5KD] ("When public officials deny benefits or make decisions affecting the public's rights, the law generally requires them to explain why. Yet many of the more advanced AI tools are not, by their structure, fully explainable.").

133 W. Nicholson Price II, *Artificial Intelligence in Health Care: Applications and Legal Issues*, 14 SCITECH LAW. 10, 10 (2017).

134 Robin C. Feldman, Ehrik Aldana & Kara Stein, *Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know*, 30 STAN. L.& POL'Y REV. 399, 401 (2019).

are "intertwined," and neither is meaningful without "opening the black box."[135] Without a clear understanding the feasibility of the glass box alternative, these fundamental problems may appear unsolvable. Thus, we view the glass box alternative as missing in prominent debates about how to regulate the use of AI, particularly in settings in which important rights are at stake, such as the criminal justice system.

B. The Glass Box Advantage

Data science problems can be grouped into two categories: those with tabular data (e.g., criminal history counts, age), and those with raw data (images, soundwaves, large bodies of text).[136] Neural networks are the best technique currently for raw data problems. But for tabular data, most modern methods are about equally accurate, including those that can produce very interpretable models. In other words, there does not appear to be any performance benefit from using complex models like neural networks for tabular data problems. Recidivism risk scoring, for instance, is a tabular data problem where black box models have not been shown to have an advantage over very small interpretable models.[137] As discussed earlier, for raw data problems such as computer vision, it is possible to design specialized neural networks that have a specialized notion of interpretability.[138]

In other words, AI need not be a black box to attain the accuracy of a black box. As one of us has put it simply: "Why Are We Using Black Box Models in AI When We Don't Need To?"[139] While a few early machine learning experimentalists noted this,[140] the arguments are subtle enough that they require clarification. The lack of a black box performance advantage has been shown to be true across fields, including

---

135 *See* Kaminsky & Urban, *supra* note 3, at 2047.

136 *See* Rudin, *supra* note 34, at 208.

137 *See* Jiaming Zeng, Berk Ustun & Cynthia Rudin, *Interpretable Classification Models for Recidivism Prediction*, 180 J. ROYAL STAT. SOC. 659, 659 (2017); Caroline Wang, Bin Han, Bhrij Patel & Cynthia Rudin, *In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction*, 39 J. QUANT. CRIM. 519, 519 (2022).

138 Chaofan Chen et al., *This Looks Like That: Deep Learning for Interpretable Image Recognition*, NEURIPS, 2019, at 3.

139 *See* Rudin & Radin, *supra* note 55, at 1.

140 Robert Holte, *Very Simple Classification Rules Perform Well on Most Commonly Used Datasets*, 11 MACHINE LEARNING 63 (1993).

computer vision,[141] recidivism risk scoring,[142] prediction of Type-2 diabetes,[143] and online marketing.[144]

Take computer vision as an example. Classifying objects is a complex task for humans. Although we sometimes do it unconsciously, we can also explain how we made a classification decision regarding an object, whether it was a species of bird or a brand of car or identifying a tumor in an X-ray scan.[145] In a 2019 study, computer scientists compared an interpretable model for classifying objects with non-interpretable counterparts.[146] They found that the interpretable system performed with the same accuracy as the black box systems.[147] Moreover, the system not only explained how it reached its results, but it provided visual justifications for it, by showing what features of a bird, for example, led it to conclude that it was a red-bellied woodpecker.[148] They found that the AI system "agrees with the way humans describe their own reasoning in classification," when people engage in such tasks (like identifying bird species).[149] A birdwatcher might trust the glass box AI decision more when seeing that the reason it identified the red-bellied woodpecker was that the AI focused on the bright red cap (and despite the name, an only faintly rusty belly). Thus, one can not only better understand how the AI reached a decision, but one also has more confidence that the system generally tracks human reasoning.

The stakes are higher when one turns from bird identification to risk assessments used in the criminal justice system to inform decisions such as whether to detain a person pretrial or reduce their sentence. Research has also shown that black box models do not perform any better in criminal law settings than simpler and interpretable models.[150] Indeed, criminal risk scoring systems that are "completely transparent and highly

---

141 *See* Chen et al., *supra* note 138.

142 *See* Jiaming Zeng, Berk Ustun & Cynthia Rudin, *Interpretable Classification Models for Recidivism Prediction*, J. Royal Stat. Soc., Mar. 26, 2015.

143 Narges Razavian et al., *Populationlevel Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors*, 3 Big Data 277, 277–87 (2015).

144 *See* Ritu Sharma, Arpit Kumar & Cindy Chuah, *Turning the Blackbox into A Glassbox: An Explainable Machine Learning Approach for Understanding Hospitality Customer*, 1 Int'l J. Info. Mgt Data Insights, Nov. 2021, at 9.

145 *See* Chen et al., *supra* note 138.

146 *See id.*

147 *See id.*

148 *See id.*

149 *Id.*

150 *See* Zeng, Ustun & Rudin, *supra* note 137. *See also* Wang, Han, Patel & Rudin, *supra* note 137 (showing that a set of glass box tools outperformed two

interpretable" have been found to be "just as accurate as the most powerful black-box machine learning models for many applications."[151] Thus: "In the criminal justice system, it has been repeatedly demonstrated . . . that complicated black box models for predicting future arrest are not any more accurate than very simple predictive models."[152]

There have been recent efforts to understand *why* interpretable models have the accuracy of black box models. One recent theory suggests that when the prediction problem is heavily influenced by randomness (e.g., whether someone will commit a crime within two years could depend on any number of circumstances and is a noisy process), there are many approximately-equally-predictive models, and in that case, it is likely that at least one of these models is interpretable.[153]

Not only does black box AI lack a performance advantage, but there are strong reasons to believe that it performs far more poorly than glass box alternatives. We have described how black box AI can lead to less accurate decision-making, because such models are harder to troubleshoot, validate, and use in practice. There is a second and deeper problem: errors may not come to light when concealed in a black box.[154]

As will be developed further in the next section, errors are common in criminal justice data, where police, clerks, judges, defense lawyers, and prosecutors are not primarily tasked with producing high quality and reliable data.[155] This is a fundamental problem in the criminal justice setting, where, as John Pepper, Carol Petrie, and Sean Sullivan have explained: "Errors are evidently pervasive, systematic, frequently related to behaviors and policies of interest, and unlikely to conform to convenient textbook assumptions."[156] Even basic typographical errors in the input to black box recidivism prediction models

---

leading criminal risk assessment instruments, the Public Safety Assessment and COMPAS, and providing a fairness assessments of these models).

151 *See* Zeng, Ustun & Rudin, *supra* note 137.

152 *See* Rudin & Radin, *supra* note 55, at 4.

153 Lesia Semenova, Cynthia Rudin & Ronald Parr, *On the Existence of Simpler Machine Learning Models*, ACM CON. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1827, 1827 (2022).

154 *See* Cynthia Rudin, Caroline Wang & Beau Coker, *The Age of Secrecy and Unfairness in Recidivism Prediction*, HARV. DATA SCI. REV., Winter 2020, at 2.

155 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 [https://perma.cc/6WE2-NNBX].

156 John Pepper, Carol Petrie & Sean Sullivan, *Measurement Error in Criminal Justice Data*, *in* HANDBOOK OF QUANTITATIVE CRIMINOLOGY, at Abstract (Alex R. Piquero

has led to catastrophic errors deeply affecting people's lives.[157] We turn next to three particular disadvantages of black box AI in criminal cases.

C. Three Challenges to Uses of AI in Criminal Justice

AI's use in the criminal system has raised a range of concerns, and in this section, we focus on scientific challenges, and not ethical or legal challenges, to which we turn in the next Part. There are three broad types of scientific challenges to the use of AI in general, and in the criminal system: (1) the problems of training and input data, or the data used to develop an AI system; (2) validation, or the efforts made to ensure that the system works as intended; and (3) interpretation and explanation, two different concerns that we distinguish and with which we refer, respectively, to the ability of users to know what the AI system actually relied on in making decisions, and the intelligibility of its outputs, which can include post hoc explanations.

In this section, we seek to bridge misunderstandings between the computer science and legal communities. First, as to data, while scientists appreciated that the quality of data matters deeply, they often fail to understand that much data in the criminal system is problematic. Second, lawyers appreciate that AI systems need to be validated, but often fail to understand that many uses of AI are by law enforcement and courts are poorly validated. Third, *both* communities fail to appreciate adequately the important distinction between interpretability and explainability and why that matters for AI in criminal justice.

1. *The Data Used to Develop Criminal Justice AI*

First, the usefulness of AI as a tool in part depends on what data we use to train and develop the AI system.[158] Inadequate or biased data distorts the AI system developed based on those inputs.[159] For that reason, Frank Pasquale has argued there should be duties of care to supply representative data when developing AI

& David Weisburd eds., 2010), https://doi.org/10.1007/978-0-387-77650-7_18 [https://perma.cc/B9CE-PQR7].

157 *See id.*

158 For an overview, see PEDRO DOMINGOS, THE MASTER ALGORITHM 7 (2015). *See also* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 693, 693 n.135 (2017).

159 *See* Frank Pasquale, *Datainformed Duties in AI Development*, 119 COLUM. L. REV. 1917, 1925–27 (2019) (providing an overview discussion).

systems.[160] AI will perform poorly if we supply it with incomplete, irrelevant, or biased data. AI systems are trained on large datasets, as described, which must be representative of the types of data we want the system to be able to analyze in the future.

For example, one might assume that bias would be minimized in detection of cats if one feeds the machine millions of photos of cats and a sufficient variety of other mammal images. However, if one only fed the machine Siamese cats, it might fail to identify tabby cats. Similarly, if wealthier people have more access to certain medical services, then AI may recommend medical support based on their past usage, and ignore others who may be in greater need of care.[161] Training an AI system on past criminal justice data raises substantial challenges due to the quality of that data.

As Justice Ruth Bader Ginsburg put it in *Herring v. United States*, although databases "form the nervous system of contemporary criminal justice operations," nevertheless, "[t]he risk of error stemming from these databases is not slim."[162] In *Herring*, the defendant was arrested based on a recalled warrant. The police database had not been updated to reflect that the warrant had been recalled months earlier. In part this was because there was "no electronic connection between the warrant database of the [Sheriff's Department] and that of the County Circuit Clerk's Office" even though they were located in the same building.[163]

Why are criminal justice databases often unreliable? Many things can go wrong with data. Data may be incomplete or missing. Data may be recorded differently at different institutions and thus hard to merge together. There may be systematic biases in how data is recorded. Data may be overwritten or lost. There may be data entry errors. All of these problems and more are exemplified in the criminal justice setting.

Data problems occur even for the most basic and critical types of data.[164] Policing is a highly localized and fragmented system, and information on outcomes, such as arrests, jail detention,

---

160 *Id.* at 1927–28 ("Both lawmakers and policymakers should hold users of such data sets responsible for making predictable errors based on defective data sets, particularly if they fail to disclose the limitations of the data used.").

