THE JUDICIAL DEMAND FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

*Ashley Deeks**

A recurrent concern about machine learning algorithms is that they operate as "black boxes," making it difficult to identify how and why the algorithms reach particular decisions, recommendations, or predictions. Yet judges are confronting machine learning algorithms with increasing frequency, including in criminal, administrative, and civil cases. This Essay argues that judges should demand explanations for these algorithmic outcomes. One way to address the "black box" problem is to design systems that explain how the algorithms reach their conclusions or predictions. If and as judges demand these explanations, they will play a seminal role in shaping the nature and form of "explainable AI" (xAI). Using the tools of the common law, courts can develop what xAI should mean in different legal contexts. There are advantages to having courts to play this role: Judicial reasoning that builds from the bottom up, using case-by-case consideration of the facts to produce nuanced decisions, is a pragmatic way to develop rules for xAI. Further, courts are likely to stimulate the production of different forms of xAI that are responsive to distinct legal settings and audiences. More generally, we should favor the greater involvement of public actors in shaping xAI, which to date has largely been left in private hands.

INTRODUCTION

A recurrent concern about machine learning algorithms is that they operate as "black boxes." Because these algorithms repeatedly adjust the way that they weigh inputs to improve the accuracy of their predictions, it can be difficult to identify how and why the algorithms reach the outcomes they do. Yet humans—and the law—often desire or demand answers to the questions "Why?" and "How do you know?" One way to address the "black box" problem is to design systems that explain how the algorithms reach their conclusions or predictions. Sometimes called "explainable AI" (xAI), legal and computer science scholarship has identified various actors who could benefit from (or who should demand) xAI. These include criminal defendants who receive long sentences based on opaque predictive algorithms,[1] military commanders who are

* E. James Kelly Jr.—Class of 1965 Research Professor of Law, University of Virginia Law School. Thanks to Danielle Citron, John Duffy, Bert Huang, Leslie Kendrick, Rich Schragger, and Rebecca Wexler for helpful comments, and to Scott Harman-Heath for excellent research assistance.

1. See, e.g., Megan T. Stevenson & Christopher Slobogin, Algorithmic Risk Assessments and the Double-Edged Sword of Youth, 36 Behav. Sci. & L. 638, 639 (2018).

considering whether to deploy autonomous weapons,[2] and doctors who worry about legal liability for using "black box" algorithms to make diagnoses.[3] At the same time, there is a robust—but largely theoretical—debate about which algorithmic decisions require an explanation and which forms these explanations should take.

Although these conversations are critically important, they ignore a key set of actors who will interact with machine learning algorithms with increasing frequency and whose lifeblood is real-world controversies: judges.[4] This Essay argues that judges will confront a variety of cases in which they should demand explanations for algorithmic decisions, recommendations, or predictions. If and as they demand these explanations, judges will play a seminal role in shaping the nature and form of xAI. Using the tools of the common law, courts can develop what xAI should mean in different legal contexts, including criminal, administrative, and civil cases. Further, there are advantages to having courts play this role: Judicial reasoning that builds from the bottom up, using case-by-case consideration of the facts to produce nuanced decisions, is a pragmatic way to develop rules for xAI.[5] In addition, courts are likely to stimulate (directly or indirectly) the production of different forms of xAI that are responsive to distinct legal settings and audiences. At a more theoretical level, we should favor the greater involvement of public actors in shaping xAI, which to date has largely been left in private hands.

Part I of this Essay introduces the idea of xAI. It identifies the types of concerns that machine learning raises and that xAI may assuage. It then considers some forms of xAI that currently exist and discusses the advantages to each form. Finally, it identifies some of the basic xAI-related choices judges will need to make when they need or wish to understand how a given algorithm operates.

---

2. See Matt Turek, Explainable Artificial Intelligence (XAI), DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence [https://perma.cc/ZNL9-86CF] (last visited Aug. 13, 2019).

3. W. Nicholson Price II, Medical Malpractice and Black-Box Medicine, *in* Big Data, Health Law, and Bioethics 295, 295–96 (I. Glenn Cohen, Holly Fernandez Lynch, Effy Vayena & Urs Gasser eds., 2018).

4. See, e.g., Lilian Edwards & Michael Veale, Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking for, 16 Duke L. & Tech. Rev. 18, 67 (2017) [hereinafter Edwards & Veale, Slave to the Algorithm] (questioning whether xAI will be useful because "[i]ndividual data subjects are not empowered to make use of the kind of algorithmic explanations they are likely to be offered" but ignoring the possible role for courts as users of xAI).

5. Cf. Andrew Tutt, An FDA for Algorithms, 69 Admin. L. Rev. 83, 109 (2017) (proposing a federal statutory standard for explainability and arguing that "[i]f explainability can be built into algorithmic design, the presence of a federal standard could nudge companies developing machine-learning algorithms into incorporating explainability from the outset"). I share Andrew Tutt's view that it is possible to provide incentives for designers to incorporate xAI into their products, but I believe that there are advantages to developing these rules using common law processes.

Against that background, the Essay then turns to two concrete areas of law in which judges are likely to play a critical role in fleshing out whether xAI is required and, if so, what forms it should take. Part II considers the use of machine learning in agency rulemaking and adjudication and argues that judges should insist on some level of xAI in evaluating the reasons an agency gives when it produces a rule or decision using algorithmic processes.[6] Further, if agencies employ advanced algorithms to help them sort through high volumes of comments on proposed rules, judges should seek explanations about those algorithms' parameters and training.[7] In both cases, if judges demand xAI as part of the agency's reason-giving process, agency heads themselves will presumably insist that their agencies regularly employ xAI in anticipation of litigation.

Part III explores the use of predictive algorithms in criminal sentencing. These algorithms predict the likelihood that a defendant will commit additional crimes in the future. Here, the judge herself is the key consumer of the algorithm's recommendations, and has a variety of incentives—including the need to give reasons for a sentence, concerns about reversal on appeal, a desire to ensure due process, and an interest in demonstrating institutional integrity—to demand explanations for how the sentencing algorithm functions.

As courts employ and develop existing case law in the face of predictive algorithms that arise in an array of litigation, they will create the "common law of xAI," law sensitive to the requirements of different audiences (judges, juries, plaintiffs, or defendants) and different uses for the explanations given (criminal, civil, or administrative law settings).[8] A nuanced common law of xAI will also provide important incentives and feedback to algorithm developers as they seek to translate what are currently theoretical debates into concrete xAI tools.[9] Courts should focus on the power of xAI to identify algorithmic error and bias and the need

---

6. For the argument that judicial review of agency rulemaking employs common law methodologies, see Jack M. Beermann, Common Law and Statute Law in Administrative Law, 63 Admin. L. Rev. 1, 3 (2011).

7. See Melissa Mortazavi, Rulemaking Ex Machina, 117 Colum. L. Rev. Online 202, 207–08 (2017), https://columbialawreview.org/wp-content/uploads/2017/09/Mortazavi-v5.0.pdf [https://perma.cc/SF8R-EG9C] (examining the possibility that agencies may deploy automated notice-and-comment review).

