RULEMAKING AND INSCRUTABLE AUTOMATED DECISION TOOLS

*Katherine J. Strandburg**

Complex machine learning models derived from personal data are increasingly used in making decisions important to peoples' lives. These automated decision tools are controversial, in part because their operation is difficult for humans to grasp or explain. While scholars and policymakers have begun grappling with these explainability concerns, the debate has focused on explanations to decision subjects. This Essay argues that explainability has equally important normative and practical ramifications for decision-system design. Automated decision tools are particularly attractive when decisionmaking responsibility is delegated *and* distributed *across multiple actors to handle large numbers of cases. Such decision systems depend on explanatory flows among those responsible for setting goals, developing decision criteria, and applying those criteria to particular cases. Inscrutable automated decision tools can disrupt all of these flows.*

This Essay focuses on explanation's role in decision-criteria development, which it analogizes to rulemaking. It analyzes whether, and how, decision tool inscrutability undermines the traditional functions of explanation in rulemaking. It concludes that providing information about the many aspects of decision tool design, function, and use that can be explained can perform many of those traditional functions. Nonetheless, the technical inscrutability of machine learning models has significant ramifications for some decision contexts. Decision tool inscrutability makes it harder, for example, to assess whether decision criteria will generalize to unusual cases or new situations and heightens communication and coordination barriers between data scientists and subject matter experts. The Essay concludes with some suggested approaches for facilitating explanatory flows for decision-system design.

INTRODUCTION

Machine learning models derived from large troves of personal data are increasingly used in making decisions important to peoples' lives.[1]

* Alfred Engelberg Professor of Law and Faculty Director of the Information Law Institute, New York University School of Law. I am grateful for excellent research assistance from Madeline Byrd and Thomas McBrien and for summer research funding from the Filomen D. Agostino and Max E. Greenberg Research Fund.

1. See Max Fisher & Amanda Taub, Is the Algorithmification of the Human Experience a Good Thing?, N.Y. Times: The Interpreter (Sept. 6, 2018), https://static.nytimes.com/email-content/INT_5362.html (on file with the *Columbia Law Review*).

These tools have stirred both hopes of improving decisionmaking by avoiding human shortcomings and concerns about their potential to amplify bias and undermine important social values.[2] It is often hard for humans to grasp or explain how or why machine-learning-based models map input features to output predictions because they often combine large numbers of input features in complicated ways.[3] This inherent inscrutability[4] has drawn the attention of data scientists,[5] legal scholars,[6] policymakers,[7] and others[8] to the explainability problem.

---

2. Compare Susan Wharton Gates, Vanessa Gail Perry & Peter M. Zorn, Automated Underwriting in Mortgage Lending: Good News for the Underserved?, 13 Housing Pol'y Debate 369, 370 (2002) (finding that automated underwriting systems more accurately predict mortgage default than humans and result in higher approval rates for underserved applicants), and Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, Human Decisions and Machine Predictions, 133 Q.J. Econ. 237, 268 (2017) (showing that applying machine learning algorithms to pretrial detention decisions could reduce the jailed population by forty-two percent without an increase in crime), with Jeffrey Dastin, Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women, Reuters (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/6SA7-R35L] ("Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.").

3. See, e.g., Finale Doshi-Velez & Mason Kortz, Accountability of AI Under the Law: The Role of Explanation 9–10 (2017), https://cyber.harvard.edu/publications/2017/11/AIExplanation [https://perma.cc/AQ5V-582E]; Jenna Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, Big Data & Soc'y, Jan.–June 2016, at 1, 3; Aaron M. Bornstein, Is Artificial Intelligence Permanently Inscrutable?, Nautilus (Sept. 1, 2016), http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable [https://perma.cc/B562-NCUN]; see also Info. Law Inst. at N.Y. Univ. Sch. of Law with Foster Provost, Krishna Gummadi, Anupam Datta, Enrico Bertini, Alexandra Chouldechova, Zachary Lipton & John Nay, Modes of Explanation in Machine Learning: What Is Possible and What Are the Tradeoffs?, *in* Algorithms and Explanations (Apr. 27, 2017), https://youtu.be/U0NsxZQTktk (on file with the *Columbia Law Review*).

4. See Andrew D. Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085, 1094 (2018) (defining "inscrutability" in this context as "a situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension").

5. See generally Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning, *in* 2018 IEEE 5th International Conference on Data Science and Advanced Analytics 1 (2018) (cataloging various ways to define and evaluate interpretability in machine learning); Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, *in* 2018 IEEE 5th International Conference on Data Science and Advanced Analytics 80 (2018) ("While interpretability is a substantial first step, these mechanisms need to *also* be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited."); Zachary C. Lipton, The Mythos of Model Interpretability, ACMQueue (July 17, 2018), https://queue.acm.org/detail.cfm?id=3241340 [https://perma.cc/CZH3-S9JG] (discussing "the feasibility and desirability of different notions of interpretability" in machine learning).

6. 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, 19–22 (2017); Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten,

This discourse has focused primarily on explanations provided to decision subjects. For example, the European Union's General Data Protection Regulation (GDPR) arguably gives decision subjects a "right to explanation,"[9] reflecting the common premise that "[t]o justify a decision-making procedure that involves or is constituted by a machine learning model, *an individual subject to that decision-making procedure* requires an explanation of how the machine learning model works."[10] Some scholars have criticized this focus, emphasizing the importance of public

---

Joel R. Reidenberg, David G. Robinson & Harlan Yu, Accountable Algorithms, 165 U. Pa. L. Rev. 633, 636–42 (2017); Selbst & Barocas, supra note 4; Andrew D. Selbst, Response, A Mild Defense of Our New Machine Overlords, 70 Vand. L. Rev. En Banc 87, 88–89 (2017), https://cdn.vanderbilt.edu/vu-wp0/wp content/uploads/sites/278/2017/05/23184939/A-Mild-Defense-of-Our-New-Machine-Overlords.pdf [https://perma.cc/MCW7-X89L]; Tal Z. Zarsky, Transparent Predictions, 2013 U. Ill. L. Rev. 1503, 1506–09; Robert H. Sloan & Richard Warner, When Is an Algorithm Transparent?: Predictive Analytics, Privacy, and Public Policy, IEEE Security & Privacy, May/June 2018, at 18, 18.

7. See, e.g., Algorithmic Accountability Act of 2019, S. 1108, 116th Cong. (2019).

8. See, e.g., Reuben Binns, Algorithmic Accountability and Public Reason, 31 Phil. & Tech. 543, 543–45 (2018); Tim Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences, 267 Artificial Intelligence 1, 1–2 (2019); Brent Mittelstadt, Chris Russell & Sandra Wachter, Explaining Explanations in AI, *in* FAT*'19 at 279, 279 (2019); Deirdre K. Mulligan, Daniel N. Kluttz & Nitin Kohli, Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions, *in* After the Digital Tornado (Kevin Werbach ed., forthcoming 2020) (manuscript at 1–2), https://ssrn.com/abstract=3311894 (on file with the *Columbia Law Review*); Sandra Wachter, Brent Mittelstadt & Chris Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, 31 Harv. J.L. & Tech. 841, 842–44 (2018); Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter & Luciano Floridi, The Ethics of Algorithms: Mapping the Debate, Big Data & Soc'y, July–Dec. 2016.

9. The GDPR requires that data subjects be informed of "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." Commission Regulation 2016/679, art. 13(2)(f), 2016 O.J. (L 119) 1.

It further provides a limited "right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." Id. art. 22(1). For the debate about what the GDPR's requirements entail, see, e.g., Bryan Casey, Ashkon Farhangi & Roland Vogl, Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise, 34 Berkeley Tech. L.J. 143, 153–68 (2019); Talia B. Gillis & Josh Simons, Explanation < Justification: GDPR and the Perils of Privacy, Pa. J.L. & Innovation (forthcoming 2019) (manuscript at 2–4), https://ssrn.com/abstract=3374668 (on file with the *Columbia Law Review*); Margot E. Kaminski, The Right to an Explanation, Explained, 34 Berkeley. Tech. L.J. 189, 192–93 (2019); Andrew D. Selbst & Julia Powles, Meaningful Information and the Right to Explanation, 7 Int'l Data Privacy L. 233, 233–34 (2017); Michael Veale & Lilian Edwards, Clarity, Surprises, and Further Questions in the Article 29 Working Part Draft Guidance on Automated Decision-Making and Profiling, 34 Computer L. & Security Rev. 398, 398–99 (2018); Wachter et al., supra note 8, at 861–65; Andy Crabtree, Lachlan Urquhart & Jiahong Chen, Right to an Explanation Considered Harmful (Apr. 8, 2019) (unpublished manuscript), https://ssrn.com/abstract=3384790 (on file with the *Columbia Law Review*).

10. Gillis & Simons, supra note 9 (manuscript at 11) (emphasis added).

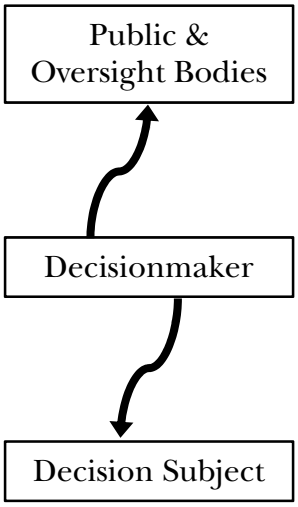
accountability.[11] Talia Gillis and Josh Simons, for example, contrast "[t]he focus on individual, technical explanation . . . driven by an uncritical bent towards transparency" with their argument that "[i]nstitutions should justify their choices about the design and integration of machine learning models not to individuals, but to empowered regulators or other forms of public oversight bodies."[12] Taken together, these threads suggest the view of explanatory flows in decisionmaking illustrated in Figure 1, in which decisionmakers justify their choices by explaining case-by-case outcomes to decision subjects and separately explaining design choices regarding automated decision tools to the public and oversight bodies.

---

11. For the most part, this emphasis is recent. See, e.g., Doshi-Velez & Kortz, supra note 3, at 3–9 (describing the explanation system's role in public accountability); Hannah Bloch-Wehba, Access to Algorithms, 88 Fordham L. Rev. (forthcoming 2019) (manuscript at 4–9), https://ssrn.com/abstract=3355776 (on file with the *Columbia Law Review*) ("These features . . . have prompted calls for new mechanisms of transparency and accountability in the age of algorithms."); Robert Brauneis & Ellen P. Goodman, Algorithmic Transparency for the Smart City, 20 Yale J.L. & Tech. 103, 132 (2018) ("Such accountability requires not *perfect* transparency . . . but . . . *meaningful* transparency."); Gillis & Simons, supra note 9 (manuscript at 11–12) ("Explanations of machine learning models are certainly not sufficient for many of the most important forms of justification in modern democracies . . . ."); Selbst & Barocas, supra note 4, at 1087 ("[F]aced with a world increasingly dominated by automated decision-making, advocates, policymakers, and legal scholars would call for machines that can explain themselves."); Jennifer Cobbe, Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making, Legal Stud. (July 9, 2019), https://www.cambridge.org/core/journals/legal-studies/article/administrative-law-and-the-machines-of-government-judicial-review-of-automated-publicsector-decisionmaking/09CD6B470DE4ADCE3EE8C94B33F46FCD/core-reader (on file with the *Columbia Law Review*) ("Legal standards and review mechanisms which are primarily concerned with decision-making processes, which examine how decisions were made, cannot easily be applied to opaque, algorithmically-produced decisions."). But, for a truly pathbreaking consideration of these issues, see Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249, 1258 (2008) ("This technological due process provides new mechanisms to replace the procedural regimes that automation endangers.").