161 Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE, Oct. 25, 2019 at 1.

162 555 U.S. 135, 155 (2009) (Ginsburg, J., dissenting).

163 *Id.* at 154.

164 For an overview, see JOHN V. PEPPER & CAROL V. PETRIE, NAT'L RSCH. COUNCIL, MEASUREMENT PROBLEMS IN CRIMINAL JUSTICE RESEARCH (2004).

sentencing, and incarceration, may be far more lacking.[165] This is especially true when organizations such as the FBI attempt to aggregate local data into nationwide databases.[166] To provide just one example, "there are no nationally representative data available on the numbers of misdemeanor arrests and convictions, let alone data about pretrial detention rates, bail, or sentencing."[167] Law enforcement agencies may report data inconsistently and incompletely, even to flagship federal efforts, such as FBI crime reporting.[168] Indeed, in 2021, during the transition from the Uniform Crime Reporting ("UCR") program to a new more detailed reporting system, the National Incident-Based Reporting System ("NIBRS"), about 40% of law enforcement agencies did not report data to the FBI.[169]

Further, many crimes go unreported, so what law enforcement does not know is substantial.[170] Even when crimes are reported, whether an activity meets a particular crime definition may be discretionary or subject to interpretation.[171] In general, criminal behavior is not only uncommon and hard to detect, but inherently involves hard-to-predict actions and "noise." In such common real-world situations, as we will discuss, there is evidence that simpler models may be more accurate than complex models.[172]

---

165 *See id.* at 2–3; *see also* James P. Lynch & John P. Jarvis, *Missing Data and Imputation in the Uniform Crime Reports and the Effects on National Estimates*, 24 J. CONTEMP. CRIM. JUST. 69, 69 (2008).

166 *See* PEPPER & PETRIE, *supra* note 164, at 2 ("Although these data collection systems do many things right, they are, like any such system, beset with the methodological problems of surveys in general as well as particular problems associated with measuring illicit, deviant, and deleterious activities. Such problems include nonreporting and false reporting, nonstandard definitions of events, difficulties associated with asking sensitive questions, sampling problems such as coverage and nonresponse, and an array of other factors involved in conducting surveys of individuals and implementing official data reporting systems.").

167 Paul Heaton, Sandra Mayson & Megan Stevenson, *The Downstream Consequences of Misdemeanor Pretrial Detention*, 69 STAN. L. REV. 711, 732 (2017).

168 Reporting to the FBI, for example, is voluntary, and as a result, missing data is commonly and issue. Michael D. Maltz, *Bridging Gaps in Police Crime Data*, BUREAU JUST. STAT. (1999), https://bjs.ojp.gov/content/pub/pdf/bgpcd.pdf [https://perma.cc/753H-C43A].

169 Weihua Li, *What Can FBI Data Say About Crime in 2021? It's Too Unreliable to Tell*, MARSHALL PROJECT (June 14, 2022), https://www.themarshallproject.org/2022/06/14/what-did-fbi-data-say-about-crime-in-2021-it-s-too-unreliable-to-tell [https://perma.cc/E8SF-V7X8].

170 *See* PEPPER & PETRIE, *supra* note 164, at 2.

171 *Id.* at 2.

172 Semenova, Rudin & Parr, *supra* note 153.

A range of other types of criminal justice data may be unavailable, or incompletely collected. In general, we lack much of the information that we need to evaluate policing.[173] For example, data on more informal interactions, like police stops, is not consistently reported or collected.[174] Police reports may include more detailed information than the basic information available from police agencies or courts that report arrest and charging statistics, but reports often are non-public, particularly as to pending cases and for individuals facing charges but not convicted.[175] Data on police use of force, and even use of deadly force, is inconsistent and incompletely collected.[176] Behavioral health data is still more lacking, despite large percentages of arrestees that have behavioral health needs.[177] Many police agencies do not collect information on police misconduct lawsuits.[178]

Turning from police data to criminal court data, we observe the same types of challenges. Outcomes in criminal cases reflect a range of subjective and discretionary decisions by various actors, including pretrial services and other social workers, prosecutors, defense lawyers, judges and jurors. Post-arrest outcomes in court often depend on negotiations between counsel, where most cases are resolved through plea bargaining, and many cases are also dismissed, while those proceeding to trial rely on judgments of jurors and judges.

---

173 *See, e.g.*, Rachel Harmon, *Why Do We (Still) Lack Data on Policing?*, 96 MARQ. L. REV. 1119, 1131–32 (2013).

174 *See It's Time to Start Collecting Stop Data: A Case For Comprehensive Statewide Legislation*, POLICING PROJECT (Sept. 30, 2019), https://www.policingproject.org/news-main/2019/9/27/its-time-to-start-collecting-stop-data-a-case-for-comprehensive-statewide-legislation [https://perma.cc/D8Q9-BJT5] ("[S]top data collection laws, even when they do exist, are far from perfect. Many don't cover both pedestrian and traffic stops, some exempt agencies from making their data public, and some contain no enforcement mechanism to ensure departments are complying. In other cases, the data itself is simply so incomplete as to be practically useless.").

175 C. Dominik Güss, Ma. Teresa Tuason & Alicia Devine, *Problems With Police Reports as Data Sources: A Researchers' Perspective*, 11 FRONT PSYCH., Oct. 22, 2020, at 1–3.

176 It has been journalists that have attempted to systematically collect data on police use of deadly force. *Fatal Force*, WASH. POST (Sept. 23, 2022), https://www.washingtonpost.com/graphics/investigations/police-shootings-database/ [https://perma.cc/J4YP-QDYW].

177 *See* JENNIFER BRONSON & MARCUS BERZOFSKY, BUREAU JUST. STAT., DEP'T JUST., INDICATORS OF MENTAL HEALTH PROBLEMS REPORTED BY PRISONERS AND JAIL INMATES, 2011–12, (2017), https://bjs.ojp.gov/content/pub/pdf/imhprpji1112.pdf [https://perma.cc/W9QJ-ARWK].

178 *See, e.g.*, Joanna C. Schwartz, *Myths and Mechanics of Deterrence: The Role of Lawsuits in Law Enforcement Decisionmaking*, 57 UCLA L. REV. 1023, 1045–52 (2010).

The plea-bargaining process is typically not documented, except for the final result, and little data exists on the process.[179] Basic case information and sentencing data may be highly incomplete as well. As just one example, the Virginia Sentencing Commission noted in 2021 that:

- 45% of all cases it examined were missing the defendant's gender;
- 35% of all cases missed the defendant's race;
- 68% of larceny cases were missing the value of the stolen items;
- 49% of drug cases were missing the type of drug; and
- 37% of assault cases were missing the description of the victim's injury.[180]

Criminal history information plays a crucial role in a range of decisions, including employment and sentencing, and yet real quality problems have long been documented regarding these basic criminal records.[181] Even whether an arrest resulted in a disposition, like a conviction, may not be automatically recorded by courts, and may depend on prosecutors' care in reporting what occurred in a criminal case.[182] Further, a crime in one jurisdiction is aggregated at the state and federal level, but may not be updated or may otherwise be incomplete.[183]

When criminal justice data is reported, there also may be basic errors in inputting information. People may not accurately report their name, age, or actions to law enforcement and law enforcement may not accurately record information in their police reports.[184] It is not the job of a police officer or court clerk to be a trained data-entry professional, and unfortunately, many still often record information by hand and rely on incomplete memory: "the timeworn practice of officers

---

179 *See, e.g.*, Brandon L. Garrett et al., *Open Prosecution*, 75 STAN. L. REV. 1365 (2023).

180 Va. Sent'g Comm'n, *The Guidelines Messenger* (June 2022), http://www.vcsc.virginia.gov/Newsletters/VCSC%20Newsletter%20Spring%202022%20Final.pdf [https://perma.cc/Z3AC-3U9Q].

181 *See, e.g.*, PETER M. BRIEN, IMPROVING ACCESS TO AND INTEGRITY OF CRIMINAL HISTORY RECORDS 7 (2005).

182 PETER BRIEN, DEP'T. JUST., REPORTING BY PROSECUTORS' OFFICES TO REPOSITORIES OF CRIMINAL HISTORY RECORDS 1 (2005) (finding that less than half of state prosecutors responding to a survey indicated they regularly submitted final disposition information for criminal history records).

183 ROBERT R. BELAIR & PAUL L. WOODWARD, USE AND MANAGEMENT OF CRIMINAL HISTORY RECORD INFORMATION: A COMPREHENSIVE REPORT 30 (1993).

184 For an overview of these challenges in the jail data setting, see William E. Crozier, Brandon L. Garrett & Arvind Krishnamurthy, *The Transparency of Jail Data*, 55 WAKE FOREST L. REV. 821, 848–49 (2020).

writing notes by hand and then typing reports hours later is fraught with potentially serious downsides."[185] Sometimes, data is missing for random reasons, but it can also be missing not-at-random (called "MNAR").[186] Such data may be submitted by many different agencies, and it may be placed in one of many data files that cannot be easily integrated into a single database. For instance, information about a single individual may be scattered in different datasets that all encode the same information differently, or agencies may collect different information about individuals, so that we cannot directly compare individuals across datasets.[187] These data entry failures can magnify in their consequences when consolidated in larger databases.[188]

Further, once data is entered, quality controls from criminal justice organizations may be lacking, which can have serious consequences for individuals if erroneous information about their past criminal history or identity is used in their present case.[189] Indeed, the U.S. Supreme Court has held that lack of police due diligence in relying on inaccurate database information to make arrests does not raise Fourth Amendment concerns.[190] Without constitutional or other legal incentives to main such data accurately, it often is not.

Thus, criminal justice data can be inaccurate or highly biased, not just because of quality control issues, but also because it reflects so many different types of discretionary decisions by different actors.[191] Researchers have made great efforts to use random assignment to judges and other quasi-experiments to try

---

185 James Careless, *Consequences of Inaccuracy in Reporting and How to Avoid Errors*, POLICE1, (July 17, 2019), https://www.police1.com/sponsored-article/articles/consequences-of-inaccuracy-in-reporting-and-how-to-avoid-errors-DM55MqqHrSmVuVDk/ [https://perma.cc/XP25-UL2G].

186 Nicholas Blasco, *Missing Data in Criminology and Criminal Justice*, *in* THE ENCYCLOPEDIA OF RESEARCH METHODS IN CRIMINOLOGY AND CRIMINAL JUSTICE 503, 504 (J.C. Barnes & David R. Forde, eds., 2021).

187 *Id.*

188 *See* Christopher Slobogin, *Government Data Mining and the Fourth Amendment*, 75 U. CHI. L. REV. 317, 323–27 (2008).

189 *See, e.g.*, Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1, 17–18 (2005).

190 *See* Herring v. United States, 555 U.S. 135, 146 (2009) (holding that relief under the Fourth Amendment is possible only "[i]f the police have been shown to be reckless in maintaining a warrant system, or to have knowingly made false entries to lay the groundwork for future false arrests").