8. See Finale Doshi-Velez & Mason Kortz, Berkman Klein Ctr. Working Grp. on Explanation & the Law, Accountability of AI Under the Law: The Role of Explanation 12 (2017), https://arxiv.org/pdf/1711.01134.pdf [https://perma.cc/LQB3-HG7L] ("As we have little data to determine the actual costs of requiring AI systems to generate explanations, the role of explanation in ensuring accountability must also be re-evaluated from time to time, to adapt with the ever-changing technology landscape.").

9. At least one scholarly piece has concluded that "there is some danger of research and legislative efforts being devoted to creating rights to a form of transparency that may not be feasible, and may not match user needs." Edwards & Veale, Slave to the Algorithm, supra note 4, at 22. A common law approach to xAI can help ensure that the solutions are both feasible and match user needs in specific cases.

for xAI to be comprehensible to the relevant audience. Further, they should be attuned to dynamic developments in xAI decisions across categories of cases when looking for relevant precedent and guidance.

I. THE WHAT AND WHY OF EXPLAINABLE AI

Artificial intelligence is a notoriously capacious and slippery term. Generally, it refers to "a set of techniques aimed at approximating some aspect of human or animal cognition using machines."[10] More concretely, scientists and scholars often use the term to encompass technologies that include machine learning, speech recognition, natural language processing, and image recognition.[11] Machine learning systems and algorithms, the driving force behind many AI developments, are valuable because of their ability to learn for themselves "how to detect useful patterns in massive data sets and put together information in ways that yield remarkably accurate predictions or estimations."[12] Many machine learning systems are trained on large amounts of data and adjust their own parameters to improve the reliability of their predictions over time.[13] Machine learning tools hold out the possibility of making more accurate decisions, faster, based on far larger quantities of data than humans can process and manipulate.[14] Importantly, though, because a machine learning system learns on its own and adjusts its parameters in ways its programmers do not specifically dictate, it often remains unclear precisely how the system reaches its predictions or recommendations.[15] This is particularly true for "deep learning" systems that use "neural networks," which are intended to replicate neural processes in the human brain.[16] Deep learning systems use nodes, arranged in multiple layers, which transfer information to each other and learn on their own how to weigh

10. Ryan Calo, Artificial Intelligence Policy: A Primer and Roadmap, 51 U.C. Davis L. Rev. 399, 404 (2017).

11. Artificial Intelligence, Lexico, https://www.lexico.com/en/definition/artificial_intelligence [https://perma.cc/MNB4-ZENF] (last visited Oct. 15, 2019) (defining "artificial intelligence" as "[t]he theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages").

12. Cary Coglianese & David Lehr, Transparency and Algorithmic Governance, 71 Admin. L. Rev. 1, 6 (2019) [hereinafter Coglianese & Lehr, Governance]; see also id. at 14–16 (describing how machine learning differs from traditional statistical techniques).

13. See Ethem Alpaydin, Machine Learning: The New AI 24–25 (2016).

14. See Coglianese & Lehr, Governance, supra note 12, at 16.

15. See Alpaydin, supra note 13, at 155; Will Knight, The Dark Secret at the Heart of AI, MIT Tech. Rev., May/June 2017, at 55, 56–57.

16. See James Farrant & Christopher M. Ford, Autonomous Weapons and Weapon Reviews: The UK Second International Weapon Review Forum, 93 Int'l L. Stud. 389, 400 (2017); 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 & n.135 (2017).

connections between nodes to correctly interpret objects in, say, a video image.[17]

Notwithstanding its potential benefits, the use of machine learning has prompted a number of concerns, especially when the systems make predictions that affect people's liberty, safety, or privacy. One strand of criticism focuses on the ways in which these algorithms can replicate and exacerbate societal biases in light of the data on which scientists train them. Another line of critiques questions the accuracy of various machine learning predictions, with objectors claiming that tools such as criminal justice algorithms predict recidivism less accurately than humans.[18]

A third concern, and the one most salient to this Essay, centers on the lack of information about how the algorithm arrives at its results—the "black box" problem.[19] The inability to parse the reasons behind the algorithm's recommendations can harm those affected by the recommendations. Opaque algorithms can undercut people's sense of fairness and trust—particularly when used by the government—and in the criminal justice setting can undercut a defendant's right to present a defense. This Essay focuses on algorithms' lack of transparency and interpretability for two related reasons. First, shedding light on how an algorithm produces its recommendations can help address the other two critiques, by allowing observers to identify biases and errors in the algorithm.[20] Second, computer scientists have begun to make promising inroads into the problem by developing what is often referred to as "explainable AI."[21]

---

17. See Farrant & Ford, supra note 16, at 400–01.

18. See Julia Dressel & Hany Farid, The Accuracy, Fairness, and Limits of Predicting Recidivism, Sci. Advances, Jan. 2018, at 1, 3, https://advances.sciencemag.org/content/4/1/eaao5580/tab-pdf (on file with the *Columbia Law Review*).

19. See, e.g., Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 3–4 (2015); Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249, 1254 (2008) (expressing concern about the "opacity of automated systems" used to inform administrative rulemaking).

20. Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning 1, 3 (2017), https://arxiv.org/pdf/1702.08608.pdf [https://perma.cc/ALR4-DM7J] ("[I]f the system can *explain* its reasoning, we then can verify whether that reasoning is sound with respect to . . . other desiderata—such as fairness, privacy, reliability, robustness, causality, usability and trust . . . .").

21. For a recent survey of developments in xAI, see Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, arXiv (May 31, 2018), https://arxiv.org/pdf/1806.00069.pdf [https://perma.cc/3SG4-G5GA] (last updated Feb. 3, 2019). One reason for recent progress in this area is the entry into force of the European Union's General Data Protection Regulation, which contains provisions that arguably give individuals affected by purely algorithmic decisions a "right to an explanation." See Regulation (EU) 2016/679 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 and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, arts. 13–15, 2016 O.J. (L 119) 41–43 (providing rights to "meaningful information about the logic involved" in certain automated decisions). The existence of these provisions, coupled with a lack of detail about what form those explanations must take, has triggered extensive discussions in

xAI encompasses a range of efforts to explain—or help humans interpret—how a particular machine learning model reached its conclusion. The concept of an explanation here "has come to refer to providing insight into the internal state of an algorithm, or to human-understandable approximations of the algorithm."[22] xAI provides a variety of benefits: It can foster trust between humans and the system,[23] identify cases in which the system appears to be biased or unfair, and bolster our own knowledge of how the world works.[24] As discussed below, in legal settings xAI can benefit judges who wish to rely on the algorithms for decisional support, litigants who seek to persuade judges that their use of algorithms is defensible, and defendants who wish to challenge predictions about their dangerousness.[25] xAI is not without costs, however. Most significantly, making an algorithm explainable may result in a decrease in its accuracy.[26] xAI may also stifle innovation, force developers to reveal trade secrets, and impose high monetary costs because xAI can be expensive to build.[27]

Fortunately, a variety of xAI currently exists, and computer scientists continue to develop new forms of it.[28] Some machine learning models are built to be intrinsically explainable, yet these models are often less

---

the legal and machine learning communities about how and in what form to explain the results of highly complex algorithms to experts and nonexperts.

22. 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); see also Doshi-Velez & Kim, supra note 20, at 2 (defining interpretability as the "ability to explain or to present in understandable terms to a human").