12. Gillis & Simons, supra note 9 (manuscript at 6–12); 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, 708–09 (2017) (emphasizing the many choices involved in implementing a machine learning model and the different sorts of explanations that could be made).

FIGURE 1: SCHEMATIC OF EXPLANATORY FLOWS IN A SIMPLE DECISION SYSTEM

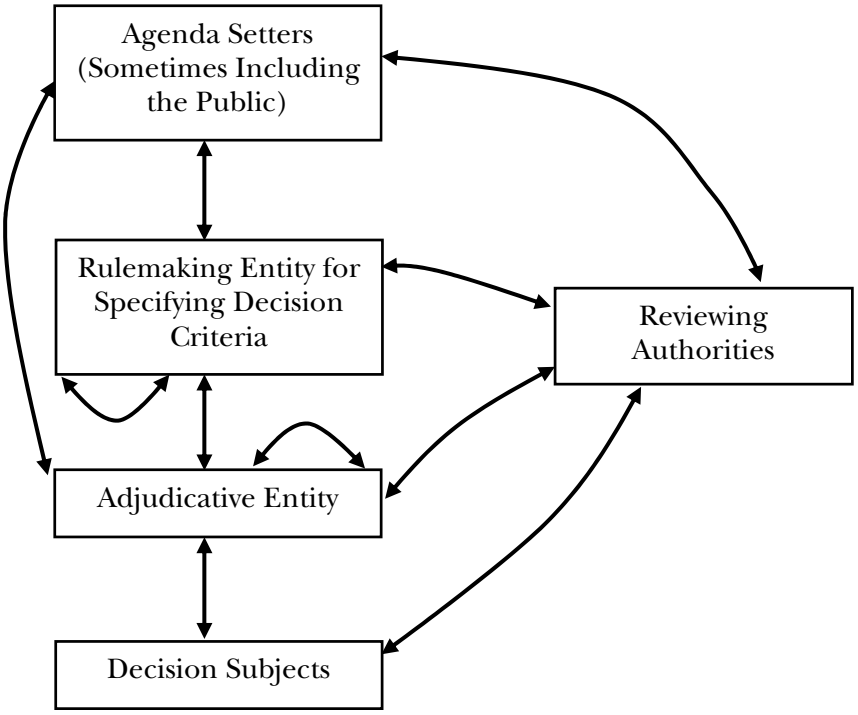


Many real-world decision systems require significantly more complex explanatory flows, however, because decisionmaking responsibility is *delegated* and *distributed* across multiple actors to handle large numbers of cases. Delegated, distributed decision systems commonly include agenda setters, who determine the goals and purposes of the systems; rulemakers tasked with translating agenda setters' goals into decision criteria; and adjudicators, who apply those criteria to particular cases.[13] In democracies, the ultimate agenda setter for government decisionmaking is the public, often represented by legislatures and courts. The public also has a role in agenda setting for many private decision systems, such as those related to employment and credit.[14] Figure 2 illustrates the explanatory flows required by a delegated, distributed decision system.

13. The terms "adjudication" and "rulemaking" are borrowed, loosely, from administrative law. See 5 U.S.C. § 551 (2012); see also, e.g., id. §§ 553–557. The general paradigm in Figure 2 also describes many private decision systems.

14. See infra section III.B.2.

FIGURE 2: SCHEMATIC OF EXPLANATORY FLOWS IN A DELEGATED, DISTRIBUTED DECISION SYSTEM



Delegation and distribution of decisionmaking authority, while often necessary and effective for dealing with agenda setters' limited time and expertise, proliferate explanatory information flows. *Delegation*, whether from the public or a private agenda setter, creates the potential for principal–agent problems and hence the need for accountability mechanisms.[15] Explanation requirements, including a duty to inform principals of facts that "the principal would wish to have" or "are material to the agent's duties," are basic mechanisms for ensuring that agents are accountable to principals.[16] *Distribution* of responsibility multiplies these principal–agent concerns, while adding an underappreciated layer of

15. See Kathleen M. Eisenhardt, Agency Theory: An Assessment and Review, 14 Acad. Mgmt. Rev. 57, 61 (1989) ("The agency problem arises because (a) the principal and the agent have different goals and (b) the principal cannot determine if the agent has behaved appropriately."); see also Gillis & Simons, supra note 9 (manuscript at 6–10) (arguing for a principal–agent framework of accountability in considering government use of machine learning).

16. Restatement (Third) of Agency § 8.11 (Am. Law Inst. 2005).

explanatory flows necessary for coordination among decision-system actors.[17]

Automated decision tools are particularly attractive to designers of delegated, distributed decision systems because their deployment promises to improve consistency, decrease bias, and lower costs.[18] For example, such tools are being used or considered for decisions involving pretrial detention,[19] sentencing,[20] child welfare,[21] credit,[22] employment,[23] and tax auditing.[24] Unfortunately, the inscrutability of many machine-learning-based decision tools creates barriers to all of the explanatory flows illustrated in Figure 2.[25] Expanding the focus of the explainability debate to include public accountability is thus only one step toward a more realistic view of the ramifications of decision tool inscrutability. Before incorporating machine-learning-based decision tools into a delegated, distributed decision system, agenda setters should have a clear-eyed view of what information is feasibly available to all of the system's actors. This would enable them to assess whether that information, combined with other mechanisms, can provide a sufficient level of accountability[26] and coordination to justify the use of a particular automated decision tool in a particular context.

---

17. See supra Figure 2.

18. See, e.g., Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 Geo. L.J. 1147, 1160 (2017) [hereinafter Coglianese & Lehr, Regulating by Robot] ("Despite this interpretive limitation, machine-learning algorithms have been implemented widely in private-sector settings. Companies desire the savings in costs and efficiency gleaned from these techniques . . . .").

19. See, e.g., Jessica M. Eaglin, Constructing Recidivism Risk, 67 Emory L.J. 59, 61 (2017).

20. See, e.g., State v. Loomis, 881 N.W.2d 749, 753 (Wis. 2016).

21. See, e.g., Dan Hurley, Can an Algorithm Tell When Kids Are in Danger?, N.Y. Times Mag. (Jan. 2, 2018), https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html (on file with *the Columbia Law Review*).

22. See, e.g., Matthew Adam Bruckner, The Promise and Perils of Algorithmic Lenders' Use of Big Data, 93 Chi.-Kent L. Rev. 3, 12–13 (2018).

23. See, e.g., Pauline T. Kim, Data-Driven Discrimination at Work, 58 Wm. & Mary L. Rev. 857, 860 (2017).

24. See, e.g., Kimberly A. Houser & Debra Sanders, The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?, 19 Vand. J. Ent. & Tech. L. 817, 819–20 (2017).

25. See infra section IV.B.

26. See, e.g., Bloch-Wehba, supra note 11 (manuscript at 27–28) (discussing the challenge of determining adequate public disclosure of algorithm-based government decision-making); Brauneis & Goodman, supra note 11, at 166–67 ("Governments should consciously generate—or demand that their vendors generate—records that will further public understanding of algorithmic processes."); Citron, supra note 11, at 1305–06 (arguing that mandatory audit trails "would ensure that agencies uniformly provide detailed notice to individuals"); Gillis & Simons, supra note 9 (manuscript at 2) ("Accountability is achieved when an institution must justify its choices about how it developed and implemented its decision-making procedure, including the use of statistical techniques or machine learning, to an individual or institution with meaningful powers of oversight and

Incorporating inscrutable automated decision tools has ramifications for all stages of delegated, distributed decisionmaking. This Essay focuses on the implications for the creation of decision criteria—or rulemaking.[27] As background for the analysis, Part I briefly compares automated, machine-learning-based decision tools to more familiar forms of decisionmaking criteria. Part II uses the explanation requirements embedded in administrative law as a springboard to analyze the functions that explanation has conventionally been expected to perform with regard to rulemaking. Part III considers how incorporating inscrutable machine-learning-based decision tools changes the potential effectiveness of explanations for these functions. Part IV concludes by suggesting approaches that may alleviate these problems in some contexts.

I. INCORPORATING MACHINE-LEARNING-BASED TOOLS INTO DELEGATED, DISTRIBUTED DECISION SYSTEMS

The design of a delegated, distributed decision system begins with an agenda setter (or agenda setters) empowered to determine the goals that should guide case-by-case decisions. To align decision outcomes with the system's goals as consistently and efficiently as possible, agenda setters task rulemakers with specifying decision criteria for adjudicators to apply. While legislators specify some decision criteria on behalf of the public, they routinely delegate rulemaking to agencies.[28] The general framework of agenda setting, rulemaking, and adjudication describes many decision systems, including in the private sector.[29]

A. *Rules, Standards, and Automated Decision Tools*

Rulemakers can devise various sorts of decision criteria, depending on the decision context. Criteria can be rule-like—specifying which case-by-case facts are to be taken into account and how—or standard-like—giving adjudicators more flexibility regarding what factual circumstances

---

enforcement."); Selbst & Barocas, supra note 4, at 1138 ("Where intuition fails, the task should be to find new ways to regulate machine learning so that it remains accountable.").

27. Elsewhere, I focus on the implications for adjudication. Katherine J. Strandburg, Adjudicating with Inscrutable Decision Rules, *in* Machine Learning and Society: Impact, Trust, Transparency (Marcello Pelillo & Teresa Scantamburlo eds., forthcoming 2020) (on file with the *Columbia Law Review*).

28. See Whitman v. Am. Trucking Ass'ns, 531 U.S. 457, 488 (2001) (Stevens, J., concurring in part and concurring in the judgment) ("[I]t would be both wiser and more faithful to what we have actually done in delegation cases to admit that agency rulemaking authority is 'legislative power.'"); see also, e.g., 5 U.S.C. § 553 (2012) (explaining the process by which agencies engage in informal rulemaking); 42 U.S.C. § 7409 (2012) (delegating determination of emissions and other air pollution standards to the Environmental Protection Agency).