191 Pepper, Petrie & Sullivan, *supra* note 156, at 4 (noting, as an example, that "police discretion in whether and how to record incidents may lead to substantial errors in the measurement of reported crimes").

to study effects of criminal decision-making, precisely because so much discretion and bias is built into the system.[192] Yet, many non-criminal practitioners do not appreciate how incomplete and error-prone such basic criminal justice data can be. Those may include developers of AI systems not sufficiently familiar with the limitations of the underlying data. If the criminal justice system data used to train AI reflects errors, discretion and biases of human decision-makers, then the outputs may similarly contain those flaws.[193] As Sandra Mayson has put it, there is a "garbage in, garbage out" problem that can result in "bias in, bias out" when relying on criminal enforcement data.[194]

Finally, when an AI system is relying on past data to form predictions about a present-moment case or situation, there is a separate data quality problem, which is that the data for the present case may also be lacking. Something as basic as the wrong address information can and does lead to an erroneous arrest.[195] Accuracy is one of the basic principles of an AI system that examines personal data.[196] In a "black box" or proprietary system, people do not know what data is relied on or how it is used.[197] There is no way, in a particular case, to assess whether data is erroneous if it is a black box. For all of the reasons just described, the state of criminal justice data collection makes a black box system particularly concerning.

2. *The Validation of Criminal Justice AI*

Second, the usefulness of AI as a tool depends on how well the system recognizes patterns in the data. We have discussed

---

192 *See, e.g.*, Will Dobbie, Jacob Goldin & Crystal S. Yang, *The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges*, 108 AM. ECON. REV. 201 (2018); Arpit Gupta, Christopher Hansman & Ethan Frenchman, *The Heavy Costs of High Bail: Evidence from Judge Randomization*, 45 J. LEGAL STUD. 471, 472–73 (2016).

193 *See* Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger & Suresh Venkatasubramanian, *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACH. LEARNING RSCH. 160 (2018).

194 *See* Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2224 (2019).

195 Wayne A. Logan & Andrew Guthrie Ferguson, *Policing Criminal Justice Data*, 101 MINN. L. REV. 541, 562–67 (2016).

196 Reuben Binns & Valeria Gallo, *Accuracy of AI System Outputs and Performance Measures*, INFO. COMM'R'S OFF. (May 2, 2019), https://web.archive.org/web/20201120063121/https://ico.org.uk/about-the-ico/news-and-events/ai-blog-accuracy-of-ai-system-outputs-and-performance-measures/ [https://perma.cc/6G6G-FDYF].

197 State v. Loomis, 881 N.W.2d 749, 769 (Wis. 2016) ("[T]he proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined.").

already how a predictive model should be evaluated using test data, and that evaluation should be replicable by others. Sometimes, the most basic types of evaluations have not occurred, and in criminal justice settings, those evaluations are not often required.

As one example, the statistics used to calculate the probative value of DNA searches in the federally-administered DNA databank system included simple mathematical errors that led to faulty calculations used for over fifteen years. Since the FBI had not made its statistics public, the errors could not be detected by lawyers or researchers.[198]

In the area of facial recognition technology, the FBI operates a system of facial recognition called FACE (Facial Analysis Comparison and Evaluation), but it has been unwilling to provide evidence of any effort to validate how accurate it is.[199] This lack of validation has been the subject of GAO inquiries, which has called on the FBI to conduct testing of the accuracy of the system.[200] The FBI has responded that under its policy, "photos cannot serve as the sole basis for law enforcement action," and that ongoing work is being done to improve the accuracy of the system.[201]

A few courts have already emphasized that the evidentiary concerns regarding reliability do not apply when the evidence is investigatory and not introduced in court.[202] Yet, even if AI

198 Spencer S. Hsu, *FBI Notifies Crime Labs of Errors in DNA Match Calculations Since 1999*, WASH. POST (May 29, 2015), https://www.washingtonpost.com/local/crime/fbi-notifies-crime-labs-of-errors-used-in-dna-match-calculations-since-1999/2015/05/29/f04234fc-0591-11e5-8bda-c7b4e9a8f7ac_story.html [https://perma.cc/KW9S-K9HF].

199 *Face Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains*, U.S. GOV'T ACCOUNTABILITY OFF. (June 4, 2019), https://www.gao.gov/products/gao-19-579t [https://perma.cc/QT34-RGWP] ("First, GAO found that the FBI conducted limited assessments of the accuracy of face recognition searches prior to accepting and deploying its face recognition system . . . Second, GAO found that FBI had not assessed the accuracy of face recognition systems operated by external partners . . . The FBI has not taken action to address these recommendations.").

200 *Id.*

201 *See* Del Greco Statement on Facial Recognition, *supra* note 83, at 3–4; *see also* P. Jonathon Phillips, Amy N. Yates, Ying Hu & Alice J. O'Toole, *Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms*, 115 PNAS 6171, 6174 (2018).

202 *See, e.g.*, Geiger v. State, 174 A.3d 954, 965 (Md. Ct. Spec. App. 2017) (emphasizing, "[r]eliability does not matter, however, because the computerized identification is not ultimately evidence in court. It is simply a guide to put the investigator on the right track").

is not formally admitted as evidence, that does not eliminate concerns about whether it is accurate. To be investigated as a criminal suspect is itself a concern, both regarding liberty and public safety. We should ask whether AI should be used, even if limited to preliminary identification purposes, which could then lead to a criminal prosecution, if we do not know how reliable the technology is.[203] If the FBI had a system paying millions of dollars to informants, whose identity was not disclosed, the public and judges would want to know if these confidential informants were reliable or not, even if their predictions were not admitted as evidence, and particularly if these leads led to criminal investigations and arrests. Investigation systems must be validated. As described earlier, in the *Arteaga* case, the New Jersey appellate court recognized as much and ordered full discovery regarding facial recognition if the prosecution sought to use eyewitness evidence in the criminal case.[204]

3. *Interpretation and Explanation of Criminal Justice AI*

Third, the distinction between interpretability and explainability has not been made clearly by many in the computer science community. This has led to broader confusion concerning the terms "open" or "transparent" or "interpretable" or "explainable" in AI. There are different meanings attached to the term "open" AI, and, unfortunately, despite their "open" branding, many uses of AI still lack interpretability. It is crucial to be precise about definitions of AI concepts. Both the theoretical computer science and the legal communities need to be consistent in the use of these definitions.

By "interpretable" AI we refer to predictive models where humans can trace the decisions step by step.[205] In contrast, by "explainable," we refer to efforts to provide post hoc descriptions of models, which could be black box models.[206] These general explanations do not let us determine how individual decisions were made. Thus, only an interpretable AI system is a glass

---

203 *See* Clare Garvie, Alvaro Bedoya & Jonathan Frankle, *The Perpetual Lineup: Unregulated Police Face Recognition in America*, GEO. L. CTR. ON PRIV. & TECH. (Oct. 18, 2016), https://www.perpetuallineup.org [https://perma.cc/BX2S-3L58].

204 *Id.*

205 Cynthia Rudin et al., *Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges*, 16 STAT. SURVS., 2022, at 2–3.

206 Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACH. INTEL. 206, 206 (2019).

box system. In an interpretable system, a person can see how the AI system works and what it relies upon in a particular instance. The predictive model is transparent.

Therefore, the preferred approach, or "interpretable machine learning," does not just explain the predictive model, but rather makes it visible to the user. The system is a "glass box" and not a "black box." It provides information regarding the model, the factors used to provide a result, and how those factors were in fact combined to provide a specific result. Such "glass box" approaches set out what matters to the AI model when it makes its predictions in ways that people can comprehend—and challenge, if necessary. For cases involving complex "raw" data-like images, the algorithm can still show its work in readily understandable ways. For instance, there are interpretable neural networks that show their calculations by highlighting not only what pixels they used, but how they compared the relevant parts of a current image to the relevant parts of training images in order to make their prediction.[207] The underlying models, or algorithms, used by the AI may be extremely complex. However, the factors that the model ultimately relies upon may be quite simple and understandable.

We note that there is an additional challenge that even if a predictive model is transparent and interpretable, the information should be conveyed in a way that is accessible to the types of people who are relying upon it (such as lawyers or judges, who likely are not computer scientists). Fortunately, some models can be so concise that they appear as if they could have been created by a human—taking the form of, for instance, a simple scorecard.[208] However, they can be as powerful as the most powerful black box models.

To explain briefly how interpretable AI systems present their results, a model may reflect a series of factors found valuable to predict a type of outcome. Some simple risk assessment instruments are depicted in a simple one- or two-page worksheet that assigns points based on certain factors, like the person's age, prior offenses, and current offense. A social worker or judge can easily see how much weight each factor has and why a person is deemed high or low risk, even if they may not understand how the data was used to generate the scheme or how accurate it is.

---

207 Chen et al., *supra* note 138.

208 Jiachang Liu, Chudi Zhong, Boxuan Li, Margo Seltzer & Cynthia Rudin, *FasterRisk: Fast and Accurate Interpretable Risk Scores*, NeurIPS, 2022, at 2.

For example, in New Mexico, the following formula was used to score the risk of new criminal activity:

> New Criminal Activity maximum total weight = 13 points
> Age at current arrest: 23 or older = 0;
> 22 or less = 2
> Pending charge at the time of the offense: Yes = 3
> Prior misdemeanor conviction: Yes = 1
> Prior felony conviction: Yes = 1
> Prior violent conviction: 1 or 2 = 1
> 3 or more = 2
> Prior failure to appear pretrial in past 2 years: 1 = 1
> 2 or more = 2
> Prior sentence to incarceration: Yes = 2[209]

This checklist is very simple. It is also potentially counterintuitive to the judicial officers using it, who might place far more weight on the current crime of arrest (which is not part of the formula), rather than other factors, such as a prior felony conviction, or a prior sentence of incarceration.

Such simple models are not what AI is known for—the stereotype is that AI must be extremely complex.[210] Yet, the contribution of AI can be in simplifying large quantities of data to produce a smaller set of useful variables. Modern interpretable machine learning techniques heavily optimize the choice of variables and how variables are combined to produce a predictive model. Computers are much faster than they were decades ago, which means that the computationally hard problems of choosing optimal variables and combining them effectively can now be solved for most reasonably-sized datasets. Even for challenging benchmark datasets, interpretable machine learning methods have been able to match the performance of black box methods.[211]

Even neural networks, which can be quite complex, need not be presented as complete black boxes. For computer vision, for instance, there are some neural networks that are designed to point out how visual information is combined from similar

---

209 *See* LAURA & JOHN ARNOLD FOUND., PUBLIC SAFETY ASSESSMENT, RISK FACTORS AND FORMULA 3, https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/PSA-Risk-Factors-and-Formula.pdf [https://perma.cc/KK9W-XLMD]; Ryan Boetel, *Courts to Implement New Risk Assessment Tool*, ALBEQUERQUE J. (May 31, 2017), https://web.archive.org/web/20220702133918/https://www.abqjournal.com/1011380/courts-to-implement-new-risk-assessment-tool-for-suspects.html [https://perma.cc/C3XU-38YV].