23. See Knight, supra note 15, at 61 (describing "explainability as the core of the evolving relationship between humans and intelligent machines"); Turek, supra note 2 ("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.").

24. See Doshi-Velez & Kim, supra note 20, at 3.

25. See, e.g., Robin A. Smith, Opening the Lid on Criminal Sentencing Software, Duke Today (July 19, 2017), https://today.duke.edu/2017/07/opening-lid-criminal-sentencing-software [https://perma.cc/F63A-VWLQ] ("Using . . . machine learning, Rudin and colleagues are training computers to build statistical models to predict future criminal behavior . . . that are just as accurate as black-box models, but more transparent and easier to interpret.").

26. See Doshi-Velez & Kortz, supra note 8, at 2 ("[E]xplanation would come at the price of system accuracy or other performance objective[s].").

27. See id. at 2, 12 ("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.").

28. This Essay's discussion of categories of xAI is necessarily simplified, because there are a wide range of approaches to categorizing xAI and the nomenclature is unsettled. For a survey of the literature on types of xAI and a detailed taxonomy thereof, see Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi & Fosca Giannotti, A Survey of Methods for Explaining Black Box Models 6–8 (2018), https://arxiv.org/pdf/1802.01933.pdf [https://perma.cc/P8PH-Z5V7].

complex as a result and tend to be less accurate in their predictions.[29] Another set of models is not intrinsically explainable. For these models, computer scientists have taken two basic approaches.[30] One type (which this Essay terms an "*exogenous* approach") does not attempt to actually explain the inner workings of (that is, the reasoning of) the machine learning algorithm. Instead, it attempts to provide relevant information to the algorithm's user or subject about how the model works using extrinsic, orthogonal methods.[31] A second type of approach actually attempts to explain or replicate the model's reasoning, and sometimes is referred to as a "*decompositional* approach."[32]

Exogenous xAI approaches can either be *model-centric* or *subject-centric*.[33] A model-centric approach, also referred to as global interpretability,[34] might involve, for instance, explaining the creator's intentions behind the modelling process, the family of model the system uses, the parameters the creators specified before training the system, qualitative descriptions of the input data the creator used to train the model, how the model performed on new data, and how the creators tested the data for undesirable properties.[35] In other words, this constitutes a thick description of the parts of the model that are knowable. A different type of model-centric approach might audit the outcomes of the machine learning system.[36] This approach would scour the system's decisions or recommendations for appearances of bias or error. Model-centric approaches attempt to explain the whole model, rather than its performance

29. These include linear, parametric, and tree-based models. Dipanjan Sarkar, The Importance of Human Interpretable Machine Learning, Towards Data Sci. (May 24, 2018), https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476 [https://perma.cc/4XD8-F7CD]. For an argument that society should use only intrinsically interpretable models for high-stakes decisions, see generally Cynthia Rudin, Please Stop Explaining Black Box Models for High-Stakes Decisions (2018), https://arxiv.org/pdf/1811.10154.pdf [https://perma.cc/Q7SF-6DYN].

30. See Guidotti et al., supra note 28, at 2 (characterizing one category of xAI as focused on describing how black boxes work and another on explaining decisions without understanding how the decision systems work); Edwards & Veale, Slave to the Algorithm, supra note 4, at 64–65 (describing two styles of algorithmic explanation: one that "opens" the black box and one that does not).

31. See Edwards & Veale, Slave to the Algorithm, supra note 4, at 65 ("[P]edagogical systems . . . can get the information they need by simply querying it, like an oracle." (emphasis omitted)).

32. Id. at 64.

33. Id. at 22.

34. See Sarkar, supra note 29.

35. See Edwards & Veale, Slave to the Algorithm, supra note 4, at 55–56.

36. Joshua A. Kroll, Solon Barocas, Edward W. Felton, Joel R. Reidenberg, David G. Robinson & Harlan Yu, Accountable Algorithms, 165 U. Pa. L. Rev. 633, 660–61 (2017) (explaining that auditing may test for discrimination in bargaining processes such as retail car negotiations).

in a particular case, and can help ensure that decisions are being made in a procedurally regular way.[37]

A subject-centric approach, also referred to as local interpretability,[38] in contrast, might provide the subject of a recommendation or decision with information about the characteristics of individuals who received similar decisions.[39] Another subject-centric approach involves the use of counterfactuals.[40] Here, people seeking to understand which factors may have most affected the algorithm's recommendation about them may, using that same algorithm, tweak the input factors to test how much a given factor mattered in the original recommendation.[41] For example, an algorithm that deems someone convicted of an offense to be at high risk of reoffending could be tested with counterfactuals to see whether the recommendation would have been different if the person were ten years older, or had one fewer arrest. The counterfactual approach could take different forms: It might present several "close possible worlds" or one "closest possible world," and it might alter one factor or several different factors.[42] One advantage of an exogenous approach is that it does "not require the data subject to understand any of the internal logic of a model in order to make use of it."[43] Subject-centric approaches can be particularly useful for individuals who are seeking to understand "if and how they might achieve a different outcome"; they empower an individual

37. See Edwards & Veale, Slave to the Algorithm, supra note 4, at 55–56.

38. See Sarkar, supra note 29 (defining local interpretability as trying to understand why the model made a particular decision in a single instance).

39. See Edwards & Veale, Slave to the Algorithm, supra note 4, at 58.

40. See Wachter et al., supra note 22, at 845 ("In the existing literature, 'explanation' typically refers to an attempt to convey the internal state or logic of an algorithm that leads to a decision. In contrast, counterfactuals describe a dependency on the external facts that led to that decision.").

41. See id. at 854, 881–82 (discussing implementation options); see also Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 Wash. L. Rev. 1, 28–29 (2014) (proposing a system to allow consumers to enter "hypothetical alterations" to their credit histories to see how the alterations affect their score).

42. See Wachter et al., supra note 22, at 848 ("Such considerations [relevant to which type of counterfactual you produce] may include the capabilities of the individual concerned, sensitivity, mutability of the variables involved in a decision, and ethical or legal requirements for disclosure."); id. at 851 (noting that one could offer "multiple diverse counterfactual explanations to the data subject"); see also Edwards & Veale, Slave to the Algorithm, supra note 4, at 63 (describing how counterfactual models can allow individuals to view and reflect upon the decisions about other users).