29. See Tony Porter & Karsten Ronit, Self-Regulation as Policy Process: The Multiple and Criss-Crossing Stages of Private Rule-Making, 39 Pol'y Sci. 41, 43 (2006) (explaining how private firms develop policy to avoid government regulation using processes such as agenda setting, problem identification, and adjudication).

they deem relevant and how they weigh those facts in coming to a decision. Decision criteria may also combine rule-like and standard-like aspects according to various schemes. For example, DWI laws in many states combine a rule-like blood alcohol threshold, above which a finding of intoxication is required, with a standard-like evaluation of intoxication at lower levels.[30] Some speed limit laws use a somewhat different scheme: Above a rule-like speed limit, there is a presumption of unsafe driving, but adjudicators may make standard-like exceptions for a narrow range of emergency circumstances.[31]

Federal sentencing guidelines illustrate another possible approach. In *United States v. Booker*, the Supreme Court held that it is unconstitutional to treat the guidelines as completely mandatory rules.[32] Judges are now "required to properly calculate and consider the guidelines when sentencing, even in an advisory guideline system."[33] The guidelines thus retain their rule-like character, but the combination scheme now gives judges the flexibility to weigh them in light of other circumstances.

Rulemakers' design choices implicate well-known trade-offs between the predictability, consistency, technical expertise, and efficiency of rule-like criteria on the one hand and the flexibility and adaptability of standard-like criteria on the other. Incorporating an automated decision tool has several implications for those design choices. First, rulemakers will need to divide decision criteria explicitly into automated and nonautomated sets, recognizing that automated assessment is utterly rule-like. Conventional narrative descriptions of decision criteria allow a spectrum from rule-like to standard-like that does not always demand such bright line allocation up front. Second, automation, especially using machine learning, distinctively constrains the sorts of rules that can be developed.[34] Third, the use of inscrutable automated decision tools limits the schemes that adjudicators can feasibly use to combine automated assessments with their assessments of nonautomated factors.[35]

Complete automation of consequential decisions is uncommon, and likely to remain so, for normative and legal reasons.[36] Human adjudicators will often be tasked with evaluating some aspects of decision criteria and combining those evaluations with automated tool outputs to make final decisions. Because different combination schemes can produce very

---

30. See Drunk Driving Laws and Penalties by State, Justia, https://www.justia.com/50-state-surveys/drunk-driving-dui-dwi/ [https://perma.cc/8B35-AKUP] (last updated July 2018). For a specific example, see N.Y. Veh. & Traf. Law § 1192 (McKinney 2019).

31. See, e.g., Speeding Tickets: How to Defend Yourself, Nolo, https://www.nolo.com/legal-encyclopedia/speeding-tickets-defending-yourself-29605.html [https://perma.cc/AQ3L-469A] (last visited Aug. 10, 2019).

32. 543 U.S. 220, 245 (2005).

33. U.S. Sentencing Comm'n, Guidelines Manual 14 (2018).

34. See Strandburg, supra note 27 (manuscript at 16); infra section II.B.

35. See Strandburg, supra note 27 (manuscript at 13–20).

36. See infra Part III.

different outcomes, rulemakers should specify a combination scheme for adjudicators to apply. The rigidity of automated assessment rules limits the feasible combination schemes, especially when the automated tool is inscrutable to adjudicators, and often to rulemakers as well.[37]

B. *Machine Learning Models as Decision Tools*

Developments in machine learning are driving the recent upsurge of interest in automated decision tools. Machine learning is designed to fit "big" training data to complex, nonlinear models that map large sets of input features to outcome variables,[38] which serve as proxies for a relevant decision criterion of interest.[39] By using large numbers of features and training data for many individual cases, machine learning can automatically "learn" nuanced distinctions between cases from the training data, thereby producing models that are both "personalized" and more "evidence based" than may be possible using more conventional rulemaking approaches.[40] The hope is that machine-learning-based decision tools can extend automated, rule-like assessment to some decision criteria that adjudicators would conventionally have been required to evaluate in a more standard-like manner.[41] The choice to incorporate a machine-learning-based decision tool constrains rulemakers' design choices in several important ways, however.

---

37. For more in-depth discussion of this point, see Strandburg, supra note 27 (manuscript at 13–19).

38. See John Nay & Katherine J. Strandburg, Generalizability: Machine Learning and Humans in the Loop, *in* Research Handbook on Big Data Law (Roland Vogl ed., forthcoming 2019) (manuscript at 15), https://ssrn.com/abstract=3417436 (on file with the *Columbia Law Review*).

39. For useful overviews of the machine learning process, see Lehr & Ohm, supra note 12, at 669–702; Burrell, supra note 3, at 5; Pedro Domingos, A Few Useful Things to Know About Machine Learning, Comms. ACM, Oct. 2012, at 78, 79–80. Note that data scientists usually separate the available data into "training" and "test" data sets to improve validation. Lehr & Ohm, supra note 12, at 685–88. There are a number of techniques for doing this, but this Essay will gloss over the distinction and refer to all of the data that is used to develop the model as "training" data.

40. See Anthony J. Casey & Anthony Niblett, A Framework for the New Personalization of Law, 86 U. Chi. L. Rev. 333, 333 (2019) ("Personalized law is an old concept. The idea that the law should be tailored to better fit the relevant context to which it applies is obvious and has been around as long as the idea of law itself."); see also P'ship for Pub. Serv., Seize the Data: Using Evidence to Transform How Public Agencies Do Business 3 (2019), https://ourpublicservice.org/wp-content/uploads/2019/06/Seize-the-Data.pdf [https://perma.cc/2UD3-D2QA] (discussing the ways in which federal agencies can utilize data to inform their decisionmaking).

41. Casey & Niblett, supra note 40, at 335 ("As technologies associated with big data, prediction algorithms, and instantaneous communication reduce the costs of discovering and communicating the relevant personal context for a law to achieve its purpose, the goal of a well-tailored, accurate, and highly contextualized law is becoming more achievable."). But see, e.g., Solon Barocas, danah boyd, Sorelle Friedler & Hanna Wallach, Editorial, Social and Technical Trade-Offs in Data Science, 5 Big Data 71 (2017) (providing an overview of several critiques of machine learning models).

1. *Data-Driven Constraints on Rule Design.* — Machine learning has the potential to create nuanced models of how outcome variables depend on many feature variables, but collecting the sort of "big data" needed to take advantage of machine learning's strengths is difficult and expensive. As a result, machine learning processes often rely on "found data,"[42] collected for some other purpose, to train the models.[43] Unfortunately, reliance on found data leaves rulemakers at the mercy of whatever feature sets and outcome variables happen to have been collected.[44] Having "big data" for an outcome variable makes it possible to train a model that effectively predicts that outcome variable, but that sort of data is often not available for the decision criteria that are truly of interest. Treating a loose or inaccurate proxy as if it were a true assessment is likely to lead to inaccurate, biased, and otherwise problematic decisions.[45] For example, a judge might like to know the likelihood that the defendant would commit a serious crime if released pending trial, but the available data might instead record arrests for any crime, which is a loose and biased proxy for the factor of interest.[46] There is thus often a trade-off between using an outcome variable for which "bigger" data is available and using a better proxy for the true criteria of interest. The need for "big" training data similarly limits the available feature sets to data types that have been recorded for large numbers of individuals.[47] Those limits constrain the sorts of factual "evidence" that can be considered by a machine-learning-based decision tool. As a result, opting to use a machine-learning-based decision tool places restrictions on decision-criteria design that may or may not be worth the trade-offs.

The limitations imposed by training data availability are related to a machine learning model's "generalizability," or ability to perform well in handling cases that were not included in the data used to train it.[48] Generalizability is also related to issues of over- or under-fitting that are associated with the extent to which a model can pick up normatively

---

42. See Matthew Salganik, Bit by Bit: Social Research in the Digital Age, ch. 2.2 (open review ed. 2019).

43. Id.

44. Id.

45. Id.; see also Emily Keddell, Substantiation Decision-Making and Risk Prediction in Child Protection Systems, 12 Pol'y Q. 46, 48 (2016) (discussing bias and other problems with using "substantiation, meaning a decision that abuse has been investigated and found to have occurred," as an outcome variable for predicting risk of child abuse).

46. See, e.g., Eaglin, supra note 19, at 75–77 (2017) ("[D]efining recidivism is less intuitive and more subjective than it may appear.").

47. See Nay & Strandburg, supra note 38 (manuscript at 14) ("Relevant information may be left out of the feature set simply because it was not prevalent enough in the training data, because it is idiosyncratic, unquantifiable or otherwise not collectible *en masse* or because it is newly available and/or newly relevant due to societal or technological changes.").

48. Id. (manuscript at 7) ("A model is generalizable to the extent it applies, and performs similarly well, beyond the particular dataset from which it was derived.").

relevant distinctions between cases.[49] A model that accurately fits its training data can fail to generalize well if new factual scenarios crop up over time, if its outcome variable is a bad proxy for some subgroups of the population, if its feature variables do not capture all normatively relevant distinctions, or if it is simply over-fitted to the training data because of the way that developers have tuned the machine learning parameters. By limiting the outcome variables and features that a machine-learning-based model can consider, data availability constraints are likely to limit the model's generalizability. There is no computational metric for generalizability because it depends on how well the model will perform on as-yet-unknown cases.

2. *Inscrutability and Decision-Criteria Design.* — Machine learning's inscrutability stems from the fact that the computational mapping from feature inputs to outcome prediction is often hard to explain in terms that are intuitively comprehensible to humans.[50] Part of what makes these mappings difficult to explain is their reliance on large numbers of features, which can make the behavior of even simple functions difficult to intuit.[51] "Deep learning" models lack explainability at a more fundamental level, in that the ways they map input features to outcome variables cannot be represented in standard forms, such as closed equations, decision trees, or graphs.[52] Even developers and subject matter experts find it difficult or impossible to interpret such models, though there is ongoing research into technical methods for producing approximate interpretations of inscrutable machine-learning models and for training sufficiently accurate explainable models.[53] Developers employ inscrutable machine learning models, despite their explainability issues, because they are often more accurate in fitting the training data.[54] Essentially, this is because a more complicated, and thus less explainable, computational mapping can always be fit more closely to the training data.[55] Discussions of this trade-off between "accuracy" and explainability have focused rather myopically on explanation's value to decision

---

49. Id. For the balance of this Essay, I will refer to both sorts of concerns as "generalizability."

50. See supra note 3 and accompanying text.

51. See, e.g., Selbst & Barocas, supra note 4, at 1100–05 (discussing explainability in credit scoring).

52. See, e.g., Bornstein, supra note 3.

53. See, e.g., Lehr & Ohm, supra note 12, at 708–10; Selbst & Barocas, supra note 4, at 1110–15.

54. See, e.g., David Weinberger, Optimization over Explanation: Maximizing the Benefits of Machine Learning Without Sacrificing Its Intelligence, Medium (Jan. 28, 2018), https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d [https://perma.cc/4U2F-5BW7].