210 Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 88 (2014).

211 Rudin et al., *supra* note 205, at 26.

cases to produce a prediction for a current case, as discussed above.[212] The outcomes are highly interpretable; the program shows, for example, which portion of a bird's wing or crest was relied upon to connect an image of a bird in the wild with the exemplar photograph of, say, a Northern Cardinal.[213] The program can explain that "it is looking at *this* region of input because *this* region is similar to that prototypical example."[214]

For high-stakes criminal cases, as discussed next, interpretability is important. However, in contrast to interpretable AI, explainable AI only provides a partial "explanation" for how the AI might have performed. It does not actually detail what the AI system did and what the predictive model in fact relied upon. Thus, the less desirable type of approach, explainable AI, develops black box models from data, and then queries the black box, in effect, to provide a speculative account of what the algorithm may have done.[215] These explanations do not open the black box and may actually further obscure its workings.[216] In effect, this approach uses proxies to guess what the AI may have done. Thus, explainable AI "pokes at" the black box, whereas interpretable AI replaces it with something we can understand.

Explainable AI might be better than shrouding the AI in complete secrecy if the only alternative were a black box. Perhaps that is why many in the AI community have emphasized that explainable AI (or XAI) is a comparatively good thing. Similarly, legal scholars have proposed laws requiring that AI be explainable[217] and proposed that judges should demand explainable AI.[218] To be sure, some of those statements are generic calls for less black box AI and do not distinguish

---

212 Chen et al., *supra* note 138, at 2.

213 *Id.* at 9.

214 *Id.* at 8.

215 *See* Leilani H. Gilpin et al., *Explaining Explanations: An Overview of Interpretability of Machine Learning*, arXiv, May 31, 2018.

216 Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 Harv. J.L. & Tech. 841, 850 (2018).

217 Andrew Tutt, *An FDA for Algorithms*, 69 Admin. L. Rev. 83, 109 (2017) (proposing federal statute requiring explainability: "If explainability can be built into algorithmic design, the presence of a federal standard could nudge companies developing machinelearning algorithms into incorporating explainability from the outset").

218 Deeks, *supra* note 121, at 1830 ("This Essay argues that judges will confront a variety of cases in which they should demand explanations for algorithmic decisions, recommendations, or predictions.").

between explainable and interpretable AI. Many use the term "explainable" to refer to a broad range of approaches, including to refer to what we call interpretability.[219] The concept is what matters most, not the terms chosen. We do agree that with explainable AI the user might better understand what might have been done than if no explanation was provided at all.[220]

However, there are serious problems with explainable (as opposed to interpretable) AI. Explanations are not always faithful to the model's calculations. In other words, explanations can often be wrong. Many explainability methods regularly disagree with each other, illustrating why we cannot trust any of them—we have no way of knowing which one(s) are correct (if any actually are).[221] In computer vision, where we aim to explain why an image was classified as containing certain content (say, a person), explanations tend to often be the same regardless of whether the prediction was right or wrong, where the explanation consisted only of highlighting pixels where there were edges in the image. This example also illustrates how explanations can also be incomplete: even if we know which pieces of information the AI is using (which pixels from the image, for instance), if we do not know how that information is being combined to form the prediction, the explanation may not be useful. Explanations also tend to be wrong on more difficult decisions (cases close to the decision boundary), which are precisely the cases where we need explanations to be correct. Explanations may not even be needed on easier decisions, because the decision may be obvious anyway.

Explanations (even wrong ones) also may lend more authority to the black box, justifying its use in the first place.[222] We view post-hoc explanations as misleading and inappropriate in high-stakes settings, like criminal justice. That is why we view

---

219 *Id.* at 1836–37. For additional helpful discussion of modelcentric interpretability, versus local interpretability of decisions made in particular instances, see *id.* at 1835–37.

220 For example, the Department of Defense explains the need for explainable AI as follows, "Explainable AI—especially explainable machine learning—will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners." *See* Matt Turek, *Explainable Artificial Intelligence*, DARPA (July 2, 2023), https://www.darpa.mil/program/explainable-artificial-intelligence [https://perma.cc/WTW2-24GM].

221 Tessa Han, Suraj Srinivas & Hima Lakkaraju, *Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations*, NEURIPS, 2022, at 3.

222 *See* Rudin & Radin, *supra* note 55.

the distinction between explainable and interpretable models as an important one. Unfortunately, many explainable AI proposals were made without testing whether interpretable AI was feasible or might perform as well as black box AI. As we will address in the next Part, constitutional and legal programs are exacerbated by black box AI as well.

III
Glass Box Constitutional Criminal Procedure

We as a society now understand that AI can affect people's lives important ways.[223] These include applications in our criminal system, where AI is already used in a host of criminal investigation, pretrial, and sentencing-related settings.[224] The particular use of AI is important and can greatly alter the accuracy, privacy, and fairness interests at stake, as well as the fair trial rights involved. In the first section that follows, we describe how any use of AI that results in evidence introduced during a criminal investigation, or in court (or perhaps to obtain a search warrant), will generally raise far more constitutional concerns than a use of AI that is not used to prosecute a person. Second, we discuss equal protection and the need for a glass box to safeguard against discrimination by AI in criminal justice. Third, we turn to the rules regarding admissibility of expert evidence, which also demand glass box AI. Fourth, we turn to the role of human consumers of AI evidence. There are deeply inaccurate and biased uses of human discretion in the criminal system, and combatting such bias raises challenges that legal systems have grappled with for decades. Yet, we cannot know whether AI improves on the accuracy of human decision-making if it is a black box system. This Part concludes that we should require, judicially or through legislation, interpretable AI in criminal cases to safeguard constitutional rights, prevent use of unreliable evidence, avoid discrimination, and ensure public safety.

---

223 Bryan Casey, Ashkan Farhand & Roland Vogt, *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, 34 Berkeley Tech. L.J. 145, 148 (2019) ("In recent years, however, society's deferential attitude toward algorithmic objectivity has begun to wane—thanks, in no small part, to a flurry of influential publications examining bias . . . .").

224 *See id.* at 149 ("Particularly in the last five years, numerous studies across multiple industry sectors and social domains have revealed the potential for algorithmic systems to produce disparate [realworld] impacts on vulnerable groups.").

A. Glass Box Fair Trial Rights

Glass box AI serves to protect a range of constitutional rights, including and especially during criminal adjudication. The most relevant constitutional provisions are the Due Process Clauses of the Fifth and Fourteenth Amendments and the Sixth Amendment.[225] We argue that due process requires glass box AI in criminal adjudication. The role of due process is less demanding during criminal investigations, proceedings used to determine bail,[226] and sentencing,[227] but we argue that there is so little justification for use of a black box that a glass box should be required in such settings as well. Before and during a trial, however, the due process protections in criminal cases include assurances that all material, exculpatory and impeaching evidence of innocence be disclosed to defendants, under *Brady v. Maryland* and its progeny.[228]

This disclosure right is especially important because the Due Process Clause ensures "against conviction except upon proof beyond a reasonable doubt of every fact necessary to constitute the crime with which [a defendant] is charged."[229] The *Brady* obligation requires that prosecutors disclose exculpatory evidence even in the absence of a request from the accused, including impeachment evidence and evidence in the possession of other government actors, including the police.[230] This is because the prosecutor serves as "'the representative . . . of a sovereignty . . . whose interest . . . in a criminal prosecution is not that it shall win a case, but that justice shall be done."[231] Recent federal rulings have increasingly

---

225 Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 10–16 (2014).

226 For an overview, see Brandon L. Garrett, *Models of Bail Reform*, 74 FLA. L. REV. 879 (2022).

227 *See, e.g.*, FED. R. EVID. 1101 (stating that the rules of evidence are not applicable during sentencing).

228 373 U.S. 83, 87 (1963). Regarding questions whether machinegenerated results are themselves "testimonial" under the Sixth Amendment Confrontation Clause, see Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 2039–48 (2017).

229 *In re* Winship, 397 U.S. 358, 364 (1970).

230 Giglio v. United States, 405 U.S. 150, 153–54 (1972) (holding that the duty includes impeachment evidence); United States v. Agurs, 427 U.S. 97, 107 (1976) (finding no defense discovery request is required to preserve *Brady* rights); Kyles v. Whitley, 514 U.S. 419, 437 (1995) (holding that prosecutors are responsible for obtaining and disclosing exculpatory evidence that law enforcement possesses).

231 *Kyles*, 514 U.S. at 439 (quoting Berger v. United States, 295 U.S. 78, 88 (1935)).

focused on the obligations to disclose forensic evidence in a range of settings.[232]

Thus, if prosecutors introduce an expert presenting the results of an AI analysis, the defense should be entitled to discovery not just regarding the ultimate result of that analysis, but also to information that could permit the defense to impeach the expert or challenge the AI calculations. No such evidence will exist, however, unless it is a glass box AI system. The Advisory Committee to the Federal Rules of Criminal Procedure notes Rule 16 is intended to require disclosure of scientific results and tests: "[T]he requirement that the government disclose documents and tangible objects 'material to the preparation of his defense' underscores the importance of disclosure of evidence favorable to the defendant."[233] It is standard practice to disclose underlying documentation of forensic experts in federal cases, although state practices are quite uneven.[234] There is a pressing need to ensure use of glass box AI, because otherwise disclosures are far less readily made in discovery.

Thus, in *State v. Chun*, the New Jersey Supreme Court ordered Draeger Safety Diagnostics Inc., which produces the Alcotest 7110 breathalyzer, to disclose its proprietary source code for independent review.[235] As a result, outside analysis revealed significant source code errors.[236] Similarly, a New Jersey appellate court ordered, in discovery, that if the state sought to use probabilistic DNA software, then the "defendant is entitled to access, under an appropriate protective order, to the software's source code and supporting software development and related documentation."[237] Such orders are important, but do not provide the same access as if the defense, with glass box AI, can observe how the system reached conclusions in a particular case.

---

232 For an overview, see Brandon L. Garrett, *Constitutional Regulation of Forensic Evidence*, 73 WASH. & LEE L. REV. 1147 (2016).

233 FED. R. CRIM. P. 16 advisory committee's note to 1974 amendment.

234 U.S. Dep't of Just., U.S. Att'ys' Manual § 95.003 (2017) ("[I]f requested by the defense, the prosecutor should provide the defense with a copy of, or access to, the laboratory or forensic expert's 'case file,' either in electronic or hardcopy form."). However, for local practices and rulings denying the defense such access to bench notes, see Paul C. Giannelli, *Criminal Discovery, Scientific Evidence, and DNA*, 44 VAND. L. REV. 791, 809–10 (1991) (discussing limitations on discovery of reports and bench notes); *see also* Paul C. Giannelli, *Bench Notes & Lab Reports*, 22 CRIM. JUST. 50, 50–51 (2007).

235 943 A.2d 114, 123 (N.J. 2008).