43. Wachter et al., supra note 22, at 851; id. at 860 ("[C]ounterfactuals bypass the substantial challenge of explaining the internal workings of complex machine learning systems," providing information that "is both easily digestible and practically useful for understanding the reasons for a decision, challenging them, and altering future behaviour for a better result.").

to more effectively navigate and challenge the process in a particular case.[44]

An alternative to these exogenous approaches is a category of xAI that attempts to explain (or "decompose") the model's reasoning. The most obvious way to do so is to reveal the source code for the machine learning model, but that approach will often prove unsatisfactory (because of the way machine learning works and because most people will not be able to understand the code).[45] More nuanced alternatives exist, however. One approach is to create a second system alongside the original "black box" model, sometimes called a "surrogate model."[46] A surrogate model works by analyzing featured input and output pairs but does not have access to the internal weights of the model itself.[47] For instance, scholars constructed a decision tree that effectively mirrored the computations of a black box model that predicted patients' risk for diabetes. The decision tree allowed computer scientists to track which factors (such as cholesterol level, nicotine dependence, and edema) the black box model weighed in making its risk assessments.[48] In a legal setting, this approach might entail creating a decision tree that accurately reconstructs the decisions of a self-driving car's black box algorithms in a product liability case, for example. These systems closely approximate the predictions made by an underlying model, while being interpretable.[49]

There are a host of ways in which machine learning algorithms will find their way into court in coming years. As a result, the courts themselves will be important actors in the machine learning ecosystem that is working to decide when, how, and in what form to develop xAI for algorithms. In specific cases, courts will need to consider a range of questions: Who is the audience for the explanation, and how simple or complex should the explanation be? How long should it take the user to understand the explanation?[50] What structure or form should the xAI take: lines of code, visual presentations, manipulable programs?[51] What

---

44. See Andrew D. Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085, 1120 (2018).

45. Kroll et al., supra note 36, at 638–39 (arguing that revealing source code is a misguided way of creating algorithmic accountability).

46. W. Andrew Pruett & Robert L. Hester, The Creation of Surrogate Models for Fast Estimation of Complex Model Outcomes, PLOS One (June 3, 2016), https://doi.org/10.1371/journal.pone.0156574 [https://perma.cc/GZ33-78MC].

47. Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, arXiv (Aug. 9, 2016), https://arxiv.org/pdf/1602.04938.pdf [https://perma.cc/8WN8-WQJF].

48. Osbert Bastani, Carolyn Kim & Hamsa Bastani, Interpreting Blackbox Models via Model Extraction, arXiv (Jan. 24, 2019), https://arxiv.org/pdf/1705.08504.pdf [https://perma.cc/L2K4-ZPVU].

49. See id.

50. See Doshi-Velez & Kim, supra note 20, at 7–8.

51. See Wachter et al., supra note 22, at 872 (noting that one could disclose the algorithm's source code, formula, weights, and full set of variables).

factors should the explanation focus on? When should xAI be model-centric and when should it be subject-centric? If there are trade secrets at issue, should the court review the algorithm in camera or request an independent peer review under a nondisclosure agreement?[52] More generally, what will constitute a "meaningful explanation"?[53] Judges are well positioned in this ecosystem to develop pragmatic approaches to xAI, even though they are not—indeed, *because* they are not—experts in machine learning technology.

To understand how these questions may arise concretely in practice, the next Parts identify and analyze two legal settings in which courts soon will need to make decisions about the types of xAI that are helpful—or that may even be legally required.

II. ALGORITHMS IN AGENCY RULEMAKING AND ADJUDICATION

Scholars have begun to consider the ways in which machine learning algorithms could advance the work of administrative agencies.[54] Cary Coglianese and David Lehr write, "[N]ational security and law enforcement agencies are starting to rely on machine learning . . . . [O]ther government agencies have also begun to explore uses of machine learning, revealing growing recognition of its promise across a variety of policy settings and at all levels of government."[55] Machine learning algorithms

52. See Coglianese & Lehr, Governance, supra note 12, at 49 (suggesting these two methods of review as ways to balance the need for transparency in administrative decision-making with the need to protect trade secrets).

53. In the national security context, judges frequently have to decide what types of classified explanations by the executive branch are sufficient. See Ashley S. Deeks, Secret Reason-Giving, 129 Yale L.J. (forthcoming 2019) (manuscript at 22–24) (on file with the *Columbia Law Review*) (discussing secret reason-giving in the context of foreign surveillance, asset freezes, state secrets, and the Freedom of Information Act).

54. See generally, e.g., Citron, supra note 19 (expressing concern that automated decisionmaking will undermine procedural safeguards and displace expert reasoning); Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 Geo. L.J. 1147 (2017) [hereinafter Coglianese & Lehr, Regulating by Robot] (arguing that the use of machine learning in administrative actions does not violate the nondelegation doctrine, due process, equal protection, or the reasoned explanation requirements of the Administrative Procedure Act); Mariano-Florentino Cuéllar, Cyberdelegation and the Administrative State, *in* Administrative Law from the Inside Out: Essays on Themes in the Work of Jerry L. Mashaw 134 (Nicholas R. Parillo ed., 2017) (highlighting the potential tradeoff between the increased precision of artificial intelligence decisionmaking and the risk of displacing agency deliberation about social welfare); Benjamin Alarie, Anthony Niblett & Albert Yoon, Regulation by Machine (Dec. 1, 2016) (unpublished manuscript), https://ssrn.com/abstract =2878950 (on file with the *Columbia Law Review*) (envisioning that agencies may deploy algorithms to predict how courts will decide administrative law cases).

55. Coglianese & Lehr, Regulating by Robot, supra note 54, at 1161; see also Coglianese & Lehr, Governance, supra note 12, at 3 ("Scholars and policy officials alike see increasing promise for the use of machine-learning algorithms by administrative agencies in a range of domestic policy areas.").

offer the potential to support agency rulemaking and also perhaps adjudications.[56] Virtually all of the scholars who have studied the issue anticipate that agencies' use of algorithms will only increase in coming years.[57]

Consider how agencies might deploy machine learning algorithms to facilitate rulemaking. Justice Mariano-Florentino Cuéllar writes, "Over time, neural networks and genetic algorithms will almost certainly inform judgments about the proper scope of a rule . . . ."[58] Coglianese and Lehr go further, envisioning truly autonomous rulemaking in areas such as SEC regulation of high-speed electronic trading or Treasury Department regulations that respond to real-time market changes suggestive of systemic risk.[59] They even envision multiagent systems, where machine learning algorithms would model different forecasts for different values to be traded off, and a separate machine learning system representing the agency would pick the model (and hence the rule) that maximizes the objective selected by humans.[60]

Another opportunity for the use of machine learning algorithms in the agency setting might be to parse and summarize voluminous public comments provided as part of notice and comment rulemaking.[61] Further, as noted above, agencies might turn to machine learning to help them conduct adjudications.[62] This could include using algorithms to predict pilot competence and grant pilot's licenses, forecast the effects of a proposed merger on competition, or decide disability claims.[63] None of these processes will exclude the human role entirely—at the very least, computer scientists must code agency "values" into the algorithms in the form of ones and zeros—but machine learning–driven rulemaking and adjudication may embody a host of decisional steps that are nontransparent and difficult to trace.

Courts are likely to confront all of these agency uses of algorithms. Under the Administrative Procedure Act (APA), courts generally may

---

56. Coglianese & Lehr, Regulating by Robot, supra note 54, at 1167 (discussing possible applications of machine learning in administrative rulemaking and adjudications).

57. See, e.g., Cuéllar, supra note 54, at 135 ("Reliance on computer programs to make administrative decisions — whether designed as conventional expert systems, more elaborate genetic or otherwise self-modifying algorithms, neural or 'deep learning' networks, or other machine learning mechanisms — will likely accelerate.").

58. Id. at 144.

59. Coglianese & Lehr, Regulating by Robot, supra note 54, at 1171–72.

60. Id. at 1174; Coglianese & Lehr, Governance, supra note 12, at 9–10; see also Cuéllar, supra note 54, at 17.

61. Mortazavi, supra note 7, at 207–08.

62. See Coglianese & Lehr, Governance, supra note 12, at 9 (noting that "the statistical tools that will facilitate adjudicating by algorithm already exist and are already being employed in analogous endeavors"); Cuéllar, supra note 54, at 137 (envisioning "sleek black boxes" administering "bureaucratic justice").