55. See id. ("[U]nderstanding and measuring AI systems in terms of their optimizations gives us a way to benefit from them even though they are imperfect and even when we cannot explain their particular outcomes.").

subjects.[56] This section briefly explores how inscrutability constrains decision-criteria design, focusing on its implications for a decision system's ability to cope with generalizability concerns. Explanations of decision criteria also have important functions associated with accountability and coordination, which are analyzed in Part II, below.

Generalizability is essentially the technical version of the long-standing concern that rule-like decision criteria will be insufficiently flexible and forward-thinking to produce good outcomes in real-world decisions. While machine-learning-based models can be more nuanced than conventional rules in taking account of many known features, they cannot avoid the limitations of their training data.[57]

Conventional decision systems cope with generalizability concerns in two ways. First, rulemakers can scrutinize the rules in advance and try to imagine how things might go wrong, so that the rules can be redesigned to avoid problems that would otherwise crop up in real-world cases.[58] This option is not available for inscrutable machine-learning-based models. While rulemakers can and should scrutinize the training data, features, outcome variables, and validation metrics, those methods are not equivalent to scrutinizing the logic of the rule.

Second, conventional rulemakers often provide adjudicators with some standard-like flexibility to use analogy, common sense, normative judgment, and so forth to cope with case-by-case circumstances that are not adequately treated by rule-like criteria. Human adjudicators' ability to generalize in this way is limited when they are faced with the output of an inscrutable automated decision tool because they cannot discern whether and how the tool has failed to consider relevant factual circumstances. These limitations on adjudicators' capacity to generalize constrain the sorts of schemes that rulemakers can design for combining automated and nonautomated factors. For example, while adjudicators can apply the per se blood alcohol limit discussed earlier without understanding its basis, they cannot sensibly consider whether to deviate from the sentencing guidelines in a particular case without understanding the basis for the suggested sentence.[59]

II. CONVENTIONAL REASONS FOR EXPLAINING RULEMAKING

When critics talk about the inscrutability of machine-learning-based decision tools, a common rejoinder is that human decisionmakers are

56. See Selbst & Barocas, supra note 4, at 1111.

57. See Nay & Strandburg, supra note 38 (manuscript at 6).

58. Id. (manuscript at 14–15).

59. For a more extensive discussion of these issues, see Strandburg, supra note 27 (manuscript at 15–17).

also "black boxes,"[60] in the sense that it is impossible to know what went on in a human decisionmaker's mind before coming to a decision.[61] This rejoinder misses the mark. Reason giving is a core requirement in conventional decision systems precisely *because* human decisionmakers are inscrutable and prone to bias and error, not because of any expectation that they will, or even can, provide accurate and detailed descriptions of their thought processes. This point sharpens when one shifts from Figure 1's decisionmaking paradigm to the more realistic paradigm of Figure 2. When the decisionmaker is a distributed, multi-actor institution, explanation requirements cannot be aimed at uncovering what went on in "the" decisionmaker's mind.

Generations of legal scholars have considered the functions that explanation and reason giving can perform in delegated, distributed decision systems. Machine-learning-based decision tools ease some of the familiar challenges posed by human black boxes and create some new ones.[62] Before focusing on what is distinctive about these tools, it makes sense to learn from our experience with legal explanation requirements for human decision systems.[63] Section II.A therefore provides a brief overview of some of the primary sources for legal explanation requirements. Section II.B then discusses the primary theoretical rationales behind these requirements, as relevant to rulemakers.

A. *Legal Reason-Giving Requirements*

The principle that government decisions should be justified by reasons is well enshrined in the law, not only in the United States but also in

---

60. See, e.g., Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 3–8 (2015) ("The term 'black box' is a useful metaphor . . . it can mean a system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other.").

61. See, e.g., Yavar Bathaee, The Artificial Intelligence Black Box and the Failure of Intent and Causation, 31 Harv. J.L. & Tech. 889, 891–92 & nn.11–12 (2018). See generally Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Machine Intelligence 206, 208–10 (2019) (arguing that explanations of black box models "often do not make sense or do not provide enough detail to understand what the black box is doing").

62. See infra Part IV.

63. My aim here is not to delve into when, or whether, the law *requires* explanations for government decisions using automated tools, though that question is of obvious importance. See, e.g., Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 Wash. L. Rev. 1, 8 (2014); Citron, supra note 11, at 1301–13; Kate Crawford & Jason Schultz, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, 55 B.C. L. Rev. 93, 121–28 (2014); Daniel J. Steinbock, Data Matching, Data Mining, and Due Process, 40 Ga. L. Rev. 1, 64–81 (2005); Cary Coglianese & David Lehr, Adjudicating by Algorithm, Regulating by Robot, The Regulatory Review (May 22, 2017), https://www.theregreview.org/2017/05/22/coglianese-lehr-adjudicating-algorithm-regulating-robot/ [https://perma.cc/7R9L-AGDR].

other democracies.[64] Under U.S. law, reason giving is a key component of the constitutional requirement that no one be deprived by government of life, liberty, or property without "due process of law."[65] Though government decisionmakers are not always required to give reasons for their decisions, reason giving is the least common denominator of due process requirements.[66]

Administrative law is especially concerned with delegated, distributed decision systems and has been described as "the progressive submission of power to reason."[67] Where agencies engage in rulemaking, explanations address not only the individual right to due process but also concerns about separation of powers and delegation of legislative power.[68] The Administrative Procedure Act (APA), along with the Constitution's due process requirement, imposes general procedural structures and constraints that apply to most federal agencies.[69] Its purposes include informing the public about the agency's activities and providing for public participation in the rulemaking process.[70] Its provisions thus exemplify the sort of explanation requirements that law imposes on rulemakers.

Explanation and justification are at the heart of notice and comment rulemaking, the most common process by which administrative agencies promulgate regulations.[71] This dialogue with the public, who are the ultimate agenda setters for government decision systems, illustrates one aspect of explanation's function within a distributed decision system. After designing a set of regulations, an agency ordinarily must

---

64. See, e.g., Jerry L. Mashaw, Reasoned Administration: The European Union, the United States, and the Project of Democratic Governance, 76 Geo. Wash. L. Rev. 99, 101 (2007) [hereinafter Mashaw, Reasoned Administration].

65. U.S. Const. amends. V, XIV.

66. See SEC v. Chenery Corp., 318 U.S. 80, 94 (1943) (holding that an administrative agency must provide an understandable reason for its action so that a court may review it); Martin Shapiro, The Giving Reasons Requirement, 1992 U. Chi. Legal F. 179, 197 (explaining that the European Economic Community Treaty's reason-giving requirement is roughly equivalent to due process requirements in the U.S. Constitution).

67. Jerry L. Mashaw, Small Things Like Reasons Are Put in a Jar: Reason and Legitimacy in the Administrative State, 70 Fordham L. Rev. 17, 26 (2001) [hereinafter Mashaw, Small Things].

68. See Kevin M. Stack, The Constitutional Foundations of *Chenery*, 116 Yale L.J. 952, 1020–21 (2007) (explaining that *Chenery* "enforces a presumption" that "require[s] Congress to condition the grant of authority to an agency on the agency's expressly stating its grounds for acting"); see also, e.g., Mashaw, Small Things, supra note 67, at 22–23.

69. See Administrative Procedure Act, 5 U.S.C. §§ 551–557 (2012); see also U.S. Const. amends. V, XIV.

70. U.S. Dep't of Justice, Attorney General's Manual on the Administrative Procedure Act 9 (1947).

71. More formal rulemaking processes are required in some situations. 5 U.S.C. §§ 553(c), 556–557. Agencies may also promulgate internal procedural rules, interpretive guidance, and general policy statements without engaging in notice and comment procedures. Id. § 553(b)(A).

publish them in the Federal Register, along with a section that "discusses the merits of the proposed solution, cites important data and other information used to develop the action, and details its choices and reasoning. The agency must also identify the legal authority for issuing the rule."[72] After publication, the public is given an opportunity to comment on the proposal.[73] The agency must then consider the comments when it finalizes the rule.[74] Final rules must be published along with a statement that "sets out the goals or problems the rule addresses, describes the facts and data the agency relies on, responds to major criticisms in the proposed rule comments, and explains why the agency did not choose other alternatives."[75] Rulemakers are also required to explain any later changes to existing regulations,[76] which helps ensure that reforms are made carefully and for appropriate reasons.

Judicial oversight is another mechanism for ensuring that rules are in accord with the agenda setter's goals.[77] The record of the rulemaking process is an important basis for judicial review. Courts generally must defer to agency legal interpretation and expertise wherever the governing statute is silent or ambiguous.[78] Nonetheless, a regulation may be overturned if a reviewing court determines that it is unconstitutional; inconsistent with the governing statutory authority; or arbitrary, capricious, or an abuse of discretion.[79]

Because judges often lack the substantive expertise that would be required for effective substantive review of agency rulemaking, courts perform a so-called "hard look" review of the rulemaking record to test whether an agency approached a given rulemaking task diligently, rationally, and without pursuing conflicting agendas. Under the hard look approach to the arbitrary and capricious standard, an agency must "demonstrate that it engaged in reasoned decisionmaking by providing an

---

72. A Guide to the Rulemaking Process, Office of the Fed. Register, https://www.federalregister.gov/uploads/2011/01/the_rulemaking_process.pdf [https://perma.cc/5GMH-RUZP] [hereinafter Guide to Rulemaking] (last visited Aug. 11, 2019).

73. 5 U.S.C. § 553(c).

74. See id.; see also United States v. Nova Scotia Food Prods. Corp., 568 F.2d 240, 252–53 (2d. Cir. 1977) (holding an agency's procedures inadequate because they ignored important considerations developed through comments).

75. Guide to Rulemaking, supra note 72.

76. See Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 42 (1983) (holding that, while an agency's change in policy does not need to be supported by reasons more substantial than those underpinning the original rule, it must still be explained); see also FCC v. Fox Television Stations, Inc., 556 U.S. 502, 514–15 (2009).

77. See Chevron, U.S.A, Inc. v. NRDC, 467 U.S. 837, 842–43 (1984) (holding that courts may strike down agency regulations that clearly fall outside of the bounds that Congress set).

78. Id. at 843.

79. 5 U.S.C. §§ 706(2)(A)–(D).

adequate explanation for its decision,"[80] "provide the 'essential facts upon which the administrative decision was based' and explain what justifies the determination with actual evidence beyond a 'conclusory statement.'"[81] A rule will also fail the test if it "is the product of 'illogical' or inconsistent reasoning; . . . fails to consider an important factor relevant to its action, such as the policy effects of its decision or vital aspects of the problem in the issue before it; or . . . fails to consider 'less restrictive, yet easily administered' regulatory alternatives."[82]

The prospect of hard look review "on the record" gives agencies incentives to create detailed records justifying the rules they promulgate, thereby also providing incentives for agencies to make rules that *can* be justified by such records. These accountability mechanisms are far from perfect and are regularly critiqued[83] but nonetheless endure as core means for addressing the unavoidable accountability problems faced by delegated, distributed decision systems.