236 *Id.* at 137, 160.

237 State v. Pickett, 246 A.3d 279, 284 (N.J. Super. Ct. App. Div. 2021).

Relatedly, glass box AI is needed for defendants to benefit from effective assistance of counsel, protected by the Sixth Amendment and the Due Process Clauses. The U.S. Supreme Court has repeatedly emphasized obligations of the defense to adequately challenge forensic evidence: "Criminal cases will arise where the only reasonable and available defense strategy requires consultation with experts or introduction of expert evidence."[238]

Similarly, the Supreme Court's Sixth Amendment Confrontation Clause rulings have emphasized the defense's right to adequately confront adverse witnesses, including forensic witnesses, in court.[239] Confrontation simply cannot occur if conclusions were reached by AI, and in a matter that cannot be disclosed to the defense, or by an expert witness for the prosecution, because the AI is not interpretable. Thus, Andrea Roth has described the Sixth Amendment concern with any use of AI that does not permit defendants to examine or cross-examine AI evidence presented by a government expert.[240] Only interpretable AI can be meaningfully disclosed by a human expert who can then be subject to cross-examination regarding the factors relied upon by the AI. As noted, in the case of risk scoring, there has been much evidence of typographical errors and other types of data errors may influence scores.[241] For example, in the case of COMPAS in Broward County, Florida, apparently the wrong scoring model was used for years: the COMPAS parole score was used to determine pretrial risk, rather than a COMPAS pretrial score designed for this purpose.[242] In our view, defense counsel cannot meaningfully defend a person regarding an AI result without glass box AI.[243]

---

238 Hinton v. Alabama, 571 U.S. 263, 273 (2014) (quoting Harrington v. Richter, 562 U.S. 86, 106 (2011)).

239 MelendezDiaz v. Massachusetts, 557 U.S. 305, 313 (2009); Bullcoming v. New Mexico, 564 U.S. 647, 652 (2011).

240 *See* Roth, *supra* note 39, at 1300–01.

241 Rudin, Wang & Coker, *supra* note 154.

242 Anthony W. Flores, Kristin Bechtel & Christopher T. Lowenkamp, *False Positives, False Negatives, And False Analyses: A Rejoinder To "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.,"* 80 FED. PROB. 38, 40 (2016); Jeff Larson, Surya Mattu, Lauren Kirchner & Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA, (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perma.cc/6ETG-QDCU]; Eugenie Jackson & Christina Mendoza, *Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not*, HARV. DATA SCI. REV., Winter 2020, at 8.

243 *See* Strickland v. Washington, 466 U.S. 668, 687, 694 (1984) (holding that a fair trial under the Sixth Amendment requires defense counsel to provide

We emphasize the importance of affirmatively adopting policies to ensure these due process and Sixth Amendment constitutional rights are safeguarded, because in practice, many have long been poorly enforced. Discovery in criminal cases can be typically quite limited, making it difficult for defendants to be aware of exculpatory evidence.[244] Even if discovery is required, discovery violations are difficult to detect.[245] Nor are evidentiary rights clearly defined in pretrial settings, during plea bargaining, or in sentencing proceedings in many jurisdictions.[246] A criminal defendant may not be aware that AI was used to generate leads or evidence.

Existing judicial rulings regarding the use of AI and defense access have been quite mixed. A few judges have begun to raise concerns regarding black box use of AI in criminal cases, but unfortunately most have not granted relief when defendants are denied access to information about AI evidence used against them. In *People v. Collins*, a New York trial judge explained, regarding a government program called the Forensic Statistical Tool ("FST") used to interpret complex DNA mixtures, that:

> [T]he fact that FST software is not open to the public, or to defense counsel, is the basis of a more general objection. This court understands the city's desire to control access to computer programming that was developed at great cost. But the FST is, as a result, truly a "black box"—a program that cannot be used by defense experts with theories of the case different from the prosecution's.[247]

The court in *Collins* then disallowed the prosecution's evidence regarding this DNA analysis, finding the FST not sufficiently reliable.[248] The prosecution had wanted its expert to tell the jurors that "one DNA mixture was 972,000 times more probable if the sample originated from defendant Collins and

---

"reasonably effective assistance," and that a violation additional requires a showing of materiality and prejudice, that "but for counsel's unprofessional errors, the result of the proceeding would have been different").

244 *See, e.g.*, Jenny Roberts, *Too Little, Too Late: Ineffective Assistance of Counsel, the Duty To Investigate, and Pretrial Discovery in Criminal Cases*, 31 FORDHAM URB. L.J. 1097, 1128, 1128 nn.141–42 (2004).

245 Ben Grunwald, *The Fragile Promise of Openfile Discovery*, 49 CONN. L. REV. 771, 781 (2017).

246 *Cf.* John G. Douglass, *Fatal Attraction? The Uneasy Courtship of Brady and Plea Bargaining*, 50 EMORY L.J. 437, 443 (2001) (explaining how *Brady* governs disclosure before a trial or a plea but is almost always enforced after the fact, when a defendant tries to overturn a conviction).

247 People v. Collins, 15 N.Y.S.3d 564, 580 (Sup. Ct. 2015).

248 *See id.* at 587.

two unknown, unrelated people than if it instead originated from three unknown, unrelated individuals."[249] FST was not a black box technology that was designed to be uninterpretable; rather, it was a computer program, with code designed by New York City's Office of the Chief Medical Examiner, and the government simply refused to disclose it. Indeed, as noted, once the code was disclosed as the result of a federal court ruling,[250] experts reviewed it and found serious flaws, including in the way the FST grouped people by race, and then the Office ceased its use.[251]

As noted, a New Jersey appellate court ordered, in discovery, that if the state sought to use probabilistic DNA software, then the defendant was entitled to access, "under an appropriate protective order," the software's "source code and supporting software development and related documentation."[252] That ruling was recently cited by another New Jersey court in support of a defense request, regarding facial recognition technology, for "the identity, design, specifications, and operation of the program or programs used for analysis, and the database or databases used for comparison."[253]

However, many other courts have addressed, but not required, basic disclosures concerning use of AI in criminal cases. Indeed, a range of other courts in New York had admitted the FST evidence before these errors came to light.[254] A Pennsylvania court rejected a defense challenge to expert evidence concerning DNA mixture analysis, in the context of evaluating whether it was a "generally accepted" scientific methodology, finding it was "proprietary software."[255] A range of courts have admitted the results of DNA mixture software into evidence, by asserting it is reliable, or relying on precedent, but not clearly explaining why it should be permitted without validation of the AI or interpretability.[256] At best, they have

---

249 *Id.* at 565.

250 Order at 1, United States v. Johnson, No. 1:15CR00565 (S.D.N.Y. June 7, 2016) (No. 15CR565).

251 *See* Lauren Kirchner, *Traces of Crime: How New York's DNA Techniques Became Tainted*, N.Y. TIMES (Sept. 4, 2017), https://www.nytimes.com/2017/09/04/nyregion/dna-analysis-evidence-new-york-disputed-techniques.html [https://perma.cc/8H33-HY43].

252 State v. Pickett, 246 A.3d 279, 284 (N.J. Super. Ct. App. Div. 2021).

253 State v. Arteaga, 296 A.3d 542, 551, 557 (N.J. Super. Ct. App. Div. 2023).

254 *See, e.g.*, People v. Lopez, 23 N.Y.S.3d 820, 825 (Sup. Ct. 2015).

255 *See* Commonwealth v. Foley, 38 A.3d 882, 888–89 (Pa. Super. Ct. 2012).

256 *See, e.g.*, United States v. Russell, No. CR142563, 2018 WL 7286831, at *8 (D.N.M. Jan. 10, 2018).

found it sufficient that the software developer claimed to have validated the software.[257]

In other situations, the question of interpretable AI has been addressed only tangentially, but not squarely, by reviewing courts, including because lawyers have not themselves directly attacked the central problem of non-interpretability. For example, the most prominent legal challenge to a black box risk assessment program was brought in Wisconsin, where a defendant argued that it violated due process rights to base the sentence on an algorithm, called COMPAS and marketed by a private company (then called Northpointe, now Equivant), whose operation and validating information was not disclosed.[258] In *State v. Loomis*, the Wisconsin Supreme Court dismissed these claims, emphasizing "the proprietary nature of COMPAS," and that judges have discretion when they consider the risk instrument.[259]

Responding to the concerns raised by the defense, the Wisconsin Supreme Court did rule that sentencing judges must be given written warnings, or a "written advisement," about the risk tool, including cautioning judges that it relies on group data.[260] Those warnings seemed designed to address some concerns about the lack of transparency. However, such warnings do not open the black box to allow one to assess the operation or accuracy of the AI as used in an individual person's case.[261] The COMPAS system is not interpretable: one cannot know how it reached its results based on the data shared with the system, so one cannot check its correctness or assess whether its approach is valid. Nor did the court address the issue of possible noise in the data, such as typographical errors, that cannot be detected if one cannot see what the AI is relying on in a particular case. In fact, effort has been made by scientists to understand how COMPAS weighs important variables like race and age, but without much success.[262]

---

257 *See* Eli Siems, Katherine J. Strandburg & Nicholas Vincent, *Trade Secrecy and Innovation in Forensic Technology*, 73 HASTINGS L.J. 773, 814 (2022) ("[J]udges have been willing to allow developers of proprietary code to rely solely on labbased inputoutput testing that is not properly designed to uncover coding errors.").

258 881 N.W.2d 749, 763–65 (Wis. 2016).

259 *See id.* at 764–65.

260 *Id.* at 769.

261 The court asserted these warnings "enable courts to better assess the accuracy of the assessment and the appropriate weight to be given to the risk score." *Id.* at 764.

262 *See* Rudin, Wang & Coker, *supra* note 154; *see also* Dressel & Faird, *supra* note 54 (showing, in agreement with other researchers, "that although COMPAS

Notably, the defense in *Loomis* sought the source code, but not information regarding how COMPAS functioned and how it was developed based on training data.[263] In other words, the defense did not seek the information necessary to make the COMPAS AI interpretable.

Thus, while not always squarely raised, in a range of settings, courts have deferred to government claims that black box use of AI is justified in use at sentencing, reliance on AI by experts, and in other contexts. The government often claims black box AI offers something advantageous. In response, judges have tended to narrowly view defense requests for discovery regarding evidentiary uses of AI. To be sure, this has been a larger problem in forensic evidence generally.[264] Often, revelations of forensic errors occur many years after a conviction.[265] This makes it especially troubling that courts have continued to uphold the rights of companies to protect black box formulas and of government actors to conceal the bases of forensic AI.[266] A glass box approach can help to prevent such harmful deployment of AI in criminal cases. With the use of interpretable AI, errors can be better detected in individual cases.