63. Coglianese & Lehr, Regulating by Robot, supra note 54, at 1170–71; Cuéllar, supra note 54, at 136–37.

review final agency actions.[64] For example, in the informal rulemaking context, courts may review agency factual determinations and discretionary decisions and set aside those actions that are arbitrary, capricious, or an abuse of discretion.[65] In that context, the Supreme Court requires an agency to "examine the relevant data and articulate a satisfactory explanation for its action including a 'rational connection between the facts found and the choice made.'"[66] More recently, the Court confirmed that the courts' role involves "examining the *reasons* for agency decisions—or, as the case may be, the absence of such reasons."[67] Agencies also are expected to address salient points raised in public comments.[68] That said, courts will give an agency particular deference when the agency is making predictions within its area of expertise that involve technical matters ("at the frontiers of science").[69]

Agency reason-giving thus plays an important role in defending the rules that agencies produce. Yet reason-giving can be complicated, if not confounded, by machine learning algorithms. An agency that has relied heavily on a machine learning algorithm prediction about the impact of a particular chemical on human health or about the population trajectory of a threatened species may need to share with the court the types of data it used, the type of machine learning model it used, the algorithm's error rate, and—possibly—the way the algorithm functioned to produce its prediction.[70] It is not yet clear precisely what courts will demand of agencies in this setting, or how agencies will respond.

Some scholars are relatively sanguine about the ease with which courts will adjust to the growing use of algorithms by agencies. For example, Coglianese and Lehr argue that current legal standards in administrative law do not demand anything close to transparency, that courts apply a deferential standard to agency rulemaking that relies on complex modelling, and that agencies will generally be able to meet that standard if they can show that the algorithm has performed as intended and achieves a justified objective.[71] Other scholars are more skeptical. Danielle Citron, for instance, worries that opaque algorithms impair meaningful judicial review because courts cannot see the rules that are

64. 5 U.S.C. § 702 (2012); Block v. Cmty. Nutrition Inst., 467 U.S. 340, 345 (1984).

65. 5 U.S.C. § 706(2)(A); see also id. § 553.

66. Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

67. Judulang v. Holder, 565 U.S. 42, 53 (2011) (emphasis added).

68. Perez v. Mortg. Bankers Ass'n, 135 S. Ct. 1199, 1203 (2015) ("An agency must consider and respond to significant comments received during the period for public comment.").

69. Balt. Gas & Elec. Co. v. NRDC, 462 U.S. 87, 103 (1983).

70. See Cuéllar, supra note 54, at 151–52 (noting that courts may want to understand how that process occurred and how users tested the system to ensure those values were fairly captured in the output).

71. See Coglianese & Lehr, Governance, supra note 12, at 35–36, 39, 47–49.

actually applied in a given case.[72] One possibility is that a court could reduce its level of deference to an agency decision when the agency deploys a black-box algorithm purchased from the private sector, because the court concludes that the agency is making a prediction based on private sector expertise, not its own.

Whether optimistic or pessimistic about the way courts will address these challenges, many scholars take comfort in xAI's possibilities. Justice Cuéllar contemplates that machine learning algorithms may help agencies withstand judicial scrutiny, because he assumes that their use could "conceivably yield greater transparency by making it easier to follow what precise considerations were used in driving a particular outcome."[73] This is only true, of course, if some form of xAI accompanies the algorithm. Likewise, Coglianese and Lehr admit that xAI will make it easier to defend an extensive use of machine learning algorithms by agencies. They highlight the "widening panoply of techniques that data scientists are developing to make learning algorithms more explainable" and note that even when the government uses algorithms to make individual-level predictions, "government agencies will likely have strategies available to them to provide individual-level explanations."[74] In short, xAI is likely to serve as an important linchpin in agencies' transition from human-dominated decisionmaking to machine-dominated decisionmaking. Yet none of these scholars focus on the direct role that the courts will play in affecting xAI itself.

As courts work through administrative law cases involving machine learning algorithms, they will play a significant role in shaping the xAI ecosystem. The extent to which courts seek information about the inputs, outputs, and reliability of agency algorithms or express interest in testing counterfactuals will give concrete form to current xAI discussions, which are happening largely in the abstract. Courts' approaches to agency algorithms in rulemaking settings might prompt developers to pursue exogenous xAI approaches, using model-centric explanations to defend the overall workings and reliability of the algorithm. Courts' approaches to agency algorithms in adjudication, in contrast, might lead developers to pursue decompositional approaches, using subject-centric explanations to defend the specific adjudicatory choices made. The healthy and growing set of xAI tools means that there is a range of choices from which to draw—and, as of now, no statutory guidance about xAI.

The prospect of courts being able to select the proper xAI tool for a given situation is a good thing, for all of the reasons that we celebrate the

72. Citron, supra note 19, at 1298.

73. Cuéllar, supra note 54, at 142, 153. Cuéllar seems less sanguine about situations in which xAI is not available, noting with concern that decisions could "be made on a basis phenomenologically different from what could easily be understood or even explained by human participants." Id. at 157.

74. Coglianese & Lehr, Governance, supra note 12, at 6, 55.

strengths of the common law.[75] Courts can move "cautiously and incrementally" as they sort out what types of xAI will be effective and realistically achievable in explaining different types of agency algorithms.[76] The courts will confront a set of concrete facts, and can, as a result, produce context-sensitive holdings that do not attempt to impose broad policies on xAI developments. Further, the courts here will build on existing case law that fleshes out the requirements of the APA, modestly adjusting that case law for situations in which the use of this new technology raises unanswered questions.[77]

xAI may also mitigate changes in the law that otherwise could result from the technological disruptions wrought by machine learning. For example, if courts become concerned about continuing to accord deference to agency decisionmakers who rely heavily on algorithms or worry about granting opaque algorithmic decisionmaking a "presumption of regularity,"[78] xAI may help assuage these concerns. Agencies may perceive the advantages of adopting xAI as a means to address judicial concerns ex ante and thus to minimize disadvantageous doctrinal changes.[79] Although common law xAI will, at least initially, offer less predictability than a federal xAI statute would, it can more easily take into account technological developments in xAI, and it can be more sensitive to what is both necessary and possible in a given setting.

III. CRIMINAL SENTENCING ALGORITHMS

In the administrative law setting, judges will sit as neutral reviewers of an agency's use of machine learning algorithms. In the criminal justice setting, judges themselves may be the ones using those algorithms.[80]

---

75. For a general discussion of the advantages of developing rules through the common law rather than by statute, see generally Jeffrey J. Rachlinski, Bottom-Up Versus Top-Down Lawmaking, 73 U. Chi. L. Rev. 933 (2006).

76. Neal Devins & David Klein, The Vanishing Common Law Judge?, 165 U. Pa. L. Rev. 595, 630 (2017) ("[A] series of such decisions will yield a refined principle or rule, resulting in fewer injustices and inefficiencies than would result if the first court's approach were followed religiously in all similar cases.").

