THE INTUITIVE APPEAL OF EXPLAINABLE MACHINES

Andrew D. Selbst & Solon Barocas***

Algorithmic decision-making has become synonymous with inexplicable decision-making, but what makes algorithms so difficult to explain? This Article examines what sets machine learning apart from other ways of developing rules for decision-making and the problem these properties pose for explanation. We show that machine learning models can be both inscrutable and nonintuitive and that these are related, but distinct, properties.

Calls for explanation have treated these problems as one and the same, but disentangling the two reveals that they demand very different responses. Dealing with inscrutability requires providing a sensible description of the rules; addressing nonintuitiveness requires providing a satisfying explanation for why the rules are what they are. Existing laws like the Fair Credit Reporting Act (FCRA), the Equal Credit Opportunity Act (ECOA),

---

* Postdoctoral Scholar, Data & Society Research Institute; Visiting Fellow, Yale Information Society Project. Selbst is grateful for the support of the National Science Foundation under grant IIS 1633400.

** Assistant Professor, Cornell University, Department of Information Science. For helpful comments and insights on earlier drafts, the authors would like to thank Jack Balkin, Rabia Belt, danah boyd, Kiel Brennan-Marquez, Albert Chang, Danielle Citron, Julie Cohen, Lilian Edwards, Sorelle Freidler, Giles Hooker, Margaret Hu, Karen Levy, Margot Kaminski, Rónán Kennedy, Been Kim, Jon Kleinberg, Brian Kreiswirth, Chandler May, Brent Mittelstadt, Deidre Mulligan, David Lehr, Paul Ohm, Helen Nissenbaum, Frank Pasquale, Nicholson Price, Manish Raghavan, Aaron Rieke, David Robinson, Ira Rubinstein, Matthew Salganik, Katherine Strandburg, Sandra Wachter, Hanna Wallach, Cody Marie Wild, Natalie Williams, Jennifer Wortman Vaughan, Michael Veale, Suresh Venkatasubramanian, and participants at the following conferences and workshops: *NYU Innovation Colloquium*, NYU School of Law, February 2017; *We Robot*, Yale Law School, March 2017; *Big Data Ethics Colloquium*, The Wharton School, Philadelphia, PA, April 2017; *NYU Algorithms and Explanations Conference*, NYU School of Law, April 2017; *TILTing Perspectives*, Tilburg University, the Netherlands, May 2017; *Privacy Law Scholars' Conference*, Berkeley, CA, June 2017; *Summer Faculty Workshop*, Georgetown University Law Center, June 2017; *Explainable and Accountable Algorithms Workshop*, Alan Turing Institute, UK, January 2018. Special thanks to Chandler May for graphics that sadly did not make it into the final draft, and to the editors of the *Fordham Law Review* for their excellent and professional work getting this Article ready for publication.

and the General Data Protection Regulation (GDPR), as well as techniques within machine learning, are focused almost entirely on the problem of inscrutability. While such techniques could allow a machine learning system to comply with existing law, doing so may not help if the goal is to assess whether the basis for decision-making is normatively defensible.

In most cases, intuition serves as the unacknowledged bridge between a descriptive account to a normative evaluation. But because machine learning is often valued for its ability to uncover statistical relationships that defy intuition, relying on intuition is not a satisfying approach. This Article thus argues for other mechanisms for normative evaluation. To know why the rules are what they are, one must seek explanations of the process behind a model's development, not just explanations of the model itself.

> There can be no total understanding and no absolutely reliable test of understanding.
>
> —Joseph Weizenbaum, "Contextual Understanding by Computers"[1]

INTRODUCTION

Algorithms increasingly inform consequential decisions about our lives, with only minimal input from the people they affect and little to no explanation as to how they work.[2] This worries people, and rightly so. The results of these algorithms can be unnerving,[3] unfair,[4] unsafe,[5] unpredictable,[6] and unaccountable.[7] How can decision makers who use algorithms be held to account for their results?

It is perhaps unsurprising that, faced with a world increasingly dominated by automated decision-making, advocates, policymakers, and legal scholars would call for machines that can explain themselves.[8] People have a natural

---

1. 10 COMM. ACM 474, 476 (1967). In the 1960s, the project of artificial intelligence (AI) was largely to mimic human intelligence. Weizenbaum was therefore actually arguing that computers will never fully understand humans. The purpose of AI research has changed drastically today, but there is a nice symmetry in the point that humans will never have total understanding of computers.

2. Aaron M. Bornstein, *Is Artificial Intelligence Permanently Inscrutable?*, NAUTILUS (Sept. 1, 2016), http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable [http://perma.cc/RW3E-5CPV]; Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [http://perma.cc/7VYF-5XR7]; Cliff Kuang, *Can A.I. Be Taught to Explain Itself?*, N.Y. TIMES (Nov. 21, 2017), https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html [http://perma.cc/3CYF-QTVC].

3. *See, e.g.*, Omer Tene & Jules Polonetsky, *A Theory of Creepy: Technology, Privacy, and Shifting Social Norms*, 16 YALE J.L. & TECH. 59, 65–66 (2013); Sara M. Watson, *Data Doppelgängers and the Uncanny Valley of Personalization*, ATLANTIC (June 16, 2014), https://www.theatlantic.com/technology/archive/2014/06/data-doppelgangers-and-the-uncanny-valley-of-personalization/372780/ [http://perma.cc/7J3X-NK3C].

4. *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677–92 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 883–89 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 126–39 (2017).

5. *See, e.g.*, David Lazer et al., *The Parable of Google Flu: Traps in Big Data Analysis*, 343 SCIENCE 1203, 1203 (2014); Jennings Brown, *IBM Watson Reportedly Recommended Cancer Treatments That Were 'Unsafe and Incorrect*,*'* GIZMODO (July 25, 2018, 3:00 PM), https://gizmodo.com/ibm-watson-reportedly-recommended-cancer-treatments-tha-1827868882 [http://perma.cc/E4RZ-NVZU].

6. *See, e.g.*, Curtis E. A. Karnow, *The Application of Traditional Tort Theory to Embodied Machine Intelligence*, *in* ROBOT LAW 51, 57–58 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016) (discussing unpredictability in autonomous systems); Jamie Condliffe, *Algorithms Probably Caused a Flash Crash of the British Pound*, MIT TECH. REV. (Oct. 7, 2016), https://www.technologyreview.com/s/602586/algorithms-probably-caused-a-flash-crash-of-the-british-pound/ [https://perma.cc/K9FM-6SJE].

7. *See, e.g.*, Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 18–27 (2014); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 636–37 (2017).