B. *Reasons for Explaining Rulemaking*

Legal scholars have identified many normative rationales for reason-giving requirements. While some of these rationales pertain primarily to explanations aimed at decision subjects,[84] many are relevant to this Essay's focus on rulemaking. One important category of rationales focuses on improving the *quality* of the rules, in the sense of how effectively they further the agenda setter's goals. While scholars have mostly viewed quality control through an accountability lens, the law's reason-giving requirements also facilitate coordination, as this section explains. Another category of rationales is founded in the special relationship citizens have with a democratic government, in that they are both decision subjects and agenda setters.

---

80. Todd Garvey, Cong. Research Serv., R41546, A Brief Overview of Rulemaking and Judicial Review 15 (2017); see also *State Farm*, 463 U.S. at 52.

81. Garvey, supra note 80, at 15 (quoting United States v. Dierckman, 201 F.3d 915, 926 (7th Cir. 2000)).

82. Id. (first quoting Am. Fed'n of Gov't Emps., Local 2924 v. Fed. Labor Relations Auth., 470 F.3d 375, 380 (D.C. Cir. 2006); then quoting Cincinnati Bell Tel. Co. v. FCC, 69 F.3d 752, 761 (6th Cir. 1995)).

83. See, e.g., Gutierrez-Brizuela v. Lynch, 834 F.3d 1142, 1152 (10th Cir. 2016) (Gorsuch, J., concurring) ("In this way, *Chevron* seems no less than a judge-made doctrine for the abdication of the judicial duty."); Stephen Breyer, Judicial Review of Questions of Law and Policy, 38 Admin. L. Rev. 363, 383 (1986) (claiming that hard look review leads to "abandonment or modification of the initial project irrespective of the merits").

84. See Henry J. Friendly, Some Kind of Hearing, 123 U. Pa. L. Rev. 1267, 1280–81 (1975) (explaining that providing notice and grounds for the proposed action helps the individual "marshal evidence and prepare his case"); Martin H. Redish & Lawrence Marshall, Adjudicatory Independence and the Values of Due Process, 95 Yale L.J. 455, 475–91 (1986) ("The instrumental conception of due process focuses on the individual's interest in having an opportunity to convince the decisionmaker that he deserves the right at issue.").

1. *Reason Giving to Improve Quality.* — Reason giving "promotes accountability by limiting the scope of available discretion and ensuring that public officials provide public-regarding justifications for their decisions" and "facilitates transparency, which, in turn, enables citizens and other public officials to evaluate, discuss, and criticize governmental action, as well as potentially to seek legal or political reform."[85] It is also a bulwark against arbitrariness.[86] For administrative agencies, "legitimacy flows primarily from a belief in the specialized knowledge that administrative decisionmakers can bring to bear on critical policy choices. And the only evidence that this specialized knowledge has in fact been deployed lies in administrators' explanations or reasons for their actions."[87] In addition to promoting quality through accountability, reason giving might be expected to improve rule quality through the disciplining effect of "showing your work" and by facilitating communication and coordination among rulemakers. Reason giving also assists in the evaluation and reform of rules.[88]

a. *The "Show Your Work" Phenomenon.* — The "show your work" phenomenon is familiar: The very process of explaining one's reasoning is likely to improve it by highlighting loopholes, inconsistencies, and weaknesses.[89] For groups, the "show your work" phenomenon includes the benefits of deliberating to jointly produce an explanation. If rulemakers anticipate that outsiders will see, and potentially critique, their explanations, the effect is heightened, since the prospect of being exposed as sloppy, ill informed, biased, or captured should provide incentives for rulemakers to devise rules that *can* be explained and justified.

By forcing rulemakers to justify their work product in terms of appropriate goals and relevant facts, the "show your work" phenomenon may also deter bias and arbitrariness. This phenomenon will presumably

---

85. Glen Staszewski, Reason-Giving and Accountability, 93 Minn. L. Rev. 1253, 1278 (2009).

86. See Lisa Schultz Bressman, Beyond Accountability: Arbitrariness and Legitimacy in the Administrative State, 78 N.Y.U. L. Rev. 461, 473–74 (2003) (noting that the initial motivation for judicial innovations such as a "reasoned consistency" requirement for agency decisions was the prevention of arbitrariness); Christine N. Cimini, Principles of Non-Arbitrariness: Lawlessness in the Administration of Welfare, 57 Rutgers L. Rev. 451, 510–12 (2005) (arguing that accountability for and reviewability of agency decisions serve to prevent arbitrary decisionmaking).

87. Mashaw, Reasoned Administration, supra note 64, at 117. But see Jodi L. Short, The Political Turn in American Administrative Law: Power, Rationality, and Reasons, 61 Duke L.J. 1811, 1814 (2012) (discussing the "gathering movement to reconceptualize the legitimacy of administrative agencies in terms of their political—and specifically, their presidential—accountability as opposed to their expertise, their fidelity to statutory commands, or their role as fora for robust citizen participation and deliberation" (footnotes omitted)).

88. See Strandburg, supra note 27 (manuscript at 12–13).

89. See In Re Expulsion of N.Y.B., 750 N.W.2d 318, 326 (Minn. Ct. App. 2008) (drawing an analogy between procedural requirements in administrative law and the "show your work" method of teaching mathematics).

be most effective when it creates self-awareness of unintentional bias and arbitrariness. But, as one commentator colorfully put it, "hypocrisy has a civilising force" in human decisionmaking.[90] Explanations facilitate scrutiny, making it more difficult to mask intentional bias.

b. *Explaining to Agenda Setters.* — Notice and comment, review by the Office of Information and Regulatory Affairs (OIRA), and judicial review exemplify the interplay of explanation and feedback between agenda setters and rulemakers. Explanations to agenda setters perform two main functions related directly to the principal–agent problems mentioned earlier.[91] The first function is accountability, which entails keeping an eye out for ways in which a rulemaking entity's bias, conflicts of interest, sloppiness, or lack of zeal might have infected the rule it devised. The second function is to catch misalignments between the agenda setter's goals and the rule's potential application to real-world case types that rulemakers may not have considered adequately (or at all). This second function relates to the generalizability concerns discussed in Part I.[92] For government decision systems, the general public is the ultimate agenda setter. Public feedback may also be vital to some private decision systems because of the value of engaging diverse perspectives in ferreting out problems of accountability and misalignment.

The benefits of a public explanation obviously depend on whether the public is willing and able to engage with it—a perennial problem. Notice and comment has been criticized because well-funded, concentrated interests are better equipped to understand the proposed rules and to use their influence to bend them to their own benefit.[93] The empirical picture is mixed. While many studies find little participation by individuals in notice and comment rulemaking,[94] some have found

---

90. Guido Noto La Diega, Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information, 9 J. Intell. Prop. Info. Tech. & Electronic Comm. L. 3, 10 (2018).

91. See supra notes 15–16 and accompanying text.

92. See supra section I.B.1.

93. See Susan Webb Yackee, Sweet-Talking the Fourth Branch: The Influence of Interest Group Comments on Federal Agency Rulemaking, 16 J. Pub. Admin. Res. & Theory 103, 105 (2005) ("[T]he notice and comment period is an important political arena where the bureaucracy frequently alters and adapts public policies to better match the preferences of interest group commenters.").

94. See Cary Coglianese, Citizen Participation in Rulemaking: Past, Present, and Future, 55 Duke L.J. 943, 958 (2006) ("Most rules still garner relatively few overall comments and even fewer comments from individual citizens."); Marissa Martino Golden, Interest Groups in the Rule-Making Process: Who Participates? Whose Voices Get Heard?, 8 J. Pub. Admin. Res. & Theory 245, 253–54 (1998) ("Neither NHTSA nor the EPA received a single comment from an individual citizen on any of the eight rules that were examined. . . . Here, fully 9 percent of the comments HUD received on this rule were submitted by individual citizens."); Nina A. Mendelson, Rulemaking, Democracy, and Torrents of E-Mail, 79 Geo. Wash. L. Rev. 1343, 1357–59 & n.79 (2011) ("Although the right to . . . submit written comments in agency rulemaking extends to every member of the public, . . . actual participation in rulemaking is not well balanced.").

substantial participation in commenting on particular sorts of regulations by citizen groups or individuals submitting form letters.[95] While citizen groups are usually not heavily resourced, they can build up significant subject matter expertise, allowing them to submit meaningful feedback and criticism.[96] This is obviously not a complete answer to the power imbalance, but it counsels against underestimating the societal benefit of public explanations. Moreover, the power imbalances in the case-by-case decision systems of interest to us here are somewhat different from those in the standard interest group story, in which powerful regulated entities use notice and comment to influence agencies propounding environmental or consumer protection regulations.[97] Here, the affected parties are individuals, who do not have outsize power to influence the design of decision systems that are critical to their opportunities in vital arenas such as employment, credit, public benefits, criminal justice, and family life.[98] Moreover, these individuals are members of the public and thus agenda setters in their own right.

c. *Communication and Coordination Among Rulemakers.* — It almost goes without saying that the quality of outcomes from a delegated, distributed decision system depends on coordination and communication between the players, including within rulemaking entities and between rulemakers and adjudicators.[99] For conventional decision systems, explanation's coordinating function has received considerably less scholarly attention than its accountability function. This is not terribly surprising for two reasons. First, the narrative form of conventional rules makes their content somewhat self-explanatory to agenda setters and adjudicators, shifting the focus toward explaining why that content is justified. Second, explanation's coordinating function piggybacks on its accountability function. Requiring rulemakers to create explanations aimed at the public, courts, or other agenda setters indirectly provides incentives for the coordinated effort necessary to create those explanations, which

---

95. See Mariano-Florentino Cuéllar, Rethinking Regulatory Democracy, 57 Admin. L. Rev. 411, 462 (2005) (studying three regulatory proceedings in which 72.1%, 98.6%, and 98.3% of comments, respectively, came from individual members of the public; in two of the proceedings, individual comments were almost exclusively form letters); Golden, supra note 94, at 253–55 (finding contributions by citizens' groups ranging from 0% to 16.7% of comments depending on the agency and regulation).

96. See Cuéllar, supra note 95, at 450–51, 458–59 (finding, in a study of two regulatory proceedings, considerably higher values for "comment sophistication" in comments from public membership or public interest organizations than from individuals); Yackee, supra note 93, at 105 ("[I]nterest group comments provide a new source of information and expertise to the bureaucracy during the rulemaking process.").

97. See Golden, supra note 95, at 255 (contrasting EPA and NHTSA rulemakings with "extremely limited participation by public interest or citizen advocacy groups" with HUD rulemakings where "commenters include citizen advocacy groups, individual citizens, and a wide range of government agencies").