B. Glass Box Equal Protection

Glass box AI systems also better safeguard rights under the Equal Protection Clause, which protects people from purposeful discrimination against protected groups, including

---

may use up to 137 features to make a prediction, the same predictive accuracy can be achieved with only two features, and that more sophisticated classifiers do not improve prediction accuracy or fairness"); Angelino, LarusStone, Alabi, Seltzer & Rudin, *supra* note 54, at 35 (finding that the accuracy of COMPAS can be predicted with a simple classifier); Sam CorbettDavies, Emma Pierson, Avi Feller & Sharad Goel, *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/2TNC-WZP8].

263 For an excellent discussion, see Ellora Israni, *Algorithmic Due Process: Mistaken Accountability and Attribution in* State v. Loomis, JOLT DIGEST (Aug. 31, 2017), https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1 [https://perma.cc/8DWG-B3WG].

264 For an overview, see GARRETT, *supra* note 41, at ch. 8.

265 *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 44–45 (2009) [hereinafter NAS Report] (describing audits and quality control failures at labs around the country).

266 Wexler, *supra* note 55.

from discrimination based on race.[267] To prove a violation of the Equal Protection Clause, the Supreme Court has held that a racially disparate impact is not sufficient; rather, a litigant must show that there was a racially discriminatory purpose to the government action.[268] Commentators have correctly pointed out that if AI is black box, then the government has plausible deniability when faced with an equal protection claim, even if outcomes using the AI can be shown to be racially disparate.[269] Criminal defendants may then be fighting an "unwinnable battle" in arguing that black box code is discriminatory, but without having any information about how the AI system works or whether it relies impermissibly on factors such as race.[270]

Yet, judges would be wrong to assume that the government has a plausible justification for using black box AI to prevent any inquiry into what its purpose was, or whether there is a disparate impact. Under the Equal Protection Clause, if strict scrutiny did apply, the government might claim a "compelling government interest" supporting the use of black box AI.[271] The court in *Loomis* relied on such a rationale.[272] Yet, the interest cannot be compelling if there is no well-supported performance advantage to the use of black box AI versus glass box AI. If the government is potentially obscuring potential discriminatory uses of race, with no well-justified and compelling benefit, then a judge should carefully inquire into how and why the government is using black box AI.

Challenges to criminal justice outcomes under the Equal Protection Clause face substantial hurdles, not least because of the latitude afforded to law enforcement under the Fourth Amendment[273] and the discretion afforded to prosecutors as executive actors.[274] The U.S. Supreme Court in *McCleskey v. Kemp* emphasized the difficulty and unwillingness to isolate,

---

267 *See* U.S. CONST. amend. XIV, § 1.

268 Washington v. Davis, 426 U.S. 229, 239 (1976).

269 For an overview of the doctrines regarding disparate impact and treatment, see Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341 (2010).

270 Leah Wisser, Note, *Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing*, 56 AM. CRIM. L. REV. 1811, 1818 (2019).

271 Palmore v. Sidoti, 466 U.S. 429, 432–33 (1984).

272 State v. Loomis, 881 N.W.2d 749, 759 (Wis. 2016) ("The need to have additional sound information is apparent . . . for sentencing courts.").

273 *See* Whren v. United States, 517 U.S. 806, 813 (1996) ("[T]he constitutional basis for objecting to intentionally discriminatory application of laws [by law enforcement] is the Equal Protection Clause, not the Fourth Amendment.").

274 United States v. Armstrong, 517 U.S. 456, 469 (1996).

even using statistical models, sources of bias in the criminal system.[275] A decision to use a black box AI model, which obscures potential discrimination, certainly raises different issues than the exercise of discretion by human decision makers.

However, even if the courts are not yet receptive to equal protection claims regarding AI, nevertheless, the risk of racial discrimination is a powerful reason not to permit black box AI in criminal cases on policy grounds. Thus, researchers have shown that in the probabilistic genotyping area, "software systems, designed to do the same job, produce different results and can have a disparate impact on different racial/ethnic groups."[276] Legislatures can insist that AI be carefully vetted to assure against discriminatory impacts on group of persons. Recent state legislation has imposed such requirements in the area of facial recognition technology.[277] It is much easier to check for the possibility of unlawful discrimination of any kind if the model is transparent.

C. *Glass Box* Daubert

At trial, an expert witness may present AI evidence; but if the AI is a black box, the parties cannot readily vet the expert to satisfy the foundational burden on the party seeking to introduce expert testimony. There are strong reasons to fear that judges will not rigorously examine black box AI evidence or insist on a glass box. Some judges have deferentially reviewed the admissibility of expert evidence in criminal cases, even after the U.S. Supreme Court's *Daubert* ruling and the 2000 amendments to Federal Rule of Evidence 702 tightened the gatekeeping requirements for expert evidence in federal court, with most states also now following the same approach.[278] The National Academy of Sciences ("NAS") explained, in a landmark 2009 report, that where judges have long failed to adequately scrutinize forensic evidence, scientific safeguards must be put into place by the legislatures.[279]

---

275 *See* 481 U.S. 279, 297, 312 (1987).

276 Jenna Neefe Mathews et al., *When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems*, 2020 PROC. AAAI/ACM CONF. ON AI ETHICS & SOC'Y 102, 108.

277 *See infra* subpart III.C.

278 For an overview, see Garrett & Fabricant, *supra* note 41, at 1599.

279 *See generally* NAS Report, *supra* note 265; *see also* Peter J. Neufeld, *The (Near) Irrelevance of* Daubert *to Criminal Justice and Some Suggestions for Reform*, 95 AM. J. PUB. HEALTH S107, S110 (2005).

The NAS report highlighted how courts routinely found admissible a range of forensic evidence of lacking in reliability, where: "With the exception of nuclear DNA analysis . . . no forensic method 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."[280] Expert testimony regarding traditional forensic evidence has resulted in tragic wrongful convictions and lab scandals.[281]

Most recently, concerns regarding the reliability of forensic evidence motivated the 2023 amendments to Rule 702, which underscore the burden on the party seeking to introduce an expert, as well as the need to examine the reliability of opinions an expert reaches.[282] The Advisory Committee to the Federal Rules of Evidence also highlighted how these revisions are "especially pertinent" to forensic expert evidence used in criminal cases.[283]

Those concerns regarding reliability of expert methods, the application of those methods, and the opinions reached, will each be important if black box AI is used in criminal cases. After all, the party with the burden of showing that the expert used reliable methods cannot readily satisfy that burden, if how the AI functioned is not interpretable. The emphasis in the rule revision regarding the burden on the party seeking to introduce the expert is also highly practically important in criminal cases. A criminal defendant, if indigent, may sometimes be denied funds to retain an expert to examine methods or technology used by a prosecution expert.[284] For black box AI, the barriers to practically challenging the evidence will be greater if the defendant has no way to independently re-examine prosecution use of AI.

In a range of contexts, judges have already permitted experts to testify regarding black box AI systems without disclosure of underlying source code or methods.[285] In the litigation concerning complex DNA mixtures, government actors have

---

280 NAS Report, *supra* note 265, at 7.

281 For an overview, see generally GARRETT, *supra* note 41, at 11–22.

282 ADVISORY COMM. ON RULES OF PRAC. & PROC., AGENDA BOOK 891–93 (2022).

283 Memorandum from John D. Bates, Chair Comm. on Rules Prac. & Proc. to Scott S. Harris, Sup. Ct. Clerk 227 (Oct. 19, 2022), https://www.uscourts.gov/sites/default/files/2022_scotus_package_0.pdf [https://perma.cc/37AJ-97SQ].

284 *See generally* Paul C. Giannelli & Sarah Antonucci, *Forensic Experts and Ineffective Assistance of Counsel*, 48 CRIM L. BULL. 1360 (2012).

285 Jennifer L. Mnookin, *Of Black Boxes, Instruments, and Experts: Testing the Validity of Forensic Science*, 5 EPISTEME 343, 352–55 (2008).

cited to the reliability of the technology as a reason to not to disclose underlying source code to defendants. They claimed there is no need to disclose information about their AI because it functioned well, and they have argued that reports by scientific groups like the NAS simply misunderstand the reliability of their AI systems.[286] Fortunately, some courts are beginning to view such claims of expert reliability as unsupported unless the black box is opened to examine and validate the AI.[287] As one court put it well, in the context of probabilistic genotyping, "affording meaningful examination of the source code, which compels the critical independent analysis necessary for a judge to make a threshold determination as to reliability at a [*Daubert*] hearing, is imperative."[288]

Further, even if a black box works on average, it may not have worked when applied in a particular case. Judges should examine, as Rule 702 requires them to do, not only whether a method used by an expert is reliable, with the burden on the party seeking to introduce the expert testimony, but whether it was reliably applied to the facts, and whether the opinions reached are sufficiently reliable.[289] Such as-applied scrutiny is particularly important when methods are not interpretable. Judges have raised concerns where experts chose not to document forensic analyses; the same concerns should be raised if an expert uses a black box AI system in a criminal case with analysis that is not interpretable.[290] Thus, properly applied, *Daubert* and Rule 702, together with state analogues, should provide substantial protections in criminal cases. However, to date, courts have often not granted relief to defendants, perhaps because they have accepted the myth of a black box performance advantage, which we seek to dispel in this Article.

IV
Toward Glass Box Regulation of AI

Ultimately, it is human decision makers who use information that an AI system provides in a criminal case. We need to ensure,

---

286 *See generally* Brief and Appendix on Behalf of the Attorney General Amicus Curiae, State v. Pickett, 246 A.3d 279 (N.J. Super. Ct. App. Div. 2021) (No. 085463).

287 *See, e.g.*, *Pickett*, 246 A.3d at 316 (ruling that TrueAllele source code regarding DNA analysis must be disclosed to the defense).

288 *Id.* at 323–24.

289 Fed. R. Evid. 702(a)–(d).

290 *See, e.g.*, Brandon L. Garrett, *The Reliable Application of Fingerprint Evidence*, 66 UCLA L. Rev. Discourse 64, 76 (2018).

though, that AI performs better and fairly or that it works at all; and further that the humans who make use of AI evidence understand its strengths and limitations. This Article has called into question any performance justification for black box AI and described the constitutional and statutory demands for glass box AI. This Part calls for regulation to require glass box AI in criminal settings, with a high justification for any departures from that norm of interpretability. We survey judicial and legislative approaches and suggest that a greater focus on interpretability could greatly strengthen regulatory efforts.

A. Glass Box Regulation

A glass box regulatory agenda is needed. A range of legal measures can ensure that black box AI is not used in the criminal system. As just discussed, far more can and should be done to apply and robustly protect the existing Bill of Rights in the U.S. Constitution, particularly when AI is used to provide evidence regarding criminal defendants. There is an unfortunate reality, however, that constitutional rights may not be enough to address these issues, where they have been unevenly enforced in criminal cases, given the challenges that largely indigent defendants face in obtaining adequate discovery and the pressures to plead guilty and waive trial rights. In Europe, where there is a "right to explanation" for AI under Article 22 of the General Data Protection Regulation ("GDPR"), a companion Law Enforcement Directive, adopted in 2016, restricts the use of AI in criminal investigations, including by requiring an assessment of its risk to "rights and freedoms" as well as data privacy.[291] Under the U.S. Constitution, criminal defendants deserve enhanced protection, not less protection than consumers receive. Having dispelled the myth that black box AI performs better, it is far easier to make both the constitutional case and the policy case that government has little justification for use of black box AI in criminal cases.