8. *See infra* Part II.

feel for explanation. We know how to offer explanations and can often agree when one is good, bad, in-between, on point, or off topic. Lawyers use explanation as their primary tradecraft: judges write opinions, administrators respond to comments, litigators write briefs, and everyone writes memos. Explanations are the difference between a system that vests authority in lawful process and one that vests it in an unaccountable person.[9]

Although we comfortably use explanations, asking someone to define the concept will often generate a blank look in response. Analytically, explanation is infinitely variable, and there can be many valid explanations for a given phenomenon or decision. Thus far, in both law and machine learning, the scholarly discourse around explanation has primarily revolved around two questions: Which kinds of explanations are most useful, and which are technically available?[10] Yet, these are the wrong questions or, at least, the wrong stopping points.

Explanations of technical systems are necessary but not sufficient to achieve law and policy goals, most of which are concerned not with explanation for its own sake, but with ensuring that there is a way to evaluate the basis of decision-making against broader normative constraints such as antidiscrimination or due process. It is therefore important to ask how exactly people engage with those machine explanations in order to connect them to the normative questions of interest to law.

This Article argues that scholars and advocates who seek to use explanation to enable justification of machine learning models are relying on intuition to connect the explanation to normative concerns. Intuition is both powerful and dangerous. While this mode of justifying decision-making remains important, we must understand the benefits and weaknesses of connecting machine explanation to intuitions. Remedying the limitations of intuition requires considering alternatives, which include institutional processes, documentation, and access to those documents.

This Article proceeds in four parts. Part I examines the various anxieties surrounding the use of automated decision-making. After discussing secrecy, lack of transparency, and lack of technical expertise, Part I argues that the two distinct, but similar, concepts that truly set machine learning decision-making apart are inscrutability and nonintuitiveness.

Part II examines laws and machine learning techniques designed specifically to address the problem of inscrutable decisions. On the legal side, Part II.A discusses the "adverse action notices" required by federal credit laws[11] and the informational requirements of the European Union's General Data Protection Regulation (GDPR).[12] On the technical side, Part

9. *See* Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633, 636–37 (1995).

10. *See infra* Part III.

11. This Article will focus on the Fair Credit Reporting Act, 15 U.S.C. §§ 1681–1681x, and Equal Credit Opportunity Act, 15 U.S.C. §§ 1691–1691f.

12. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 1 (EU) [hereinafter GDPR].

II.B discusses various techniques used by computer scientists to make machine learning models interpretable, including designing for simplicity, approximating complex models in simpler form, extracting the most important factors in a particular decision, and allowing some degree of interaction with the models to see how changes in inputs affect outputs. These techniques can be useful in meeting the requirements of the law, but such explanations, even when they comply with the law, may be of limited practical utility.

Part III builds the connection between explanation and intuition before evaluating the merits of an intuition-centered approach to justification. It canvasses reasons besides justification that one might want explainable machines—dignity or autonomy on the one hand and consumer or data-subject education on the other—before concluding that neither is adequate to fully address the concerns with automated decision-making. Interrogating the assumptions behind a third reason—that explanation will reveal problems with the basis for decision-making—demonstrates the reliance on intuition. The remainder of Part III examines the value and limitations of intuition. With respect to machine learning in particular, although intuition can root out obviously good or bad cases, it cannot capture the cases that give machine learning its greatest value: true patterns that exceed human imagination. These cases are not obviously right or wrong, but simply strange.

Part IV aims to provide another way. Once outside the black box, all that is left is to question the process surrounding its development and use. There are large parts of the process of machine learning that do not show up in a model but can contextualize its operation, such as paths considered but not taken and the constraints that influence these choices. Where intuition is insufficient to determine whether the model's rules are reasonable or rest on valid relationships, justification can sometimes be achieved by demonstrating and documenting due care and thoughtfulness.

I. INSCRUTABLE AND NONINTUITIVE

Scholarly and policy debates about regulating a world controlled by algorithms have been mired in difficult questions about how to observe, access, audit, or understand those algorithms.[13] The difficulty has been attributed to a diverse set of problems, specifically that algorithms are

13. *See, e.g.*, Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM. & SOC'Y 14 (2017) (evaluating methods of researching algorithms); Malte Ziewitz, *Governing Algorithms: Myth, Mess, and Methods*, 41 SCI. TECH. & HUM. VALUES 3 (2016); Solon Barocas, Sophie Hood & Malte Ziewitz, Governing Algorithms: A Provocation Piece (Mar. 29, 2013) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2245322 [https://perma.cc/DB7Z-C9A6]; Nick Seaver, Knowing Algorithms (Feb. 2014) (unpublished manuscript), https://static1.squarespace.com/static/55eb004ee4b0518639d59d9b/t/55ece1bfe4b030b2e8302e1e/1441587647177/seaverMiT8.pdf [https://perma.cc/7HG3-74U3].

"secret"[14] and "opaque"[15] "black boxes"[16] that are rarely, if ever, made "transparent";[17] that they operate on the basis of correlation rather than "causality"[18] and produce "predictions"[19] rather than "explanations";[20] that their behavior may lack "intelligibility"[21] and "foreseeability";[22] and that they challenge established ways of being "informed"[23] or "knowing."[24] These terms are frequently used interchangeably or assumed to have overlapping meanings. For example, opacity is often seen as a synonym for secrecy,[25] an antonym for transparency,[26] and, by implication, an impediment to understanding.[27] Yet the perceived equivalence of these terms has obscured important differences between distinct problems that frustrate attempts at regulating algorithms—problems that must be disentangled before the question of regulation can even be addressed.

This Part argues that many of these challenges are not unique to algorithms or machine learning. We seek here to parse the problems raised by machine learning models more precisely and argue that they have little to do with the fact that their very existence may be unknown, that their inner workings may be opaque, or that an understanding of their operations may require specialized knowledge. What sets machine learning models apart from other decision-making mechanisms are their *inscrutability* and *nonintuitiveness*.

---

14. *See, e.g.*, Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. ON TELECOMM. & HIGH TECH. L. 235, 236–37 (2011) (recounting the origins of using trade-secret protections for algorithms); Brenda Reddix-Smalls, *Credit Scoring and Trade Secrecy: An Algorithmic Quagmire or How the Lack of Transparency in Complex Financial Models Scuttled the Finance Market*, 12 U.C. DAVIS BUS. L.J. 87, 88–90 (2011) (discussing the use of trade-secret protections for algorithms, which result in lack of transparency concerning algorithmic decision-making).

15. Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan.–June 2016, at 1, 3–5; Roger Allan Ford & W. Nicholson Price II, *Privacy and Accountability in Black-Box Medicine*, 23 MICH. TELECOMM. & TECH. L. REV. 1, 11–12 (2016); Tal Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCI. TECH. & HUM. VALUES 118, 129 (2016).

16. *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY 8 (2015).

17. *See, e.g.*, Citron & Pasquale, *supra* note 7, at 27; Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1506.

18. *See, e.g.*, Kim, *supra* note 4, at 875.

19. Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1267–68 (2017).

20. *See, e.g.*, Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation,"* 38 AI MAG., Fall 2017, at 50, 55.

21. *See, e.g.*, Brennan-Marquez, *supra* note 19, at 1253.

22. *See, e.g.*, Karnow, *supra* note 6, at 52.

23. *See, e.g.*, Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 89–90 (2017).

24. Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 974–77 (2018).

25. *See, e.g.*, Burrell, *supra* note 15, at 3–4.

26. *See, e.g.*, Ford & Price, *supra* note 15, at 12; Zarsky, *supra* note 15, at 124.

27. *See, e.g.*, Burrell, *supra* note 15, at 4–5.

We adapt and extend a taxonomy first proposed by Jenna Burrell,[28] where our primary purpose is to emphasize these last two properties and clear up confusion.[29] Inscrutability and nonintuitiveness have been conflated in the past: where the property of inscrutability suggests that models available for direct inspection may defy understanding, nonintuitiveness suggests that even where models are understandable, they may rest on apparent statistical relationships that defy intuition.[30]

A. Secret

The first common critique of algorithmic decision-making is secrecy. Secrecy captures two related, but distinct, concerns: (1) secrecy of the model's existence and (2) secrecy of its operation.

The first concern is as old as the original Code of Fair Information Practices (FIPs), the conceptual basis for the majority of privacy laws:[31] "There must be no personal-data record-keeping systems whose very existence is secret."[32] This principle underlies more recent calls to "end secret profiling" involving algorithms and machine learning, where secrecy is understood as a purposeful attempt to maintain ignorance of the very fact of profiling.[33]

While such worries are particularly pronounced when the government engages in algorithmic decision-making,[34] similar objections arise in the commercial sector, where there are a remarkable number of scoring systems

28. *See generally id*.

29. Our parsing of the issues is similar to the taxonomy proposed by Ed Felten in a short blog post on *Freedom to Tinker*. Ed Felten, *What Does It Mean to Ask for an "Explainable" Algorithm?*, FREEDOM TO TINKER (May 31, 2017), https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/ [https://perma.cc/QF7B-RTC6].

30. We intentionally use the term "nonintuitive" rather the word "unintuitive" or "counterintuitive." In our view, "unintuitive" implies a result that would not be expected but is easily understood once explained, and "counterintuitive" suggests a phenomenon that is opposite one's expectations. Instead, we intend to refer to a phenomenon about which intuitive reasoning is not possible.

31. WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES 56 (2018); Robert Gellman, Fair Information Practices: A Basic History 3 (Apr. 10, 2017) (unpublished manuscript), https://bobgellman.com/rg-docs/rg-FIPshistory.pdf [https://perma.cc/CQ9E-HK9A] (discussing the history of the FIPs).

32. SEC'Y'S ADVISORY COMM. ON AUTOMATED PERS. DATA SYS., U.S. DEP'T OF HEALTH, EDUC. & WELFARE, RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS 41 (1973), https://www.justice.gov/opcl/docs/rec-com-rights.pdf [https://perma.cc/8TG8-FBL9]. In fact, the newly effective GDPR requires, among other things, disclosure of the "existence" of an automated decision-making tool. *See infra* note 142 and accompanying text.

33. *Algorithmic Transparency: End Secret Profiling*, ELECTRONIC PRIVACY INFO. CTR., https://epic.org/algorithmic-transparency/ [https://perma.cc/ZW4W-HKTM] (last visited Nov. 15, 2018); *see also* Margaret Hu, *Big Data Blacklisting*, 67 FLA. L. REV. 1735, 1745–46 (2015).

34. *See* Ira S. Rubinstein, Ronald D. Lee & Paul M. Schwartz, *Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches*, 75 U. CHI. L. REV. 261, 262–70 (2008); Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives*, 116 PENN ST. L. REV. 285, 295–97 (2011).

of which consumers are simply unaware.[35] In many cases, this ignorance exists because the companies engaged in such scoring are serving other businesses rather than consumers.[36] But the fact that more recent forms of hidden decision-making involve algorithms or machine learning does not change the fundamental secrecy objection—that affected parties are not aware of the existence of the decision-making process.[37]

The second secrecy concern arises where the existence of a decision-making process is known, but its actual operation is not. Affected parties might be aware that they are subject to such decision-making but have limited or no knowledge of how the decision-making process works.[38] Among the many terms used to describe this situation, "opacity" seems most apt, as there is enough visibility to see that the model exists but not enough to discern any of its details.

While this is perhaps the most frequent critique of algorithms and machine learning—that their inner workings remain undisclosed or inaccessible[39]—it, too, has little to do with the technology specifically. It is an objection to being subject to a decision where the basis of decision-making remains secret, which is a situation that can easily occur without algorithms or machine learning.[40]

There are sometimes valid reasons for companies to withhold details about a decision-making process. Where a decision-making process holds financial and competitive value and where its discovery entails significant investment or ingenuity, firms may claim protection for its discovery as a trade secret.[41] Trade-secret protection applies only when firms purposefully restrict disclosure of proprietary methods,[42] which creates incentives for firms to maintain secrecy around the basis for decision-making. If the use of algorithms or machine learning uniquely increases up-front investment or competitive advantage, then the incentives to restrict access to the details of

35. *See* PAM DIXON & ROBERT GELLMAN, THE SCORING OF AMERICA: HOW SECRET CONSUMER SCORES THREATEN YOUR PRIVACY AND YOUR FUTURE 84 (2014), http://www.worldprivacyforum.org/wp-content/uploads/2014/04/WPF_Scoring_of_America_April2014_fs.pdf [https://perma.cc/39RJ-97M6].

36. *See* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY i (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [https://perma.cc/8HQY-6WVP].

37. SEC'Y'S ADVISORY COMM. ON AUTOMATED PERS. DATA SYS., *supra* note 32, at 29 (discussing the lack of awareness of record keeping and use of personal data).

38. This could refer to secrecy around what data is considered or how it is used. See *infra* Part II.A for a discussion of these concerns with respect to the Fair Credit Reporting Act.

39. *See, e.g.*, Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 107–08 (2018); Citron & Pasquale, *supra* note 7, at 10–11. *See generally* PASQUALE, *supra* note 16.