98. See supra notes 19–23 and accompanying text.

99. For more on explanations between rulemakers and adjudicators, see Strandburg, supra note 27 (manuscript at 10–13).

in turn activates the "show your work" phenomenon.[100] Similarly, the record creation incentivized by hard look review[101] requires internal coordination, while the resulting record can facilitate further communication and coordination. In addition, and partly to ensure that the required explanations and record will pass muster, rulemaking bodies often impose procedures that amount to internal explanation requirements.[102]

2. *Reason Giving, Democracy, and Respect.* — Reason giving legitimates governmental decisionmaking in a democracy because, as one scholar puts it, "[a]uthority without reason is literally dehumanizing. It is, therefore, fundamentally at war with the promise of democracy, which is, after all, self-government."[103] Particularly in the context of rulemaking by unelected administrative agencies, reason-giving requirements ensure that members of the public are treated as citizens, rather than subjects: "[T]o be subject to administrative authority that is unreasoned is to be treated as a mere object of the law or political power, not a subject with independent rational capacities."[104]

Explanations also empower citizens in their agenda-setting role, by helping them to understand what the rules require, providing bases for individual and group opinion formation and advocacy, and helping minorities to identify rules that ignore or undermine their interests.[105] Reason giving thus "embodies, and provides the preconditions for, a deliberative democracy that seeks to achieve consensus on ways of promoting the public good that take the views of political minorities into account."[106]

These rationales do not have the same force for private-sector decisions, where decision subjects ordinarily do not have similar agenda-setting rights. But explaining the rationale behind decisionmaking criteria also comports with more general societal norms of fair and nonarbitrary treatment. Moreover, the public has an interest as citizens and individuals, both legally and ethically, in the fairness and reasonableness of

---

100. See supra section II.B.1.a.

101. See supra section II.A.

102. See, e.g., Jennifer Nou, Intra-Agency Coordination, 129 Harv. L. Rev. 421, 436–37, 451 (2015) (explaining that "agency heads fac[ing] greater uncertainty regarding how to formulate and draft their regulations in ways that would withstand judicial challenge . . . can respond by creating structures and processes that lower the costs of internal information processing"); Thomas O. McGarity, The Internal Structure of EPA Rulemaking, Law & Contemp. Probs., Autumn 1991, at 57, 58–59, 90–94 (detailing the internal procedures and "team" approach used by the EPA to respond to "[t]he exigencies of external review and the practical necessities of bringing multiple perspectives within EPA to bear on the decisionmaking process").

103. Mashaw, Reasoned Administration, supra note 64, at 118.

104. Id. at 104.

105. See, e.g., Staszewski, supra note 85, at 1278–84.

106. Id. at 1278.

private decision systems that fundamentally affect people's lives.[107] Indeed, private decision systems do not operate in a legal vacuum but are subject to legal protections including, for example, antidiscrimination laws and protections against fraud. In addition, as a practical matter, some subjects of private-sector decision systems are also users or customers, whose market relationships to decisionmakers give them some leverage to demand explanations of the rules that govern those relationships.

III. EXPLAINING MACHINE-LEARNING-BASED DECISION TOOLS

This Part builds on Part II's brief sketch of the purposes of reason-giving requirements by considering how the limited explainability of machine-learning-based decision tools affects the functions that explanations have conventionally been expected to perform in connection with rulemaking. Section III.A begins by taking a more precise look at which aspects of a machine-learning-based decision tool are unexplainable. Section III.B then reflects on how each of the explanation functions described in Part II is affected by the incorporation of an inscrutable machine-learning-based decision tool.

A. *Cabining Machine Learning's Explainability Problems*

Machine learning's explainability problems reside in the inscrutability of a machine learning model's computational mapping of input features to outcome variables.[108] There are, however, many aspects of the development of machine-learning-based decision tools, and of the decision rules embedded in those tools, that are just as explainable as a rule in conventional narrative form. To assess the impact of inscrutable machine-learning-based decision tools, it is important to be precise about what can and cannot be explained.

1. *Explainable Components of a Machine-Learning-Based Decision Tool.* — In some respects, the touted "black box" nature of machine learning models[109] is not nearly all that it is cracked up to be. Many choices made in the process of creating an automated decision tool are not so different from choices made in more traditional rulemaking processes. Moreover, some of those choices are *embedded as components of the rules* of the ultimate decision system, just as similar choices are reflected in narrative rules, and can be explained in conventional fashion. *Explainable components* include:

- Separation of decision criteria into automated and non-automated aspects;

---

107. Cf. Gillis & Simons, supra note 9 (manuscript at 8–9) (making a similar point about accountability).

108. See supra Introduction.

109. See supra notes 3–8 and accompanying text.

- Definitions of decision criteria to be assessed by the automated tool;
- Definitions of outcome variables to be used as proxies for decision criteria;
- Definitions of feature variables to be used as factual evidence in automated decision criteria assessments; and
- Combination schemes governing how adjudicators should combine automated assessments with other relevant information to make decisions.

Whether, under what circumstances, and to whom the law *requires* rulemakers to explain these components is outside the scope of this analysis, but there are no *technical* barriers to requiring such explanations.

2. *Explainable Rulemaking Record.* — Other important choices involved in creating a machine learning model are not reflected on the face of the ultimate automated decision rule but can be described and explained in a record of the rulemaking process.[110] Such choices include selecting training data, determining machine learning algorithms and technical parameters, devising validation protocols, and evaluating whether a model has been adequately validated to justify using it in a decision rule. All of these choices, and the reasons for them, could be included in a record of the development of a machine-learning-based decision tool. Most importantly, such a record could include information about the sources, demographics, and other characteristics of the training data sample; definitions of validation metrics; and results of validations and performance tests. This information plays much the same role as information about statistical and more specialized technical bases for rules that are routinely included in agency rulemaking records and facilitate hard look review by courts and the cost–benefit analysis required for some rules by OIRA.[111]

B. *Explanation, Decision System Quality, and Machine-Learning-Based Tools*

In light of the previous section's parsing of explainable and unexplainable aspects of machine-learning-based decision tools, this section explores how and why incorporating such tools into a decision system is likely to affect the functions of explanation,[112] identifying where the inscrutability of a machine learning model's computational mapping from input features to outcome variables is likely to create serious problems.

1. *The "Show Your Work" Phenomenon.* — The "show your work" phenomenon carries over straightforwardly to an automated decision tool's

110. See Selbst & Barocas, supra note 4, at 1130–33, for a similar argument in terms of "documentation."

111. See supra section II.A.

112. See supra Part II.

explainable components and recordable information.[113] In essence, developers' design *choices* are all explainable, and the benefits of the "show your work" phenomenon will apply to those choices.[114] The full benefits of the "show your work" phenomenon may not be retained, however, for two reasons. First, the "show your work" phenomenon is effectuated primarily through self-awareness and thus depends on developers having sufficient incentives to create detailed and persuasive explanations. Unfortunately, common practices for developing automated decision tools undermine those incentives. Because many rulemaking entities do not have data scientists on staff, they outsource development or purchase off-the-shelf products.[115] Many of these outsourced machine-learning-based decision tools are burdened with confidentiality agreements that severely limit the explanations and records of development that are provided to rulemaking entities and may block public disclosure almost entirely.[116] Such secrecy undermines the "show your work" phenomenon. Second, the "show your work" phenomenon will not aid in resolving problems that developers cannot avoid through careful design choices and validation, as discussed further in section III.B.2.c, below.

2. *Explaining to Agenda Setters.* — This section considers how the functions of explanation to agenda setters depend on access to (i) the explainable components;[117] (ii) information about data selection, sources, and validation that could be available in a rulemaking record;[118] and (iii) a conventional narrative explanation of the way that the rule maps input features to outcome variables. Explanations to agenda setters serve accountability functions but can also be important for generalizability.[119]

---

113. See supra section II.B.1.a.

114. See Anupam Chander, The Racist Algorithm?, 115 Mich. L. Rev. 1023, 1028–29 (2017) ("[E]ven for programmers or companies who intend to discriminate, the process of coding itself is likely to cause programmers to shy away from actually encoding the discrimination.").

115. See AI Now Inst., Algorithmic Accountability Policy Toolkit 7–9 (2018), https://ainowinstitute.org/aap-toolkit.pdf [https://perma.cc/K9BQ-VQG6] (providing an overview of algorithms used by governments and developed by private companies, such as a Medicaid eligibility tool built by IBM, surveillance technologies built by Palantir, and parole term software developed by Northpointe); see also Allegheny Cty. Dep't of Human Servs., Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions (2017), https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf [https://perma.cc/YNY3-RTCD] (highlighting a child welfare service's predictive risk model built through a public–private partnership).

116. See, e.g., Bloch-Wehba, supra note 11 (manuscript at 9–26) (describing confidentiality constraints on the use of algorithms in Medicaid, education, and criminal law enforcement); Brauneis & Goodman, supra note 11, at 137–59 (describing the use of proprietary algorithms in policing, child welfare, public safety, and teacher performance determinations).

117. See supra section III.A.1.

118. See supra section III.A.2.

119. See supra section II.B.1.b.

As noted earlier, the benefits of public explanation are often effectuated through advocacy groups.[120] To isolate the unique issues stemming from machine-learning-based decision tools, it is thus helpful to consider whether such tools can be satisfactorily explained to advocacy groups with significant substantive expertise and moderate resources, assuming that most other agenda setters, such as legislatures, courts, OIRA, or private businesses, will have at least the capacity of such groups.

a. *Explainable Components.* — The explainable components identified in section III.A.1 will be understandable to a public advocacy group with sufficient expertise and resources and can facilitate extremely valuable checks on the decision system's accountability and generalizability. For example, such a group might assess whether the proxy outcome variable is biased or unlikely to generalize to some sorts of cases; consider whether the use of some feature variables is normatively unacceptable or whether important features are missing from the list; or evaluate whether the amount of flexibility given to adjudicators in combining the automated tool output with other information is appropriate. These agenda setters can help to evaluate whether it is normatively appropriate to use a rule-like automated tool to evaluate certain decision criteria or whether a more flexible, standard-like approach should be required.[121] Though rulemakers presumably will also have considered this question, they may be prone to view automation's potential through rose-colored glasses for various reasons, such as a bias toward cost-cutting measures.[122]

b. *Data Sources and Validation.* — Explanations of data sources and validation in a rulemaking record are potentially useful for uncovering bias or sloppiness, detecting holes in the coverage of the sample set, and ensuring that all normatively relevant performance metrics have been examined. For example, unrepresentative training data is one important source of generalizability problems.[123] The public's diverse perspectives may give it an edge over rulemakers in identifying forms of representativeness that might matter for the decision criteria in question.[124]