The legislative response to the use of black box AI in criminal cases has only just begun, and a focus of the first wave of local and state legislation in the United States has been police use of facial recognition technology. In the United States, several

---

[291] *See* LED, *supra* note 42, at art. 27; *see also* Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, art. 22, 2016 O.J. (L 119) 1, 4.

dozen localities have stopped using FRT systems, which is an appropriate response absent a glass box system. Today, a large coalition of groups seeks to ban facial recognition, calling it "unreliable, unjust, and a threat to basic rights and safety."[292] So far, ten states have passed restrictions on certain law enforcement uses of FRT, but these restrictions are not all likely to address the central problem. Two states, Vermont and Virginia, largely bar all use by law enforcement absent future statutory authorization.[293] Seven states have adopted regulations limiting use of FRT. Maine restricts the use of FRT by government officials to investigations of a "serious crime," or in identifying a deceased or missing person, and it permits use of facial recognition evidence in court.[294] Massachusetts restricts use of FRT by law enforcement and requires a warrant to use it in criminal cases.[295] New Hampshire limited the use of FRT, but not by law enforcement.[296] Maryland barred use of "facial recognition service[s]" during employment interviews.[297] Utah regulates the conditions under which FRT is used.[298] Three states, California, New York and Oregon, have adopted specific moratoria. California banned law enforcement from using FRT on body cameras until 2023, while New York passed a four-year moratorium on use of FRT in schools.[299] In perhaps the farthest-reaching legislation, Washington imposes detailed conditions and transparency requirements on all government use of facial recognition.[300]

Importantly, none of those laws require glass box use of AI for facial recognition, and we are aware of no proposals to do so. Thus, the particularly detailed Washington statute requires disclosure to criminal defendants and provides: "A state or local government agency must disclose use of a facial recognition

---

292 *See Ban Facial Recognition*, FIGHT FOR THE FUTURE, https://www.banfacialrecognition.com [https://perma.cc/PCJ7-YU53].

293 *See* VA. CODE ANN. § 15.21723.2 (2023) ("No local lawenforcement agency shall purchase or deploy facial recognition technology unless such purchase or deployment of facial recognition technology is expressly authorized by statute."); VT. STAT. ANN. TIT. 20, § 4622 (2023). Virginia permits airport uses of FRT, however, and Vermont permits use for suspectspecific drone-captured images.

294 *See* ME. STAT. tit. 25, § 6001 (2023).

295 MASS. GEN. LAWS CH. 6, § 220 (2023).

296 N.H. REV. STAT. ANN. § 359–N:2 (2023).

297 MD. CODE ANN., LAB. & EMPL. § 3–717 (LexisNexis 2023).

298 UTAH CODE ANN. § 77–23e–103. The provision requires disclosure of the use of facial recognition to a prosecutor. *Id.*

299 CAL. PENAL CODE § 832.19 (West 2020) (repealed Jan. 1, 2023); N.Y. EDUC. LAW § 2–e (Consol. 2020) (repealed).

300 WASH. REV. CODE § 43.386.070 (2023).

service on a criminal defendant in a timely manner prior to trial."[301] Government agencies must produce accountability reports that explain what data the FRT system uses and "how that data is generated, collected, and processed."[302] The type of vetting, review, and accountability under the Washington statute provides a good model for regulation of government use of AI in the criminal justice system regarding validation of the system. However, it does not address interpretability. It permits black box AI and does not require disclosure of anything interpretable to defendants, lawyers, and judges. Nor does it address the need for adequate criminal justice data to develop and assess the accuracy of an FRT AI system. Sound regulation of AI should address the underlying data collection and architecture needs as well.

In the consumer rights area at the federal level, the Federal Trade Commission has issued guidance regarding uses of AI in private industry to prevent unfair and deceptive practices.[303] For example, the FTC points to the concern that large sets of data used to train AI can raise privacy concerns.[304] The FTC does not discuss the possibility that glass box approaches be substituted for black box ones. Nor do any federal efforts address the federal government's own uses of AI in criminal cases.

We propose that legislation require glass box or interpretable AI be mandatory for most uses by law enforcement agencies in criminal investigations. So long as the use of AI could result in material or information used to investigate and potentially convict a person, it should be fully interpretable. Further, all law enforcement systems should be validated based on adequate data. Validation and interpretability should be required by statute. To our knowledge, no such proposals have been introduced, to date, in the United States, apart from the regulations discussed regarding FRT specifically.

One model for more searching regulation of AI, however, comes from the European Union, where the 2016 revision to the European Union's Law Enforcement Directive ("LED") limited

---

301 *Id.*

302 *Id.* § 43.386.020(2).

303 Press Release, Fed. Trade Comm'n, FTC Report Warns About Using Artificial Intelligence to Combat Online Problems, (June 16, 2022), https://www.ftc.gov/news-events/news/press-releases/2022/06/ftc-report-warns-about-using-artificial-intelligence-combat-online-problems [https://perma.cc/6X4G-AQFE].

304 AI tools can incentivize and enable invasive commercial surveillance and data extraction practices because these technologies require vast amounts of data to be developed, trained, and used. *Id.*

the use of AI in criminal cases.[305] Following adoption of that directive, a European Parliament report on the application of AI in criminal cases emphasized that any use of AI in such cases should "respect the principles of fairness, data minimization, accountability, transparency, non-discrimination and explainability," as well as use being subject to "risk assessment and strict necessity and proportionality testing."[306] The report, which resulted in a resolution by the European Parliament, clearly recognizes the need to address risks through careful testing, verification, and transparency and explainability (although we use the term interpretability).[307]

The AI Act in Europe will provide a model for even more substantial regulation of AI systems in a range of high-stakes settings, including vetting and review of systems put into use.[308] The Act emphasizes that all uses of AI by law enforcement are "high-risk" and subject to the enhanced pre-approval and oversight provisions, because "in the law enforcement context . . . accuracy, reliability and transparency is particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress."[309] Similarly, an AI that implicates the administration of justice and fair trials is considered "high-risk," in order "to address the risks of potential biases, errors and opacity."[310] Regarding the ability of persons to understand high-risk uses of AI, the Act provides:

> To address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information,

---

305 *See* LED, *supra* note 42.

306 *See* Report on Artificial Intelligence in Criminal Law and Its Use by the Police and Judicial Authorities in Criminal Matters, EUR. PARL. DOC. A90232/2021, ¶ 4 (2021).

307 *Id.* at ¶ 17 ("[The European Parliament] call[ed] for algorithmic explainability, transparency, traceability and verification as a necessary part of oversight, in order to ensure that the development, deployment and use of AI systems for the judiciary and law enforcement comply with fundamental rights, and are trusted by citizens, as well as in order to ensure that results generated by AI algorithms can be rendered intelligible to users and to those subject to these systems.").

308 *See* Artificial Intelligence Act, *supra* note 43.

309 *Id.* art. 38.

310 *Id.* art. 40.

> including in relation to possible risks to fundamental rights and discrimination, where appropriate.[311]

We hope that accompanying regulations explain in more detail the need for interpretability. The draft Act does state that users should be "able to interpret" outputs in a way that they understand and use "appropriately."[312] That language should be operationalized to insist on interpretability regarding the factors the AI used in a particular instance, and not just post-hoc explainability.

In the United States, where no such rules exist, legislative efforts also should be aimed at ensuring government agencies at local, state, and federal levels do not violate constitutional criminal procedure rights through non-transparent and unfair AI practices. The U.S. House of Representatives considered a "Justice in Forensic Algorithms Act," which would ensure that any algorithms used in criminal cases be unrestricted by any claim of proprietary or trade secrets protection, and vetted by NIST—but the bill has not been enacted into law.[313] Congressman Dwight Evans said: "Opening the secrets of these algorithms to people accused of crimes is just common sense and a matter of basic fairness and justice. People's freedom from unjust imprisonment is at stake, and that's far more important than any company's claim of 'trade secrets.'"[314] The law would have provided an important starting place. More recently, Congressional hearings and calls for regulation have mounted, and perhaps we will eventually see more progress in legislation.[315]

B. Towards a Right to Glass-Box AI

A strong presumption of interpretability for criminal courtroom uses of AI should be recognized and grounded in

---

311 *Id.* art. 47.

312 *Id.*

313 *See* Press Release, Mark Takano, Congressman, House of Representatives, Reps. Takano and Evans Reintroduce the Justice in Forensic Algorithms Act to Protect Defendants' Due Process Rights in the Criminal Justice System (Apr. 8, 2021), https://takano.house.gov/newsroom/press-releases/reps-takano-and-evans-reintroduce-the-justice-in-forensic-algorithms-act-to-protect-defendants-due-process-rights-in-the-criminal-justice-system [https://perma.cc/Z5B9-ZMMG].

314 *Id.*

315 Claudia Grisales, *Despite Many Briefings and Hearings, Lawmakers Have a Long Way to Go to Regulate AI*, NPR, (July 26, 2023), https://www.npr.org/2023/07/26/1190327582/despite-many-briefings-and-hearings-lawmakers-have-a-long-way-to-go-to-regulate. [https://perma.cc/8T3U-8BZL].

existing constitutional criminal procedure rights, and it could be enhanced by legislation and regulation. This presumption should be used by judges when conducting due process analysis and by policymakers when deciding whether to deploy or regulate AI in a criminal system. Just as the European Union suggests that "high-risk" uses of AI by law enforcement or by courts merits far more stringent oversight, uses of AI in any criminal context should require interpretability, among other restrictions on its design and use.

This is not to say that black box AI is never possible to use in criminal settings, but rather that the presumption against it should be strong and the government should have to show something like a compelling state interest to support its use. Thus, there may be situations in which the government can offer a sufficient basis to protect certain types of AI systems from disclosure, for which this strong presumption may be overcome. For example, one could imagine a national security justification for not making public aspects of a particular AI model. However, at a minimum, the AI should be carefully vetted by independent researchers with appropriate security safeguards. Further, for court users such as criminal defense lawyers, a glass box is necessary to safeguard defendants' rights and assure both fairness and public safety.

Relatedly, given the rapid pace of technological change, it is possible that in the near future, new AI systems will show a powerful performance advantage, even in criminal justice settings that pose so many challenges. Given what we have seen in the past, it is likely that many of those systems will be "black box" systems not designed to display to users what factors are relied upon to make predictions. If a high burden can be met to show why an interpretable model cannot be used, then the AI could be used, with explainable rather than fully interpretable outputs. We emphasize, though, that in criminal cases, the burden should be placed on the government to justify the use of black box AI systems that lack interpretability.