40. *See, e.g.*, Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1407, 1410 (2001) (discussing the private database industry and corporate decision-making based on consumer data).

41. Brauneis & Goodman, *supra* note 39, at 153–60. *See generally* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018).

42. Pasquale, *supra* note 14, at 237.

the decision-making process might be understood as peculiar to algorithms or machine learning. But if other attempts to develop decision-making processes without algorithms or machine learning involve similar costs and competitive advantage, then there is nothing special about the relationship between these technologies, trade secrets, and resistance to disclosure.[43]

Firms may also reject requests for further details about the basis for decision-making if they anticipate that such details may enable strategic manipulation, or "gaming," of the inputs to the decision-making process.[44] If the costs of manipulating one's characteristics or behavior are lower than the expected benefits, rational actors would have good incentive to do so.[45] Yet these dynamics, too, apply outside algorithms and machine learning; in the face of some fixed decision procedure, people will find ways to engage in strategic manipulation. The question is whether decision procedures developed with machine learning are easier or harder to game than those developed using other methods—this is not a question that can be answered in general.

B. Requiring Specialized Knowledge

One common approach to ensuring accountability for software-reliant decision-making is to require the disclosure of the underlying source code.[46] While such disclosure might seem helpful in figuring out how automated decisions are rendered, the ability to make sense of the disclosed source code will depend on one's level of technical literacy. Some minimal degree of training in computer programming is necessary to read code, although even that might not be enough.[47] The problem, then, is greater than disclosure; in

43. *See, e.g.*, David S. Levine, *Secrecy and Unaccountability: Trade Secrets in Our Public Infrastructure*, 59 FLA. L. REV. 135, 139 (2007) (describing the growing application of trade secrecy in various technologies used in public infrastructure).

44. Jane Bambauer & Tal Z. Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV (forthcoming 2018) (manuscript at 10), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3135949 [http://perma.cc/N62U-3UUK].

45. Whether such manipulation is even possible will vary from case to case, depending on the degree to which the decision considers immutable characteristics and nonvolitional behavior. At the same time, it is unclear how easily one could even change the appearance of one's characteristics without genuinely changing those characteristics in the process. Altering behavior to game the system might involve adjustments that actually change a person's likelihood of having the sought-after quality or experiencing the event that such behavior is meant to predict. To the extent that "gaming" is a term used to describe validating rather than defeating the objectives of a decision system, this outcome should probably not be considered "gaming" at all. *See* Bambauer & Zarsky, *supra* note 44.

46. Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 10 (2017); Kroll et al., *supra* note 7, at 647–50; *Algorithmic Transparency: End Secret Profiling*, *supra* note 33. Draft legislation in New York City also specifically focused on this issue, but the eventual bill convened a more general task force to consider different approaches. *See* Jim Dwyer, *Showing the Algorithms Behind New York City Services*, N.Y. TIMES (Aug. 24, 2017), https://www.nytimes.com/2017/08/24/nyregion/showing-the-algorithms-behind-new-york-city-services.html [https://perma.cc/38V5-P3EE].

47. Desai & Kroll, *supra* note 46, at 5 ("[F]undamental limitations on the analysis of software meaningfully limit the interpretability of even full disclosures of software source code."); Kroll et al., *supra* note 7, at 647.

the absence of the specialized knowledge required to understand source code, disclosure may offer little value to affected parties and regulators.

As Mike Ananny and Kate Crawford have observed, "Transparency concerns are commonly driven by a certain chain of logic: observation produces insights which create the knowledge required to govern and hold systems accountable."[48] The process of moving from observation to knowledge to accountability cannot be assumed; in many cases, the ability to leverage observations for accountability requires *preexisting* knowledge that allows observers to appreciate the significance of a disclosure.[49] Transparency of systems of decision-making is important, but incomplete.[50] But while cultivating the knowledge necessary to read source code requires time and effort, the problem of expertise—like the problem of secrecy—is not unique to algorithms.

C. Inscrutable

Rather than programming computers by hand with explicit rules, machine learning relies on pattern-recognition algorithms and a large set of examples to uncover relationships in the data that might serve as a reliable basis for decision-making.[51] The power of machine learning lies not only in its ability to relieve programmers of the difficult task of producing explicit instructions for computers, but in its capacity to learn subtle relationships in data that humans might overlook or cannot recognize. This power can render the models developed with machine learning exceedingly complex and, therefore, impossible for a human to parse.

We define this difficulty as "inscrutability"—a situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension. While there is a long history to such concerns, evidenced most obviously by the term "byzantine," the complexity of rules that result from machine learning can far exceed those of the most elaborate bureaucracy.[52] The challenge in such circumstances is not a lack of awareness, disclosure, or expertise, but the sheer scope and sophistication of the model.[53]

Intuitively, complexity would seem to depend on the number of rules encoded by a model, the length of a rule (i.e., the number of factors that figure into the rule), and the logical operations involved in the rule. These properties, however, can be specified more precisely. Four mathematical

48. Ananny & Crawford, *supra* note 24, at 974.

49. Burrell, *supra* note 15, at 4.

50. *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1254–55 (2008); Kroll et al., *supra* note 7, at 639, 657–60.

51. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 655 (2017).

52. *Byzantine*, MERRIAM-WEBSTER, http://www.merriam-webster.com/dictionary/Byzantine [https://perma.cc/97CM-KNT2] (last visited Nov. 15, 2018) (defining the term as "intricately involved").

53. Burrell, *supra* note 15, at 4–5.

properties related to model complexity are linearity, monotonicity, continuity, and dimensionality.

A linear model is one in which there is a steady change in the value of the output as the value of the input changes.[54] Linear models tend to be easier for humans to understand and interpret because the relationship between variables is stable and lends itself to straightforward extrapolation.[55] In contrast, the behavior of nonlinear models can be far more difficult to predict, even when they involve simple mathematical operations like exponential growth.[56]

A monotonic relationship between variables is a relationship that is either always positive or always negative.[57] That is, for every change in input value, the direction of the corresponding change in output value will remain consistent, whether an increase or decrease.[58] Monotonicity aids interpretability because it too permits extrapolation and guarantees that the value of the output only moves in one direction.[59] If, however, the value of the output goes up and down haphazardly as the value of the input moves steadily upward, the relationship between variables can be difficult to grasp or predict.