77. See Aharon Barak, The Judge in a Democracy 156 (2006) (noting that expansive judicial case law hangs on narrow statutory hooks and that judges develop common law within the frameworks of statutes).

78. Cuéllar, supra note 54, at 154–56 (discussing varying levels of deference depending on the seniority of an agency decisionmaker); id. at 158 (asking whether courts should revisit the presumption of regularity to "ensure that decisionmakers recognize the risks of relying on automated analytical techniques they do not entirely understand").

79. David A. Strauss, Common Law Constitutional Interpretation, 63 U. Chi. L. Rev. 877, 895 (1996) ("Everyone recognizes that law . . . is in substantial part about following precedent and otherwise maintaining continuity with the past.").

80. Many describe these tools as employing machine learning, though the companies developing the algorithms often invoke "trade secrets," which prevents both defendants and judges from knowing precisely how the algorithms function. See, e.g., Ric Simmons, Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal

Here, too, they may—and should—demand certain explanations for how those algorithms work, to ensure that the algorithms are trustworthy and fair. Defense counsel also are likely to press prosecutors and algorithm developers for explanations, which in turn may stimulate judges to do the same.

Officials in the criminal justice system often need to predict how likely a person is to commit a dangerous act.[81] In the bail context, for example, judges must assess whether individuals are likely to return to court for trial and whether they are likely to engage in criminal acts if they are not kept in detention before trial.[82] When sentencing a defendant, the judge considers in part how likely it is that the person will reoffend if released after a particular period.[83] These data-driven algorithms have the potential to help decisionmakers avoid relying on intuition and personal biases and to allow governments to reduce jail populations without affecting public safety.[84] As a result, the criminal justice system has seen a widespread and growing use of predictive algorithms in the bail, sentencing, and parole contexts.[85]

Notwithstanding their potential, these algorithms have come under intense criticism. Some critiques focus on the idea that the data on which

---

Justice System, 2016 Mich. St. L. Rev. 947, 997 (discussing machine learning criminal justice algorithms and noting that judges need to understand the factors that the algorithm used and the historical accuracy of the algorithm's results). Even if criminal justice algorithms currently do not use advanced machine learning, scholars have argued that they soon will. See, e.g., Richard Berk, Criminal Justice Forecasts of Risk: A Machine Learning Approach 110–11 (2012) ("Actuarial methods are changing rapidly. Forecasts increasingly exploit enormous datasets that are routinely available in real time."); Richard Berk & Jordan Hyatt, Machine Learning Forecasts of Risk to Inform Sentencing Decisions, 27 Fed. Sent'g Rep. 222, 222 (2015) (arguing that machine learning forecasting methods will produce more accurate forecasts than more traditional regression analyses).

81. See Ashley Deeks, Predicting Enemies, 104 Va. L. Rev. 1529, 1538 (2018).

82. Samuel R. Wiseman, Fixing Bail, 84 Geo. Wash. L. Rev. 417, 420–21 (2016) (discussing dangerousness and flight risk as two key considerations in bail decisionmaking).

83. See Sonja B. Starr, Evidence-Based Sentencing and the Scientific Rationalization of Discrimination, 66 Stan. L. Rev. 803, 809 (2014) (noting that evidence-based sentencing is designed to assist judges in pursuit of sentencing objectives that are centered on reducing the defendant's future crime risk).

84. Sam Corbett-Davies, Sharad Goel & Sandra González-Bailón, Even Imperfect Algorithms Can Improve the Criminal Justice System, N.Y. Times (Dec. 20, 2017), https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html [https://perma.cc/E7BQ-EDCK].

85. See Algorithms in the Criminal Justice System, Elec. Privacy Info. Ctr., https://epic.org/algorithmic-transparency/crim-justice/ [https://perma.cc/3BES-5P2L] (last visited Aug. 3, 2019) (listing different states' uses of algorithmic tools for sentencing, probation, and parole decisions). For a recent example of a state's decision to require the use of algorithms in the bail setting, see Dave Gershgorn, California Just Replaced Cash Bail with Algorithms, Quartz (Sept. 4, 2018), https://qz.com/1375820/california-just-replaced-cash-bail-with-algorithms/ [https://perma.cc/JRM4-RX5Z].

computer scientists train the algorithms are racially biased.[86] Others argue that the algorithms are no better at predicting recidivism than are humans who lack criminal justice expertise.[87] Finally, many object to the fact that the algorithms' structure, contents, and testing are opaque.[88] This latter concern came to a head in *State v. Loomis*, a case in which a defendant challenged the judge's use of a sentencing algorithm called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) that had categorized him as posing a "high risk of recidivism."[89] The defendant argued that the court's use of the risk assessment violated his due process rights, in part because he was not able to assess COMPAS's accuracy.[90]

The Wisconsin Supreme Court upheld his sentence.[91] Nevertheless, the majority and a concurring Justice expressed caution about the use of opaque sentencing algorithms. The majority required future presentence investigation reports to contain warnings about the limitations of COMPAS, in order to avoid potential due process violations.[92] In concurrence, Justice Shirley Abrahamson stated that "this court's lack of understanding of COMPAS was a significant problem in the instant case."[93] She noted that "making a record, including a record explaining consideration of the evidence-based tools and the limitations and strengths thereof, is part of the long-standing, basic requirement that a circuit court explain its exercise of discretion at sentencing."[94] Even the U.S. government brief, filed to oppose the defendant's petition for writ of certiorari in the U.S. Supreme Court, conceded that "[s]ome uses of an undisclosed risk-assessment algorithm might raise due process concerns—if, for example, a defendant is denied access to the factual inputs

86. See, e.g., Andrew Guthrie Ferguson, The Rise of Big Data Policing 131–32 (2017) ("Police data remains colored by explicit and implicit bias.").

87. See, e.g., State v. Loomis, 881 N.W.2d 749, 775 n.3 (Wis. 2016) (Abrahamson, J., concurring) (acknowledging that studies differ on the accuracy of the recidivism scores of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)); Dressel & Farid, supra note 18, at 1 (concluding that nonexperts are "as accurate and fair as COMPAS at predicting recidivism" and noting the inefficacy of its more sophisticated features).

88. See, e.g., Andrea Roth, Trial by Machine, 104 Geo. L.J. 1245, 1270 (2016) (noting concerns about obscuring hidden subjectivities and errors in criminal justice algorithms); Rebecca Wexler, Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System, 70 Stan. L. Rev. 1343, 1349–50 (2018) [hereinafter Wexler, Life, Liberty] ("Developers often assert that details about how their tools function are trade secrets.").

89. *Loomis*, 881 N.W.2d at 753, 755.

90. Id. at 757.

91. Id. at 772.

92. See id. at 769–70.

93. Id. at 774 (Abrahamson, J., concurring) (noting that, "[a]t oral argument, the court repeatedly questioned both the State's and defendant's counsel about how COMPAS works" but that "[f]ew answers were available").