The technical knowledge about data science that is required to understand this information may currently be beyond the capacity of many advocacy groups and other agenda setters.[125] Over time, however, advocacy groups, particularly the larger and better resourced among them, will undoubtedly upgrade their technical expertise by involving data scientists in their work, as advocacy groups have done in other

120. See supra section II.B.1.b.

121. See supra section I.A.

122. See Coglianese & Lehr, Regulating by Robot, supra note 18, at 1160.

123. See supra section I.B.1.

124. See supra notes 105–106 and accompanying text.

125. See supra notes 95, 116 and accompanying text.

technical arenas.[126] One concern is that there are so many decision systems—national, state, local, and private—incorporating machine-learning-based decision tools that it may be difficult for advocacy groups, many of which might be small and otherwise nontechnical in nature, to keep up with all of them. For the most part, though, if characteristics about the training data, results from performance tests, and other information discussed in section III.A.2 are included in the rulemaking record, they can be expected to perform the same explanation functions as the information in a more conventional rulemaking record.

c. *Inscrutability of the Computational Mapping from Input Features to Outcome Variable.* — Information about the explainable components, data sources, and validation studies may be sufficient for the accountability function of explanation to agenda setters, in part because those information sources provide access to the most important information available to the rulemaking entity itself. The inscrutability of machine learning models creates more fundamental problems, however, regarding the extent to which explanation can help detect generalizability problems and other unintentional misalignments between the decision system's purposes and the automated criteria.[127] In some respects, the generalizability of a rule is always a guessing game—nobody can be certain how any rule will perform "out in the wild" because there may be cases that neither agenda setters nor rulemakers could have anticipated.[128] Conventional rules, with their narrative format, nonetheless allow human readers to anticipate and identify some generalizability issues using logical inference, analogy, and common sense.

These reasoning methods are not applicable to inscrutable machine learning models, however. Moreover, computational validation tools and other statistical and mathematical analyses cannot provide the same sorts of insights about generalizability, which depend on a grasp of the logic of the rule. Researchers have invented various approaches for creating approximate explanations for a machine learning model's opaque mapping.[129] While many of these methods are designed to explain the specific

126. See, e.g., Shobita Parthasarathy, Breaking the Expertise Barrier: Understanding Activist Strategies in Science and Technology Policy Domains, 37 Sci. & Pub. Pol'y 355, 358–60 (2010) (describing how breast cancer patient advocates found sympathetic experts to educate them about the technical complexities of their causes in order to advance their advocacy). Indeed, some advocacy groups are already beginning to do this. See, e.g., AI Now Inst., Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems 4–5 (2018), https://ainowinstitute.org/litigatingalgorithms.pdf [https://perma.cc/M82T-LR9H] (noting that organizers of a recent workshop examining litigation involving the government's use of algorithmic systems featured participation by relevant legal and scientific experts).

127. See supra section II.B.1.b.

128. Indeed, this is a primary justification for using standards rather than rules. See Strandburg, supra note 27 (manuscript at 13).

129. See generally Lipton, supra note 5 (surveying the academic literature of techniques designed to render machine learning models interpretable).

results of individual cases,[130] some attempt to create more general approximate explanations, which might be useful for probing generalizability issues.[131] For example, a model trained to distinguish wolves from dogs in photographs worked well on its training data but failed on a larger set of photos.[132] The problem was that the training data was skewed—nearly all of the wolves were in snowy landscapes, so the model used the presence of snow to distinguish wolves from dogs.[133] Techniques for creating approximate explanations of the machine logic helped to identify that generalizability problem because, after receiving the explanations, nearly all human observers were able to recognize that "snow" played a key role in that logic.[134]

On the whole, though, it remains uncertain whether any of these technical approaches can replace human analysis of narrative rules. Though machine learning models are trained to reproduce the outputs that human beings assigned to the training data, the mappings they create are not likely to be similar to human mental models.[135] While the association of wolves with snow ran throughout the training data, tougher generalizability issues may arise from unanticipated or uncommon "edge" cases. Humans are reasonably good at reading rules and thinking about whether they are mistaken or have blind spots but are not similarly good at predicting an inscrutable machine learning model's blind spots. For example, a deep learning model trained to triage pneumonia patients performed very well on validation tests.[136] Researchers also created a less accurate, but explainable, model based on the same data.[137] Scrutiny of the explainable model identified a problem in the data: Pneumonia patients with asthma are *high* risk, but because they had routinely been treated in the ICU, their outcomes were good,

130. See id. at 40–42.

131. See Doshi-Velez & Kim, supra note 5, at 7 ("Global interpretability implies knowing what patterns are present in general (such as key features governing galaxy formation), while local interpretability implies knowing the reasons for a specific decision (such as why a particular loan application was rejected).").

132. Marco Tulio Ribeiro, Sameer Singh & Garlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *in* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135, 1142–43 (2016).

133. See id.

134. See id.

135. See, e.g., Kevin Hartnett, Machine Learning Confronts the Elephant in the Room, Quanta (Sept. 20, 2018), https://www.quantamagazine.org/machine-learning-confronts-the-elephant-in-the-room-20180920/ [https://perma.cc/V5DY-HW2Y] (explaining that neural networks may encounter difficulty with fundamental human tasks because of their inability to process confusing and incongruous information).

136. See Rich Caruana, Paul Koch, Yin Lou, Marc Sturm, Johannes Gehrke & Noémie Elhadad, Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission, *in* Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1721, 1730 (2015).

137. See id. at 1721–22.

fooling the model into treating them as *low* risk.[138] The data scientists and medical experts working on the project *could*, in principle, have foreseen that the data for asthma sufferers might be misleading, but they didn't. They identified the asthma problem only after scrutinizing the explainable model.[139] The problem with the asthma data presumably also affected the inscrutable machine learning model, but researchers would not have been able to detect it. Moreover, as the study authors noted, because the inscrutable machine-learning-based model was fit more tightly to the training data than the explainable model, "it was possible that the neural nets had learned other patterns that could put some kinds of patients at risk" that did not show up in the explainable version.[140] Without an intuitive window into the logic of the machine learning model, there was simply no way to tell.

In sum, while the explainable components and rulemaking record can give agenda setters a good grasp on accountability and some handle on potential generalizability problems, there is no doubt that both rulemakers and agenda setters will more effectively anticipate generalizability problems if they can simply read the rule. Whether such lingering generalizability concerns outweigh the benefits of using an inscrutable machine-learning-based tool for particular decision criteria in a particular context can only be a normative judgment. Agenda setters—including, where appropriate, the public—should have the final say on that trade-off.

3. *Communication and Coordination Among Rulemakers.* — In conventional rulemaking, explanations created for agenda setters may be sufficient to facilitate communication and coordination among rulemakers. Incorporating a machine-learning-based tool into a decision system increases the challenges of communication and coordination, however, because of the disciplinary barriers between substantive experts and data scientists. Those barriers both heighten the importance of explanation and increase its difficulty. Data scientists differ from traditional rulemakers in three respects: (i) they are tool-building specialists, rather than subject matter specialists; (ii) they often do not work for the rulemaking entity;[141] and (iii) trade secrecy claims and confidentiality agreements can constrain their interactions with substantive rulemakers.[142]

Because data scientists are information technologists, rulemakers may be tempted to view their work as a technical task akin to those as-

---

138. See id.

139. See id.

140. Id. at 1722.

141. See supra note 116 and accompanying text.

142. See Brauneis & Goodman, supra note 11, at 153 ("The owners of proprietary algorithms will often require nondisclosure agreements from their public agency customers and assert trade secret protection over the algorithm and associated development and deployment processes.").

signed to an IT department.[143] Machine learning model development is deeply entangled with subject matter expertise and normative choices, however.[144] Data scientists' role is thus more like that of empirical economists, who are also technical specialists whose methods have broad application. Good economic modeling requires considerable substantive knowledge, however, which economists must access by collaborating with substantive experts or acquiring substantive expertise. Because they are highly contextual, economic models cannot simply be used off the shelf. Before porting them over to new situations, their underpinnings must be scrutinized to determine whether they can be appropriately adapted for use in those situations. Data scientists' design decisions are even more substantively fraught because the inscrutable models they create are used directly for assessing decision criteria. As a result, the substantive, normative, and policy assumptions underlying these choices have a direct impact on decision outcomes.

Though close communication and coordination between data scientists and substantive experts is critical, each group's unfamiliarity with the other's disciplinary knowledge will tend to impede it. When the development of automated decision tools is outsourced, those difficulties inevitably mount. Confidentiality agreements and trade secrecy claims keep information from rulemakers and discourage open communication, which only makes matters worse.[145] Documentation, user manuals, and training are traditional forms of explanation between software engineers and their clients.[146] While they may be sufficient for *users*, those explanatory forms are unlikely to facilitate the close communication and coordination required for *codevelopment* of decision criteria that incorporate machine-learning-based decision tools.

C. *Reason Giving, Democracy, Respect, and Machine Learning*

Some view the use of automated decision tools as inherently dehumanizing or disrespectful, at least in some contexts.[147] Here I do not adopt that view and hence consider whether the inscrutability of machine-learning-based decision tools creates problems for democratic and human values even when conventional rule-like decision criteria would

---

143. See Kate Crawford, The Hidden Biases in Big Data, Harv. Bus. Rev. (Apr. 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data [https://perma.cc/2KH9-5SFB] (explaining the tendency to view data science as objective and infallible).

144. See id. ("Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations.").

145. See Brauneis & Goodman, supra note 11, at 153–54.

146. See Technical Documentation in Software Development: Types, Best Practices, and Tools, AltexSoft, https://www.altexsoft.com/blog/business/technical-documentation-in-software-development-types-best-practices-and-tools/ [https://perma.cc/S92L-99AH] (last updated Mar. 26, 2019).

147. See, e.g., Noto La Diega, supra note 90, at 10–11.

have been acceptable. Though democratic legitimacy and dignitary concerns are part of the standard reasons for requiring government decisionmakers to provide explanations,[148] complete explanations of all government decisions have never been required. Machine-learning-based decision tools can be explained in a limited sense, as just discussed. The question, then, is when those limited explanations are sufficient in light of these sorts of concerns.

The answer to this question is likely to depend on many contextual factors, including the nature of the decision and what is at stake, the justification for automating a particular aspect of the decision criteria, the way in which adjudicators are expected to use the automated output in coming to final decisions, and, crucially, the extent to which explainable aspects of the automated tool are, in fact, explained. Citizens will likely have or develop a sense of whether the limited explanations available in a particular context are sufficient in light of these legitimacy and dignitary values. This sort of evaluation is likely to incite controversy but is not terribly different from the normative assessments that currently go into determining whether rules are appropriately employed in various contexts or what level of "due process" is appropriate for a particular decision. Rulemakers should, however, be prepared for the possibility that using inscrutable machine learning models for some sorts of decision criteria will be normatively unacceptable to the citizenry, regardless of how well-validated the machine learning model might be.