Discontinuous models include relationships where changes in the value of one variable do not lead to a smooth change in the associated value of another.[60] Discontinuities can render models far less intuitive because they make it impossible to think in terms of incremental change. A small change in input may typically lead to small changes in outputs, except for occasional and seemingly arbitrary large jumps.[61]

The dimensionality of a model is the number of variables it considers.[62] Two-dimensional models are easy to understand because they can be visualized graphically with a standard plot (with the familiar *x* and *y* axes).[63] Three-dimensional models also lend themselves to effective visualization (by adding a *z* axis), but humans have no way to visualize models with more than three dimensions.[64] While people can grasp relationships between multiple

54. Mathematically, this means that the function is described by a constant slope, which can be represented by a line. Yin Lou et al., *Intelligible Models for Classification and Regression*, *in* PROCEEDINGS OF THE 18TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 150, 150 (2012).

55. *See id.* at 151.

56. *Cf.* DEMI, ONE GRAIN OF RICE: A MATHEMATICAL FOLKTALE (1997).

57. *See Monotonicity Function*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

58. *See id.*

59. *See id.*

60. *See Continuous Function*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014) (noting that a continuous function does not suddenly jump at a given point or take widely differing values arbitrarily close to that point).

61. *See Discontinuity*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

62. *See Dimension (Dimensionality)*, A DICTIONARY OF COMPUTER SCIENCE (7th ed. 2016).

63. *See Cartesian Plane*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

64. *See Cartesian Space*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014); n-*Dimensional Space*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

variables without the aid of a graph, we will struggle to understand the full set of relationships that the model has uncovered as the number of dimensions grows. The more variables that the model includes, the more difficult it will be to keep all the interactions between variables in mind and thus predict how the model would behave given any particular input.[65]

In describing how these properties of models might frustrate human understanding, we have relied on terms like intuition, extrapolation, and prediction. The same cognitive capacity underlies all three: mentally simulating how a model turns inputs into outputs.[66] As computer scientist Zachary Lipton explains, simulatability—the ability to practically execute a model in one's mind—is an important form of understanding a model.[67] Such simulations can be either complete or partial. In the former, a person is able to turn any combination of inputs into the correct outputs, while in the latter, understanding might be limited to the relationships between a subset of input and output variables (i.e., how changes in certain inputs affect the output).

Simulation is a remarkably flat and functional definition of understanding, but it seems like a minimum requirement for any more elaborate definition.[68] This notion of understanding has nothing to say about *why* the model behaves the way it does; it is simply a way to account for the facility with which a person can play out how a model would behave under different circumstances. When models are too complex for humans to perform this task, they have reached the point of inscrutability.

D. Nonintuitive

A different line of criticism has developed that takes issue with disclosures that reveal some basis for decision-making that defies human intuition about the relevance of certain variables.[69] The problem in such cases is not

65. *See* Lehr & Ohm, *supra* note 51, at 700.

66. Zachary C. Lipton, *The Mythos of Model Interpretability*, *in* PROCEEDINGS OF THE 2016 ICML WORKSHOP ON HUMAN INTERPRETABILITY IN MACHINE LEARNING 96, 98 (2016).

67. *Id.*

68. While we limit our discussion to simulatability, inscrutability is really a broader concept. In particular, models might be difficult to understand if they consider features or perform operations that do not have some ready semantic meaning. Burrell, *supra* note 15, at 10. For example, a deep-learning algorithm can learn on its own which features in an image are characteristic of different objects (the standard example being cats). Bornstein, *supra* note 2. Part III.A.3, *infra*, returns to one such example that involves distinguishing between wolves and huskies. *See infra* notes 246–47 and accompanying text. An algorithm will usually learn to detect edges that differentiate an object from its background, but it might also engineer features on its own that have no equivalent in human cognition and therefore defy description. *See* Lipton, *supra* note 66, at 98 (discussing decomposability). This aspect of inscrutability, however, is of slightly less concern for this Article. Most methods that are common in the kinds of applications that apportion important opportunities (e.g., credit) involve features that have been handcrafted by experts in the domain (e.g., length of employment) and accordingly will usually not face this problem. *See infra* note 120 and accompanying text.

69. Deborah Gage, *Big Data Uncovers Some Weird Correlations*, WALL ST. J. (Mar. 23, 2014, 4:36 PM), https://www.wsj.com/articles/big-data-helps-companies-find-some-surprising-correlations-1395168255 [https://perma.cc/8KYB-LP9W]; Quentin Hardy,

inscrutability, but an inability to weave a sensible story to account for the statistical relationships in the model.[70] Although the statistical relationship that serves as the basis for decision-making might be readily identifiable, that relationship may defy intuitive expectations about the relevance of certain criteria to the decision.[71] As Paul Ohm explains:

> We are embarking on the age of the impossible-to-understand reason, when marketers will know which style of shoe to advertise to us online based on the type of fruit we most often eat for breakfast, or when the police know which group in a public park is most likely to do mischief based on the way they do their hair or how far from one another they walk.[72]

Even though it is clear which statistical relationships serve as the basis for decision-making in this case, why such statistical relationships exist is mystifying. This is a crucial and consistent point of confusion. The demand for intuitive relationships is not the demand for disclosure or accessible explanations; it is a demand that decision-making rely on reasoning that comports with intuitive understanding of the phenomenon in question. In social science, similar expectations are referred to as "face validity"—the subjective sense that some measure is credible because it squares with our existing understanding of the phenomenon.[73] While such demands are not unique to algorithms and machine learning, the fact that such computational tools are designed to uncover relationships that defy human intuition explains why the problem will be particularly pronounced in these cases.

Critics have pinned this problem on the use of "[m]ere correlation"[74] in machine learning, which frees it to uncover reliable, if incidental, relationships in the data that can then serve as the basis for consequential decision-making.[75] Despite being framed as an indictment of correlational analysis, however, it is really an objection to decision-making that rests on particular correlations that defy familiar causal stories[76]—even though these stories may be incorrect.[77] This has led to the mistaken belief that forcing

Bizarre Insights from Big Data, N.Y. TIMES: BITS (Mar. 28, 2012, 8:17 PM), https://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/ [https://perma.cc/GKW2-KN8T].

70. *See* Brennan-Marquez, *supra* note 19, at 1280–97.

71. *See* Paul Ohm, *The Fourth Amendment in a World Without Privacy*, 81 MISS. L.J. 1309, 1318 (2012).

72. *Id.*

73. *See generally* Ronald R. Holden, *Face Validity*, *in* 2 CORSINI ENCYCLOPEDIA OF PSYCHOLOGY 637 (Irving B. Weiner & W. Edward Craighead eds., 4th ed. 2010).

74. Kim, *supra* note 4, at 875, 883.

75. *Id.*; *see also* James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 173 (2016).

76. *See* Brennan-Marquez, *supra* note 19, at 1280–97.