Instead, judges have often imposed a presumption in favor of black box AI, citing to the need to protect proprietary interests of software developments. These private companies may advance a for-profit innovation justification for keeping an AI system proprietary. Others have responded at length to such claims, noting proprietary methods can readily be disclosed under seal and no established privilege applies, while in contrast, constitutional rights are implicated by defense access

to evidence.[316] In the DNA mixture context, the report by the President's Council of Advisors on Science and Technology has highlighted deep reasons for concern regarding the accuracy of probabilistic genotyping software and called for independent validation of such software.[317]

We highlight a different point responsive to the claims of business justifications for keeping AI as a black box. It may be the case that companies would have a harder time profiting by selling open and interpretable AI. However, the government should not incentivize profits if there is not any substantial showing that the products performs better than the interpretable alternatives. Nor should the government pay for something black box that could instead be interpretable, given the special risks to the public and constitutional rights, discussed next. Unfortunately, courts have often admitted use of black box software that is not independently validated, interpretable, and that is not disclosed in court.[318] As Erin Murphy argues, "courts should disallow statistical evidence generated by probabilistic software whose operators refuse to reveal their code."[319] As discussed next, maintaining a black box can violate a range of constitutional criminal procedure rights.

Further, researchers, government agencies, and non-profits can readily develop glass-box AI systems, and they have increasingly done so. Researchers, for example, developed a screener with simple factors for police to forecast domestic violence.[320] Other researchers developed a simple scoring system to address unnecessary use of stop and frisk by the New York City Police Department.[321] Pretrial risk assessments commonly involve simple scoring systems, focusing on factors like age and prior convictions.[322] Researchers have created free, open-source probabilistic genotyping software for interpreting

---

316 *See* Wexler, *supra* note 38; Katherine Kwong, Note, *The Algorithm Says You Did It: The Use of Black Box Algorithms to Analyze Complex DNA Evidence*, 31 HARV. J.L. & TECH. 275, 290 (2017).

317 *See* PCAST Report, *supra* note 11, at 78–80.

318 ERIN MURPHY, INSIDE THE CELL: THE DARK SIDE OF FORENSIC DNA 299 (2015).

319 *Id.* at 300.

320 Richard A Berk, Yan He & Susan B Sorenson, *Developing a Practical Forecasting Screener for Domestic Violence Incidents*, 29 EVAL. REV. 358, 358 (2005).

321 Sharad Goel, Justin M. Rao & Ravi Shroff, *Precinct or Prejudice? Understanding Racial Disparities in New York City's StopAndFrisk Policy*, 10 EVAL. REV 365, 365 (2015).

322 *See* Desmarais, Zottola, Duhart Clarke & Lowder, *supra* note 72 at 416.

DNA mixtures.[323] Simple AI systems can perform better and provide far more understandable information to criminal justice actors, without concealing errors inside a black box.

The use of facial recognition technology by the federal government provides a troubling illustration of what happens when no regulation clearly imposes any burden of justification on the use of AI by agencies or any rules concerning its use. As noted, the federal government has failed to open the black box on its FRT programs, shared with local law enforcement agencies around the country and used by federal agencies.[324] What we do know suggests that when FRT is trained on limited data, and depending on the design, FRT can be racially biased.[325] Yet no law or regulation clearly applies to require a glass box approach to FRT. Absent any regulations assuring that use of FRT is limited solely to preliminary investigations, an informal assurance to that effect is not adequate. To be sure, if the government uses FRT in wholly non-law-enforcement settings, then the burden of justification may be different. For example, the growing use of FRT to expedite entering security at an airport may be a welcome convenience for some travelers.[326] Even then, however, the public should be shown that the government is using reliable and non-discriminatory systems. Moreover, there may be criminal procedure concerns if the images are retained and used for law enforcement purposes.

Unfortunately, much of the usage by government has been ad hoc, providing no assurance that AI is being used in ways that protect rights. For example, some federal agencies do not even know what FRT systems they are using and have

---

323 *See* LRMIX STUDIO, https://github.com/smartrank/lrmixstudio [https://perma.cc/46ZD-K8MU] (last visited Feb. 9, 2024).

324 *See* U.S. GOV'T ACCOUNTABILITY OFF., GAO25526, FACIAL RECOGNITION TECHNOLOGY: CURRENT AND PLANNED USES BY FEDERAL AGENCIES (2021) (noting "18 of the 24 surveyed agencies reported using an FRT system, for one or more purposes").

325 *See About Face: Examining the Department of Homeland Security's Use of Facial Recognition and Other Biometric Technologies, Part II, Hearing Before the Comm. on Homeland Sec.*, 116th Cong. 6 (2020) (statement of Charles Romine, Dir., Info. Tech. Lab'y, U.S. Dep't Com.). Securing large datasets raises still other concerns, like Clearview AI's technology that uses biometric information from Internet users who did not give permission to the company. Kashmir Hill, *The Secretive Company that Might End Privacy as We Know it*, N.Y. TIMES (Jan 18, 2020), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html [https://perma.cc/9U67-PD3F]

326 Geoffrey Fowler, *TSA Now Wants to Scan Your Face at Security. Here are Your Rights*, WASH. POST (Dec. 2, 2022), https://www.washingtonpost.com/technology/2022/12/02/tsa-security-face-recognition/ [https://perma.cc/F9Z5-9BQ4].

not assessed the possible risks, according to a report by the Government Accountability Office.[327] This example also illustrates a different policy point: the burden of justification to use black box AI may vary depending on the purposes to which the AI system is put.[328]

Further, if a system brings in data for *multiple* purposes, such as if a biometrics database is used by the government in a non-criminal context, but law enforcement may still have access to it in criminal cases, then the need for glass box AI is particularly pressing. If it is a black box system, it will not be possible to adapt to the legal and constitutional demands that accompany use of evidence in court, even if the data is often used for non-court uses. Constitutional violations will inevitably follow. This makes it all the more important that careful regulations govern how agencies use AI.

We also emphasize that a key reason judges and other legal actors have failed to adequately scrutinize AI, in addition to the legal standards and precedent just discussed, is precisely because it has mainly been a black box. Judges have had no way to appreciate its limitations. Had they carefully examined the burden of justification for using black box AI, they might have reached different conclusions.

One final cautionary legislative and regulatory lesson comes from the First Step Act, in which federal lawmakers sought to require more open uses of AI but did not do so clearly or carefully enough. In the Act, Congress required the use of risk assessments to determine when federal prisoners would be eligible for release and to allocate prison programming.[329] The Act called for researchers to design this new risk assessment tool, a panel of researchers to vet the research design, annual validation, and "a requirement that [BOP staff] demonstrate

---

327 U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 324.

328 State legislation has reflected these distinctions. For example, Washington State lawmakers provided: "(1) Unconstrained use of facial recognition services by state and local government agencies poses broad social ramifications that should be considered and addressed. Accordingly, legislation is required to establish safeguards that will allow state and local government agencies to use facial recognition services in a manner that benefits society while prohibiting uses that threaten our democratic freedoms and put our civil liberties at risk. (2) However, state and local government agencies may use facial recognition services to locate or identify missing persons, and identify deceased persons, including missing or murdered indigenous women, subjects of Amber alerts and silver alerts, and other possible crime victims, for the purposes of keeping the public safe." WASH. REV. CODE § 43.386.900 (2023).

329 First Step Act of 2018, Pub. L. No. 115–391, 132 Stat. 5194, 5197 (2018). (codified in scattered sections of 18, 21, 34, and 42 U.S.C).

competence in administering the System, including interrater reliability, on a biannual basis."[330]

Unfortunately, the First Step Act did not require glass box AI. The Act resulted in the development of a risk assessment instrument, where the developers, as well as the Department of Justice in approving it, did not justify key choices regarding the selection of risk thresholds.[331] The Act did not provide guidance on the key issue of what should be deemed high, medium, or low risk, nor did the Department of Justice when implementing it.[332] The Act provided even less information about how treatment-related "needs" items should be used.[333] Since then, the risk assessment's developers have shared only limited information, but have disclosed that they uncovered design errors.[334] A glass box approach could have avoided an ongoing string of mishaps.

We advocate a right to glass box AI, requiring glass box AI in criminal justice settings including in non-trial settings. AI should be interpretable and accessible to criminal justice actors. Even absent legislation, a glass box approach should be adopted by government agencies at the federal, state, and local levels. Basic interpretability requirements should be adhered to by all government agencies, law enforcement, and judges using AI in criminal cases. All should face a high burden of justification for departure from the glass box norm.

CONCLUSION

Black box AI has already infiltrated far too many important criminal justice settings. Judges, lawmakers, and executive actors have been too often misled by a myth of black box performance. When they scrutinize AI, each of these actors should place a high burden of justification on those proposing to maintain non-transparent, black box use of AI in criminal law settings, and they should view unsubstantiated claims of superior black box AI accuracy with deep suspicion.

The U.S. Constitution safeguards rights to a fair trial under the Due Process Clauses and Sixth Amendment confrontation

---

330 Garrett & Stevenson, *supra* note 82, at 280.

331 *Id.* at 282.

332 *Id.* at 280.

333 *Id.*

334 NAT'L INST. JUST., 2020 REVIEW AND REVALIDATION OF THE FIRST STEP ACT RISK ASSESSMENT TOOL 5 (2021), https://www.ojp.gov/pdffiles1/nij/256084.pdf [https://perma.cc/AB46-2F38].

rights, and prohibits discrimination in violation of the Equal Protection Clause and implementing civil rights acts. Expert evidence rules should ensure that scientific evidence is carefully vetted before being admitted in a trial.

These constitutional and statutory protections have been tested as black box AI is deployed in criminal settings. The early judicial responses have not been very reassuring. Judges have rarely intervened, sometimes because they have credited claims that proprietary AI is needed to generate investment in technology, or because they have assumed it is simply not practically possible to open black box technology. Perhaps this will change as awareness is raised among litigants and judges of the risks of uncritical acceptance of AI. Judges have already faced and they will increasingly confront pressing questions about whether black box AI is authorized, justified, and constitutional.

In this Article, we seek to puncture the myth of superhuman or even superior black box AI performance over glass box, fully interpretable AI. Only for glass box AI can one truly evaluate the costs and the benefits of using an AI system. In fact, glass box AI can perform better than the black box alternatives. Further, both the benefits of glass box AI and the costs of black box AI are heightened in criminal cases, given concerns with poor data quality, biased data, and uninformed user discretion. Finally, interpretable AI can far better protect a range of constitutional rights. That is why we recommend that judges and lawmakers should largely require glass box AI and ban black box AI, while we acknowledge that in specific situations, and perhaps based on new technology, the substantial burden of justifying black box AI could potentially be overcome by compelling interests and a detailed showing regarding performance. Regulations and statutes regarding the deployment of AI in criminal law settings should directly require interpretable and validated AI.[335] In short, it is time to recognize in criminal cases a right to glass box AI.

---

[335] The U.S. Department of Justice now has an opportunity to recognize such an approach, in recommendations that the Attorney General must set out in response to the Executive Order regarding AI. *See* Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023).