IV. EXPLANATION FOR RULEMAKING

This concluding Part considers how to obtain as many of the traditional benefits of explanation as possible for decision systems that incorporate machine-learning-based decision tools. As noted earlier, machine learning's now-canonical "explainability" problem pertains only to the model's computational mapping between features and outcome variables.[149] While this inscrutability is significant, many societally significant aspects of the development of a machine-learning-based decision tool, its final form, and its integration into a decision system are just as explainable as conventional narrative rules and their underpinnings. In particular, the choice to employ a machine-learning-based decision tool to evaluate particular decision criteria is fully explainable and has significant normative and policy implications that should be open to scrutiny.

Section IV.A thus argues for applying traditional explanation requirements to the explainable aspects of such systems. Section IV.B focuses on the less-discussed issues of communication and coordination within the rulemaking entity, pointing out that these issues require more attention when automated decision tools are introduced because of the

148. See supra Part II.

149. See supra section I.B.2.

disciplinary barriers between subject matter experts and data scientists within the rulemaking entity. Section IV.C suggests mechanisms for improving the capacity for rulemaking entities and advocacy groups to make full use of the explanations that would be made available to them under the explanation requirements proposed in section IV.A.[150] While large rulemaking entities and advocacy groups may have sufficient resources to obtain the fairly minimal data science expertise necessary for this purpose, smaller rulemaking entities and advocacy groups might consider pooling resources—though perhaps not with one another—to gain access to it.

A. *Explaining the Incorporation of Machine-Learning-Based Decision Tools*

What sort of explanation should be required when a machine-learning-based decision tool is incorporated into a decision system? In particular, how should this question be answered when the tool is incorporated into decision criteria that operate as a "rule" under the APA?

When a conventional narrative rule is published in the Federal Register for comment, the public receives full notice of its terms.[151] For rule-like criteria, publication allows the public to determine—and critique—how cases of any imaginable sort would be handled.[152] Because inscrutable machine-learning-based decision tools cannot be summarized in narrative form (or even in understandable mathematical or graphical form), there is no way to provide an equivalently detailed mapping from cases to outcomes. If notice and comment demands this sort of detailed mapping, inscrutable decision tools simply cannot be incorporated into APA rules.

While "just say no" to inscrutable decision tools is certainly an appropriate approach in some decision contexts, we should be wary of adopting it as a general response to notice and comment requirements or other explanation mandates. Because machine-learning-based decision tools are attractive to policymakers,[153] an overly expansive interpretation of *what* explanation requires might backfire by motivating rulemakers and courts to adopt narrower interpretations of *whether* such requirements apply at all. Moreover, preemptively depriving society of all such tools for all purposes in all significant decision contexts seems questionable as a policy matter, given the advantages of machine-learning-based decision tools in some contexts.

---

150. Proposals for algorithmic impact assessments would produce similar results. See, e.g., Dillon Reisman, Jason Schultz, Kate Crawford & Meredith Whittaker, AI Now Inst., Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability 7–20 (2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/A3QZ-PCY6]; Andrew D. Selbst, Disparate Impact in Big Data Policing, 52 Ga. L. Rev. 109, 169–82 (2017); Selbst & Barocas, supra note 4, at 1133–38.

151. Guide to Rulemaking, supra note 72.

152. See id.

153. See supra note 18 and accompanying text.

An opposite approach, which may be closer to what is happening on the ground,[154] is to pretend that machine-learning-based decision tools are not really rules at all but something else that does not have to be explained.[155] This approach is, if anything, worse because it deprives society of the benefits of explanation for the aspects that can be explained and ignores the true rule-like nature of the tools.

The analysis here suggests an intermediate approach: Define what constitutes an adequate explanation of a machine-learning-based decision tool and require such an explanation, thus subjecting the incorporation of inscrutable machine learning models to scrutiny while not barring it entirely. This section proposes a framework for adequate explanation composed of two parts: (1) information required to *describe the rule* and (2) information treated as part of the *record* for backing up the rule, as in hard look review.[156] Standard administrative law requirements of notice and recordkeeping could be interpreted in these terms.

1. *Describing the Rule.* — An adequate description of machine-learning-based decision criteria—as would be published in the Federal Register for notice and comment—would include all of the "explainable components" of the rule.[157] Those components are part and parcel of the decisionmaking rule and should be treated as such. They are no more difficult to explain or to understand than conventional narrative rules, and their disclosure fulfills the intended functions of explanation requirements.[158] When disclosure and explanation of a rule is legally re-

---

154. The opacity surrounding current government practices makes it difficult to know precisely how these tools are treated. See, e.g., David Curie, AI in the Regulatory State: Stanford Project Maps the Use of Machine Learning and Other AI Technologies in Federal Agencies, Thomson Reuters (June 20, 2019), https://blogs.thomsonreuters.com/answerson/ai-in-the-regulatory-state/ [https://perma.cc/L4QV-UK6G] (noting the relative differences in the integration of artificial intelligence across government agencies); Colin Lecher, New York City's Algorithm Task Force Is Fracturing, The Verge (Apr. 15, 2019), https://www.theverge.com/2019/4/15/18309437/new-york-city-accountability-task-force-law-algorithm-transparency-automation [https://perma.cc/3LNR-U67A] (noting the lack of transparency around the work of the algorithm task force in New York City).

155. In sentencing proceedings, for example, recidivism risk assessments seem to be treated as a sort of factual evidence. See, e.g., State v. Loomis, 881 N.W.2d 749, 772 (Wis. 2016) (holding that the use of a risk assessment tool at sentencing did not violate defendant's due process rights). For a critique of the treatment of recidivism risk assessment in *Loomis*, see generally Anne L. Washington, How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate, 17 Colo. Tech. L.J. 131 (2018) (arguing for a "new form of reasoning . . . to explain and justify algorithmic results"). For a critique of the treatment of predictive risk assessments as evidence, see generally Steve T. Mckinlay, Evidence, Explanation and Predictive Data Modelling, 30 Phil. & Tech. 461 (2017) ("[T]he claim that [predictive risk models] provide anything close to epistemically justified evidence in a traditional philosophical sense is dubious at best.").

156. Gillis and Simons come to similar conclusions under the GDPR, though by a different path. See Gillis & Simons, supra note 9 (manuscript at 20–26).

157. See supra section III.A.1.

158. See supra section III.A.1.

quired, trade secrecy should not excuse explanation of these aspects, which reflect critical, policy-relevant rulemaking choices.

2. *The Rulemaking Record.* — Selecting training data and validating the tool's performance bear the same relationship to developing a machine-learning-based decision tool that more familiar sorts of factual inquiry and statistical analysis bear to the development and justification of a conventional rule. Summary information about the training data, explanations of how it was sourced, descriptions of the validation process, and validation results should thus be part of the rulemaking record and made available on the same terms as other parts of the record. Rulemaking entities should not sign confidentiality agreements regarding this information. However, the training data set itself should ordinarily be kept confidential for privacy reasons. Confidentiality agreements regarding certain technical parameters and details of the machine learning process might also be appropriate.

B. *Coordination and Communication Between Data Scientists and Substantive Rulemakers*

To promote informed, effective overall decision-criteria design, data scientists need a deep understanding of both the overall goals of the decision system and how the criteria evaluated by the automated tools will be incorporated into the ultimate decision. Concomitantly, substantive rulemakers need to acquire a basic understanding of the machine learning process so that they can make appropriate choices about whether to automate particular decision criteria, interact meaningfully with data scientists throughout the development process, and design appropriate combination schemes for adjudicators to follow when using the outputs of machine-learning-based tools.

The incentives provided by the proposed explanation requirements will go some way toward facilitating the requisite communication and coordination between data scientists and substantive experts. But because these interactions are of such vital importance to decision quality and face significant barriers, more may be necessary. Rulemakers who are considering incorporating a machine-learning-based tool into a decision system would therefore be well advised to adopt a prospective "by design" plan aimed at ensuring the necessary level of cooperation and communication between data scientists and substantive rulemakers.

Ideally, the development of machine-learning-based decision tools would be brought in-house, so that dedicated data scientists could develop substantive expertise to support their work. That approach is probably overly ambitious for most rulemaking entities, who would only be undertaking such development on a sporadic basis. Larger entities should still consider hiring an in-house data scientist whose role would be not only to advise the rulemaking entity in its interactions with outside

contractors but also to initiate and facilitate the necessary close interactions between rulemakers and outside data scientists.

At a minimum, where an automated decision tool is procured from outside data scientists, substantive rulemakers must demand clear and thorough explanations of the aspects described in section III.A so that they can understand the outputs of the automated decision tool and create appropriate combination schemes for adjudicators to use. The depth of information that is available about the automated tool constrains the sorts of combination schemes that adjudicators can implement. Clear communication and coordination between data scientists and substantive rulemakers are critical to assessing the severity of those constraints. As discussed above, inscrutability is especially likely to limit the extent to which adjudicators can serve the role in addressing generalizability issues that is commonly assigned to them in conventional decision systems. Rulemakers must understand and confront these and other trade-offs involved in using inscrutable decision tools.

C. *Centers of Data Science Expertise for Rulemaking Entities and Advocacy Groups*

While larger rulemaking entities may be able to develop data science expertise to help them communicate and coordinate with the data science contractors who will probably continue to do most development of machine-learning-based decision tools, smaller rulemaking entities will likely be strapped to find the necessary resources. This is a problem because smaller entities are perhaps most likely to be attracted to the potential cost-savings of automation, while also being least able to afford to acquire data science expertise, increasing the temptation to use off-the-shelf solutions. It would be wise for smaller rulemaking entities to develop mechanisms for pooling resources with similarly situated entities to provide access to data science expertise. Ideally, such pooling would bring rulemaking entities in similar substantive arenas together so that the data scientists they work with could also build up substantive expertise. This proposal is tentative; its feasibility would depend on working out the details. If this sort of resource pooling is feasible, it should receive public support, and participation might even be mandated. Advocacy groups and other agenda setters in a given arena may also benefit from creating similar centers of data science expertise to assist them in understanding the explanations provided by rulemakers and ensuring accountability.

CONCLUSION

Delegated, distributed decision systems—which are responsible for many highly consequential decisions affecting individuals—confront issues of cost, efficiency, and consistency that make automated decision tools particularly attractive. Though scholars and policymakers have fo-

cused on explanations to decision subjects and accountability to the public, the inscrutability of automated decision tools has significant, and underappreciated, implications for the explanatory flows required to develop and implement such systems. While sharing the explainable aspects of these tools can replicate some of explanation's traditional functions, using inscrutable automated decision tools inevitably degrades decision-criteria development in some respects. Thus, in weighing the advantages and disadvantages of such tools for a given decision context, policymakers and system designers should consider how inscrutability affects rulemakers and, as I discuss elsewhere, adjudicators, along with its direct impact on decision subjects.