77. *See* DANIEL KAHNEMAN, THINKING FAST AND SLOW 199–200 (2011) (discussing the "narrative fallacy"); *id.* at 224 ("Several studies have shown that human decision makers are inferior to a prediction formula even when they are given the score suggested by the formula! They feel that they can overrule the formula because they have additional information about the case, but they are wrong more often than not.").

decision-making to rest on causal mechanisms rather than mere correlations will ensure intuitive models.[78]

Causal relationships can be exceedingly complex and nonintuitive, especially when dealing with human behavior.[79] Indeed, causal relationships uncovered through careful experimentation can be as elaborate and unexpected as the kinds of correlations uncovered in historical data with machine learning.[80] If we consider all the different factors that cause a person to take an action—mood, amount of sleep, food consumption, rational choice, and many other things—it quickly becomes clear that causality is not particularly straightforward.[81] The only advantage of models that rely on causal mechanisms in such cases would be the reliability of their predictions (because the models would be deterministic rather than probabilistic), not the ability to interrogate whether the identified causal relationships comport with human intuitions and values. Given that much of the interest in causality stems from an unwillingness to simply defer to predictive accuracy as a justification for models, improved reliability will not be a satisfying answer.

* * *

The demand for intuitive relationships reflects a desire to ensure that there is a way to assess whether the basis of decision-making is sound, as a matter of validity and as a normative matter. We want to be able to do more than simply simulate a model; we want to be able to *evaluate* it. One way to ensure this possibility is to force a model to rely exclusively on features that bear a manifest relationship to the outcome of interest, on the belief that well-justified decisions are those that rest on relationships that conform to familiar and permissible patterns.

Achieving this type of intuitiveness requires addressing inscrutability as a starting point. An understandable model is necessary because there can be nothing intuitive about a model that resists all interrogation. But addressing inscrutability is not sufficient. A simple, straightforward model might still defy intuition if it has not been constrained to only use variables with an intuitive relationship to the outcome.[82]

---

78. These critiques also presume that causal mechanisms that exhaustively account for the outcomes of interest actually exist (e.g., performance on the job, default, etc.), yet certain phenomena might not be so deterministic; extrinsic random factors may account for some of the difference in the outcomes of interest. Jake M. Hofman, Amit Sharma & Duncan J. Watts, *Prediction and Explanation in Social Systems*, 355 SCIENCE 486, 488 (2017).

79. *Id.*

80. *See id.*

81. Attempts to model causation require limiting the features considered as potential causes because, to a certain extent, almost any preceding event could conceivably be causally related to the later one. JUDEA PEARL, CAUSALITY: MODELS, REASONING AND INFERENCE 401–28 (2d ed. 2009).

82. *See, e.g.*, Jiaming Zeng, Berk Ustun & Cynthia Rudin, *Interpretable Classification Models for Recidivism Prediction*, 180 J. ROYAL STAT. SOC'Y 689 (2017). Note that in this and related work, the researchers limit themselves to features that are individually and intuitively related to the outcome of interest. *See id.* at 693–97. If these methods begin with features that do not have such a relationship, the model might be simple enough to inspect but too strange to square with intuition. *See infra* Part III.B.

Insisting on intuitive relationships is not the only way to make a model evaluable. To the extent that intuitiveness is taken to be an end in itself rather than a particular means to the end of ensuring sound decision-making, its proponents risk overlooking other, potentially more effective, ways to achieve the same goal. The remainder of this Article considers the different paths we might take to use explanations of machine learning models to regulate them.

II. LEGAL AND TECHNICAL APPROACHES TO INSCRUTABILITY

This moment is not the first time that law and computer science have attempted to address algorithmic decision-making with explanation requirements. Credit scoring has long been regulated, in part, by requiring "adverse action notices," which explain adverse decisions to consumers.[83] In Europe, concern about automated decisions has been a neglected part of data protection law for more than two decades, with interest in them reinvigorated by the GDPR.[84] On the machine learning side, the subfield of "interpretability"—within which researchers have been attempting to find ways to understand complex models—is over thirty years old.[85]

What seems to emerge from the law and computer science is a focus on two kinds of explanation. The first concerns accounting for outcomes—how particular inputs lead to a particular output. The second concerns the logic of decision-making—full or partial descriptions of the rules of the system. This Part reviews the legal and technical approaches to outcome and logic-based explanations.

A. Legal Requirements for Explanation

Though much of the current concern over inscrutable systems stems from the growing importance of machine learning, inscrutable systems predate this technique. As a result, regulations that require certain systems to explain themselves already exist. This section discusses two examples of legal systems and strategies that rely on different types of explanations: credit reporting statutes, which rely on outcome-based explanations, and the GDPR, which mandates logic-based explanations. Credit scoring predates machine learning, and is governed by two statutes: the Fair Credit Reporting Act (FCRA)[86] and the Equal Credit Opportunity Act (ECOA).[87] Statistical credit-scoring systems take information about consumers as inputs, give the

83. *See infra* notes 100–01 and accompanying text.

84. *See* Directive 95/46 of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, art. 3(1), 1995 O.J. (L 281) 31, 39 (EC) [hereinafter Data Protection Directive].

85. *See, e.g.*, William van Melle, Edward H. Shortliffe & Bruce G. Buchanan, *EMYCIN: A Knowledge Engineer's Tool for Constructing Rule-Based Expert Systems*, *in* RULE-BASED EXPERT SYSTEMS: THE MYCIN EXPERIMENT OF THE STANFORD HEURISTIC PROGRAMMING PROJECT 302 (Bruce G. Buchanan & Edward H. Shortliffe eds., 1984).

86. 15 U.S.C. §§ 1681–1681x (2012).

87. 15 U.S.C. §§ 1691–1691f (2012).

inputs certain point values, add them to obtain a total score, and then make decisions based on that score. Each of these statutes require "adverse action notices" that must include a statement of reasons for denials of credit or other credit-based outcomes.[88] This is an example of what we call outcome-based explanations: a description of the facts that proved relevant to a decision, but not a description of the decision-making rules themselves.

Articles 13–15 of the GDPR require data subjects to have access to "meaningful information about the logic involved" in any automated decision-making that significantly affects them.[89] As the law is still new, the import and proper interpretation of this requirement remain unclear. In advance of a definitive interpretation, the GDPR appears to ask for a functional description of the model—enough of a description of the rules governing decision-making such that a data subject can vindicate her substantive rights under the GDPR and human rights laws.[90] This is an example of logic-based explanations: a description of the reasoning behind a decision, not just the relevant inputs to the decision.