94. Id. at 775.

about his criminal and personal history, or if his risk scores form part of a sentencing 'matrix' or establish a 'presumptive' term of imprisonment."[95]

Perhaps not surprisingly, some jurisdictions are shifting away from opaque commercial algorithms such as the one used in *Loomis* and toward algorithms that use public data and publicly available source codes.[96] Even those jurisdictions may retain an interest in xAI, because source codes alone are typically not self-explanatory. In particular, though, it is the courts in jurisdictions that continue to rely on opaque predictive algorithms that may—and should—become more aggressive in demanding xAI. There are a host of reasons why they might do so. First, both federal and state courts often face statutory requirements to justify the sentences they impose.[97] This allows the public to evaluate the reasonableness of the sentence and see what factual findings the judge made; it also permits review by appellate courts.[98] Judges who rely in part on sentencing algorithms might believe that they need to understand the parameters of the algorithms to articulate the reasons for using them. Second, judges serve as a bulwark to ensure accuracy and fairness in sentencing; demanding xAI will help judges evaluate whether the algorithms meet that standard or contain significant errors.[99] Third, judges might demand xAI to ensure the institutional integrity of the courts, which is undercut if courts use unreliable sources of guidance. Fourth, judges

---

95. Brief for the United States as Amicus Curiae at 18, Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (No. 16-6387), 2017 WL 2333897.

96. See, e.g., Elaine Angelino, Nicholas Larus-Stone, Daniel Alibi, Margo Seltzer & Cynthia Rudin, Learning Certifiably Optimal Rule Lists for Categorical Data, J. Machine Learning Res., June 2018, at 1, 2, http://www.jmlr.org/papers/volume18/17-716/17-716.pdf [https://perma.cc/LA46-XA6R] (developing an open-source machine learning model that accurately predicts a person's likelihood of rearrest); Creating a Fairer Pretrial System, Arnold Ventures (Dec. 1, 2017), https://www.arnoldventures.org/stories/creating-a-fairer-pretrial-system [https://perma.cc/TP87-G2L6] (stating that roughly forty jurisdictions have adopted or are in the process of implementing a pretrial risk assessment tool called the Public Safety Assessment); Public Safety Assessment: Risk Factors and Formula, Pub. Safety Assessment, https://www.psapretrial.org/about/factors [https://perma.cc/W3P3-4UJN] (last visited Aug. 3, 2019) (disclosing the risk factors and formula undergirding the Public Safety Assessment).

97. See, e.g., 18 U.S.C. § 3553(c) (2012) (requiring that the "court, at the time of sentencing, shall state in open court the reasons for its imposition of the particular sentence," including specific reasons for sentencing outside the range); Heather Young Keagle, Appellate Div., N.J. Superior Court, Manual on New Jersey Sentencing Law 11 (rev. 2019), https://njcourts.gov/attorneys/assets/attyresources/manualsentencinglaw.pdf [https://perma.cc/UQ77-7U32] ("At the time of sentencing, the court must 'state reasons for imposing such sentence including . . . the factual basis supporting a finding of particular aggravating or mitigating factors affecting sentence.'" (quoting State v. Fuentes, 85 A.3d 923, 932 (N.J. 2014))).

98. Toby D. Slawsky, The Importance of Statements of Reasons in Guideline Sentencing, 28 Fed. Sent'g Rep. 174, 174 (1990).

99. See Rebecca Wexler, Code of Silence, Wash. Monthly (June/July/Aug. 2017), https://washingtonmonthly.com/magazine/junejulyaugust-2017/code-of-silence/ [https://perma.cc/5PHM-6SPV] (describing how opacity can conceal flaws in criminal justice algorithms).

might insist on some form of xAI because they are worried about being reversed on appeal for relying on a flawed or poorly understood tool. Finally, judges may demand xAI at the behest of defense counsel, to facilitate adversarial challenges and promote procedural fairness.

What form is xAI likely to take here? In the administrative law context, the audiences for the xAI (executive agencies, judges, and corporate or interest-group plaintiffs) are likely to be sophisticated actors. In the criminal justice setting, there are three main audiences: (1) judges, (2) defendants, and (3) their lawyers. Some judges and defense counsel will be sophisticated repeat players, but the defendants themselves are likely to have little experience with algorithms—and indeed judges themselves will have different levels of experience with tools such as regression analyses.[100] Judges might be more interested in model-centric explanations, while recognizing that defendants may need subject-centric xAI. Both audiences might benefit from being able to run counterfactuals through the system as well. Judges will have to decide whether to demand one or the other forms of xAI—or both.

Judges will encounter pushback from the producers of proprietary algorithms, who have resisted revealing information about the workings of their algorithms on the basis of trade secrets claims.[101] There are ways to protect such secrets, however, including by issuing protective orders.[102] Further, it might be possible to build a surrogate model of the sentencing algorithm that sheds light on its functioning without forcing the producer to reveal trade secrets. In those cases, xAI may play an important role in counterbalancing trade secrets claims such as those in play in *Loomis*.

There is another, less obvious advantage to judges' use of xAI in the criminal justice setting. A persistent concern about machine learning algorithms is that they produce "automation bias"—a tendency to unduly accept a machine's recommendation.[103] Putting xAI in front of judges

100. This level of experience presumably will increase over time as more machine learning tools find their way into the practice of law.

101. See Wexler, Life, Liberty, supra note 88, at 1349–50.

102. See id. at 1409–10.

103. Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators, 19 J. Am. Med. Informatics Ass'n 121, 121 (2012); Raja Parasuraman & Dietrich H. Manzey, Complacency and Bias in Human Use of Automation: An Attentional Integration, 52 Hum. Factors 381, 397 (2010) (concluding that both expert and inexpert participants suffer from complacency and bias in their interactions with automated systems).

may lead them to question an algorithm's conclusions in a way that helps them avoid succumbing to automation bias.[104]

In light of the various benefits of xAI and a growing number of xAI tools in the toolbox, one puzzle is why courts have not already begun to insist on xAI when confronted with machine learning algorithms in criminal justice settings. Is it because the idea of xAI is nascent? Because the use of algorithms in the criminal justice context is only now starting to receive widespread scrutiny and criticism? Because of trade secrets hurdles? Or because the courts themselves currently lack the confidence to understand and use xAI?[105] It is likely a combination of all of these factors. However, as the use of machine learning and, concomitantly, xAI spreads, the courtroom is a fertile ground in which to connect xAI to real-world challenges.

CONCLUSION

Agency rulemaking and criminal justice are hardly the only areas of law in which courts will confront machine learning algorithms. Other possible legal contexts include product liability litigation involving self-driving cars or the internet of things,[106] litigation challenging school districts' use of algorithms for teacher evaluations,[107] malpractice litigation against doctors who rely on medical algorithms for diagnoses,[108] individual challenges to governmental decisions to freeze people's assets based on algorithmic recommendations,[109] defendants' challenges to police

104. See Matt O'Brien & Dake Kang, AI in the Court: When Algorithms Rule on Jail Time, Phys.org (Jan. 31, 2018), https://phys.org/news/2018-01-ai-court-algorithms.html [https://perma.cc/84R8-B2X4] (discussing automation bias in judges); see also id. (quoting a Northwestern University computer scientist as arguing that judges need "boxes that give [them] answers and explanations and ask [them] if there's anything [they] want to change").

105. Lilian Edwards & Michael Veale, Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?, 16 IEEE Security & Privacy 46, 53 (2018) [hereinafter Edwards & Veale, Enslaving the Algorithm] ("It seems quite likely that courts will be reluctant to become activists about disclosures of source code, let alone algorithmic training sets and models, until they feel more confident of their ability to comprehend and use such evidence—which may take some time.").