1. FCRA, ECOA, and Regulation B

The most straightforward legal requirement to explain inscrutable decision-making is the adverse action notice. In 1970, Congress passed FCRA[91] to begin to rein in the unregulated credit industry. FCRA was "the first information privacy legislation in the United States."[92] It limits to whom and for what purposes credit reports can be disclosed,[93] allows consumers access to their credit reports,[94] and requires credit reporting agencies (CRAs)—for example, Experian, Transunion, and Equifax—to employ procedures to ensure accuracy and govern dispute resolution.[95] FCRA was not initially concerned with how decisions were made, but rather with the then-new phenomenon of amassing large quantities of information.[96] Four years later, however, Congress passed ECOA[97] and took aim at the decision-

88. 15 U.S.C. §§ 1681m, 1691d(2).

89. GDPR, *supra* note 12, arts. 13(f)(2), 14(g)(2), 15(1)(h) (requiring access to "meaningful information about the logic" of automated decisions).

90. *See* Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233, 236 (2017). There is a vigorous debate in the literature about the "right to explanation" in the GDPR. *See infra* notes 143–45 and accompanying text. As a discussion of positive law, this debate is connected to, but different than, the point we seek to make about the GDPR—that it is one example of a law that operates by asking for the logic of a system. Even if there is held to be no "right to explanation" in the GDPR, one could imagine an equivalent law that encodes such a requirement.

91. Fair Credit Reporting Act, Pub. L. No. 91-508, 84 Stat. 1127 (1970) (codified as amended at 15 U.S.C. §§ 1681–1681x (2012)).

92. PRISCILLA M. REGAN, LEGISLATING PRIVACY: TECHNOLOGY, SOCIAL VALUES, AND PUBLIC POLICY 101 (1995).

93. 15 U.S.C. § 1681b.

94. *Id.* § 1681g.

95. *Id.* §§ 1681e(b), 1681i.

96. 115 CONG. REC. 2410 (1969).

97. Equal Credit Opportunity Act, Pub. L. No. 93-495, 88 Stat. 1521 (1974) (codified as amended at 15 U.S.C. §§ 1691–1691f (2012)).

making process.[98] ECOA prohibits discrimination in credit decisions on the basis of race, color, religion, national origin, sex, marital status, age (for adults), receipt of public assistance income, or exercise in good faith of the rights guaranteed under the Consumer Credit Protection Act.[99]

ECOA introduced the adverse action notice requirement.[100] When a creditor takes an adverse action against an applicant, the creditor must give a statement of "specific reasons" for the denial.[101] When FCRA later adopted a similar requirement, it expanded the notice to cover uses of credit information beyond decisions made by creditors, including the use of such information in employment decisions.[102]

ECOA's notice requirement was implemented by the Federal Reserve Board via Regulation B,[103] which mandates that the "statement of reasons . . . must be specific and indicate the principal reason(s) for the adverse action."[104] The regulation also notes that it is insufficient to "state[] that the adverse action was based on the creditor's internal standards or policies or that the applicant . . . failed to achieve a qualifying score on the creditor's credit scoring system."[105] An appendix to Regulation B offers a sample notification form designed to satisfy both the rule's and FCRA's notification requirements. Sample Form 1 offers twenty-four reason codes, including such varied explanations as "no credit file," "length of employment," or "income insufficient for amount of credit requested."[106] Though it is not

98. *Id.* § 502, 88 Stat. at 1521 (noting that the purpose of the legislation is to ensure credit is extended fairly, impartially, and without regard to certain protected classes).

99. 15 U.S.C. § 1691 (2012).

100. *Id.* § 1691(d)(2)(B); Winnie F. Taylor, *Meeting the Equal Credit Opportunity Act's Specificity Requirement: Judgmental and Statistical Scoring Systems*, 29 BUFF. L. REV. 73, 82 (1980) ("For the first time, federal legislation afforded rejected credit applicants an automatic right to discover why adverse action was taken.").

101. 15 U.S.C. § 1691(d)(2)–(3).

102. 15 U.S.C. § 1681m (2012).

103. Regulation B, 12 C.F.R. §§ 1002.1–.16 (2018).

104. 12 C.F.R. § 202.9(b)(2) (2018).

105. *Id.*

106. 12 C.F.R. pt. 1002, app. C (2018). The form's listed options are:

__Credit application incomplete
__Insufficient number of credit references provided
__Unacceptable type of credit references provided
__Unable to verify credit references
__Temporary or irregular employment
__Unable to verify employment
__Length of employment
__Income insufficient for amount of credit requested
__Excessive obligations in relation to income
__Unable to verify income
__Length of residence
__Temporary residence
__Unable to verify residence
__No credit file
__Limited credit experience
__Poor credit performance with us
__Delinquent past or present credit obligations with others
__Collection action or judgment

necessary to use the form, most creditors tend to report reasons contained on that form because there is a safe harbor for "proper use" of the form.[107]

Adverse action notices aim to serve three purposes: (1) to alert a consumer that an adverse action has occurred;[108] (2) to educate the consumer about how such a result could be changed in the future;[109] and (3) to prevent discrimination.[110] As the rest of this section will show, these are commonly cited reasons for relying on explanations as a means of regulation as a general matter. The first rationale, consumer awareness, is straightforward enough. It is a basic requirement of any information-regulation regime that consumers be aware of systems using their information.[111] But the relationship between adverse action notices and the other two rationales—consumer education and antidiscrimination—requires further exploration.

Adverse action notices can be helpful for consumer education. As Winnie Taylor pointed out shortly after the passage of ECOA, some reasons—"no credit file" and "unable to verify income"—are self-explanatory and would allow a consumer to take appropriate actions to adjust.[112] Conversely, some explanations, such as "length of employment" and home ownership, are harder to understand or act on.[113] This suggests that an explanation of a specific decision may be informative, but it may not reveal an obvious path to an alternative outcome.

There are also situations in which it may not even be informative. Taylor imagined a hypothetical additive credit-scoring system with eight different features—including whether an applicant owns or rents, whether he has a home phone, and what type of occupation he has, among other things—each assigned different point values.[114] In a system like that, someone who comes up one point short could find himself with every factor listed as a "principal

---

__Garnishment or attachment
__Foreclosure or repossession
__Bankruptcy
__Number of recent inquiries on credit bureau report
__Value or type of collateral not sufficient
__Other, specify: _____

Id.

107. Equal Credit Opportunity, 76 Fed. Reg. 41,590, 41,592 (July 15, 2011) ("A creditor receives a safe harbor for compliance with Regulation B for proper use of the model forms.").

108. *See* S. REP. NO. 94-589, at 4 (1976).