106. See Ian Bogost, Can You Sue a Robocar?, Atlantic (Mar. 20, 2018), https://www.theatlantic.com/technology/archive/2018/03/can-you-sue-a-robocar/556007/ [https://perma.cc/84LK-AQ9V] (discussing the legal implications of accidents caused by self-driving cars)

107. See Coglianese & Lehr, Governance, supra note 12, at 37–38 (discussing litigation by teachers over a school district's use of algorithms to rate teachers' performance).

108. See Shailin Thomas, Artificial Intelligence, Medical Malpractice, and the End of Defensive Medicine, Bill of Health (Jan. 26, 2017), http://blog.petrieflom.law.harvard.edu/2017/01/26/artificial-intelligence-medical-malpractice-and-the-end-of-defensive-medicine/ [https://perma.cc/7AU6-P3QJ] (discussing the interaction between malpractice litigation and use of machine learning algorithms).

109. See Cuéllar, supra note 54, at 144 (describing the potential use of algorithms to include decisions to freeze individuals' assets).

stops based on the use of "automated suspicion" algorithms,[110] government requests for Foreign Intelligence Surveillance Act orders based on algorithmic predictions about who is a foreign agent,[111] or challenges to algorithm-driven forensic testing.[112] These cases might implicate questions of substantive or procedural due process,[113] require "arbitrary and capricious" review, or force courts to decide whether to allow expert testimony about how a given algorithm functions.[114] Some scholars have proposed the kinds of explanations courts should seek in certain types of cases,[115] but the rubber will hit the road when the courts themselves decide what is needed. Using the tools of the common law, judges can and will productively drive the advancement and fine-tuning of xAI. When deciding xAI-related questions, courts should focus on two principles that can further public law values: maximizing xAI's ability to help identify errors and biases within the algorithm, and aligning the form of xAI in a given case with the needs of the relevant audiences.

The interest in xAI is not simply a U.S. phenomenon. The European Union's General Data Protection Regulation (GDPR), which applies to

110. See Michael Rich, Automated Suspicion Algorithms and the Fourth Amendment, 164 U. Pa. L. Rev. 871, 875–76 (2016) ("Machine learning provides a way to go one step further and use data to identify likely criminals among the general population.").

111. See 50 U.S.C. § 1805(a) (2012) (outlining the findings necessary for a judge to enter an ex parte order approving electronic surveillance); Jim Baker, Counterintelligence Implications of Artificial Intelligence—Part III, Lawfare (Oct. 10, 2018), https://www.lawfareblog.com/counterintelligence-implications-artificial-intelligence-part-iii [https://perma.cc/M6D3-KNW9] (discussing the use of AI in counterintelligence).

112. See Symposium on Forensic Expert Testimony, *Daubert*, and Rule 702, 86 Fordham L. Rev. 1463, 1513–15 (2018) (presenting discussion of the application of a *Daubert*-style test to algorithm-driven DNA testing by Professor Erin Murphy of the Advisory Committee on Evidence Rules); Drew Harwell, Oregon Became a Testing Ground for Amazon's Facial-Recognition Policing. But What if Rekognition Gets It Wrong?, Wash. Post (Apr. 30, 2019), https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/ (on file with the *Columbia Law Review*) (discussing how lawyers are preparing to litigate the admissibility of facial-recognition system evidence in court).

113. See Coglianese & Lehr, Governance, supra note 12, at 38–43.

114. Courts already have confronted *Daubert* issues in the face of a juvenile sentencing algorithm and algorithm-assisted discovery processes. See Moore v. Publicis Groupe, 287 F.R.D. 182, 182–84, 188–89 (S.D.N.Y. 2012) (concluding that using sophisticated algorithmic tools to search for electronically stored information is acceptable and that *Daubert* did not apply at the search stage of discovery because the documents were not yet being introduced as evidence); AI Now Inst., Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems 1, 13–14 (2018), https://ainowinstitute.org/litigatingalgorithms.pdf [https://perma.cc/Y5V4-NVRF] (noting that counsel persuaded the judge that past studies had not sufficiently validated a juvenile sentencing algorithm).

115. See, e.g., Kroll et al., supra note 36, at 637 (describing verification methods to ensure algorithms' procedural regularity).

countries and companies in the European Union, contains provisions[116] requiring what some have termed a "right to an explanation."[117] Some scholars have interpreted the GDPR to require data controllers who make decisions about individuals based "solely on automated processing" to provide those individuals with meaningful information about the logic involved in that automated decisionmaking.[118] But it remains unclear precisely what the GDPR requires and what steps states and companies must take to meet those requirements. Other countries have enacted their own domestic "explainability" requirements. France, for instance, in its Digital Republic Act, gives individuals a right to an explanation for administrative algorithmic decisions made about those individuals.[119] That law requires the administrative decisionmaker to provide a range of information about the "degree and the mode of contribution of the algorithmic processing to the decision making," including what data were processed, what the system's parameters were, and how the algorithm weighted factors.[120] Thus, U.S. common law decisions about xAI are likely to be of interest not only to U.S. federal and state judges but to foreign judges and administrative officials as well.

Nor are courts the only government actors that must navigate the costs and benefits of xAI. Congress may demand and shape the use of xAI across industries or within government via legislation, and it may also demand the use of xAI in briefings by executive agencies, including the intelligence community. Any statute regulating the use of xAI, however, necessarily must be crafted at a high level of generality. That statute may capture the basic values that Congress wants xAI to advance, but such a statute may struggle to endure in this quickly shifting landscape. Further, the likelihood that Congress will be able to act in this space is limited, if its recent actions on complicated technology issues are any guide.[121]

---

116. Regulation (EU) 2016/679 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 and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), arts. 15(1)(h), 22, 2016 O.J. (L 119) 43, 46.

117. See, e.g., Bryce Goodman & Seth Flaxman, European Union Regulations on Algorithmic Decision Making and a "Right to Explanation," AI Mag., Fall 2017, at 50.

118. See, e.g., id. at 55. But see Sandra Wachter, Brett Mittelstadt & Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, 7 Int'l Data Privacy L. 76, 77, 79–90 (2017) (arguing that the GDPR implements a limited "right to be informed" rather than a "right to explanation").

119. Edwards & Veale, Enslaving the Algorithm, supra note 105, at 48–49.

120. Id. at 48.

121. See Ashley Deeks, Facebook Unbound?, 105 Va. L. Rev. Online 1, 6–8 (2019), http://www.virginialawreview.org/sites/virginialawreview.org/files/01.%20Final%20Deeks.pdf [https://perma.cc/QHE5-SJ7Q] ("[Congress] has failed in its efforts to legislate on the use of encryption, election security . . . , 'hacking back,' and drone safety, and it has not tried to regulate facial-recognition software. Efforts to impose federal data-privacy laws on companies are just getting underway.").

Common law xAI thus offers real promise as we head deeper into the age of algorithms. Courts will only be able to work xAI issues at the edges, looking across legal categories to draw on xAI developments in different doctrinal areas, but that work—and the response to that work by the creators and users of machine learning algorithms—may get us where we need to be.