109. *Id.* ("[R]ejected credit applicants will now be able to learn where and how their credit status is deficient and this information should have a pervasive and valuable educational benefit. Instead of being told only that they do not meet a particular creditor's standards, consumers particularly should benefit from knowing, for example, that the reason for the denial is their short residence in the area, or their recent change of employment, or their already over-extended financial situation.").

110. *Id.* ("The requirement that creditors give reasons for adverse action is . . . a strong and necessary adjunct to the antidiscrimination purpose of the legislation, for only if creditors know they must explain their decisions will they effectively be discouraged from discriminatory practices.").

111. *See supra* note 32 and accompanying text.

112. Taylor, *supra* note 100, at 97.

113. *Id.* at 95.

114. *Id.* at 105–07.

reason"[115] for the denial. In one sense, this must be correct because a positive change in any factor at all would change the outcome. In another sense, however, choosing arbitrarily among equivalently valid reasons runs counter to the instruction to give specific and actionable notice.

Taylor also described a real system from that era, complex in all the various ways described in Part I—nonlinear, nonmonotonic, discontinuous, and multidimensional:

> [A]pplicants who have lived at their present address for less than six months are awarded 39 points, a level which they could not reach again until they had maintained the same residence for seven and one-half years. Furthermore, applicants who have been residents for between six months and 1 year 5 months (30 points) are considered more creditworthy than those who have been residents for between 1 and 1/2 years and 3 years 5 months (27 points).[116]

If the creditor tried to explain these rules simply, it would leave information out, but if the creditor were to explain in complete detail, it would likely overwhelm a credit applicant. This is an equivalent problem to simply disclosing how a model works under the banner of transparency; access to the model is not the same as understanding.[117]

The Federal Reserve Board recognized this problem, observing that, although all the principal reasons must be disclosed, "disclosure of more than four reasons is not likely to be helpful to the applicant."[118] The difficulty is that there will be situations where complexity cannot be avoided in a faithful representation of the scoring system, and listing factors alone will fail to accurately explain the decision, especially when the list is limited to four.[119] It is worth noting that modern credit systems appear not to be based on such complex models,[120] likely due to the very existence of FCRA and ECOA. Credit predictions tend to rely on features that bear an intuitive relationship to default, such as past payment history.[121] But the point is more general:

---

115. *See supra* note 104 and accompanying text.

116. Taylor, *supra* note 100, at 123.

117. *See* Ananny & Crawford, *supra* note 24, at 979 ("Transparency can intentionally occlude.").

118. 12 C.F.R. pt. 1002 supp. I, para. 9(b)(2) (2018). FCRA later codified the same limitation. 15 U.S.C. § 1681g(f)(1)(C) (2012).

119. The document also states that the "specific reasons . . . must relate to and accurately describe the factors actually considered or scored by a creditor . . . . A creditor need not describe how or why a factor adversely affected an applicant . . . . If a creditor bases the . . . adverse action on a credit scoring system, the reasons disclosed must relate only to those factors actually scored in the system." 12 C.F.R. pt. 1002 supp. I, para. 9(b)(2).

120. Patrick Hall, Wen Phan & SriSatish Ambati, *Ideas on Interpreting Machine Learning*, O'REILLY (Mar. 15, 2017), https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning [https://perma.cc/57XK-NU7G].

121. Carol A. Evans, *Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks*, CONSUMER COMPLIANCE OUTLOOK (Fed. Res. Sys., Phila., Pa.), 2017, at 4–5, https://consumercomplianceoutlook.org/assets/2017/second-issue/ccoi22017.pdf [https://perma.cc/52XP-PQN4]; *see also* ROBINSON + YU, KNOWING THE SCORE: NEW DATA, UNDERWRITING, AND MARKETING IN THE CONSUMER CREDIT MARKETPLACE 21 (2014), https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf [https://perma.cc/9FCY-4K2K].

approaches based on giving specific reasons for outcomes can fail where the system is too complex.

The adverse action notice fares worse as an antidiscrimination measure. By 1974, forcing hidden intentions into the open was a common technique for addressing discrimination.[122] Just one year before ECOA's passage, *McDonnell Douglas Corp. v. Green*[123] laid out the canonical Title VII burden-shifting framework for disparate treatment, which requires a defendant to rebut a prima facie case of employment discrimination with a nondiscriminatory reason and gives plaintiffs a chance to prove that the proffered reason is pretextual.[124] Just two years before that, the U.S. Supreme Court in *Griggs v. Duke Power Co.*[125] recognized disparate impact doctrine.[126] Disparate impact attributes liability for a facially neutral decision that has a disproportionate adverse effect on a protected class unless the decision maker can provide a legitimate business reason for the decision and no equally effective but less discriminatory alternative exists.[127] Its initial purpose was arguably to smoke out intentional discrimination where intent was hidden.[128] Thus, ECOA pursued the same goal—to prevent discrimination by forcing decision-making into the open.

While forcing stated reasons into the open captures the most egregious forms of intentional discrimination, it does not capture much else. Although, in some cases, Regulation B bars collection of protected-class information,[129] race, gender, and other features can be reliably inferred from sufficiently rich datasets.[130] Should creditors seek to discriminate intentionally by considering membership in a protected class, they would have to affirmatively lie about such behavior lest they reveal obvious wrongdoing. This form of intentional discrimination is thus addressed by disclosure. Should creditors rely on known proxies for membership in a protected class, however, while they would have to withhold the true relevance of these features in predicting creditworthiness, they could cite them honestly as reasons for the adverse action. The notice requirement therefore does not place meaningful constraints on creditors, nor does it create additional or

122. *See* Olatunde C. A. Johnson, *The Agency Roots of Disparate Impact*, 49 HARV. C.R.-C.L.L. REV. 125, 140 (2014) (tracing the history of agency use of disparate impact analysis to address latent discrimination).

123. 411 U.S. 792 (1973).

124. *Id.* at 805. The Supreme Court later found that a jury may presume that if all the employer had was pretext, that itself is evidence of discrimination. St. Mary's Honor Ctr. v. Hicks, 509 U.S. 502, 511 (1993) ("The factfinder's disbelief of the reasons put forward by the defendant (particularly if disbelief is accompanied by a suspicion of mendacity) may, together with the elements of the prima facie case, suffice to show intentional discrimination.").

125. 401 U.S. 424 (1971).

126. *Id.* at 431.

127. 42 U.S.C. § 2000e-2(k)(1)(A) (2012). This description ignores the word "refuse" in the statute, but is probably the more common reading. Barocas & Selbst, *supra* note 4, at 709.

128. Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 518–21 (2003) (discussing the "evidentiary dragnet" theory of disparate impact).

129. 12 C.F.R. § 1002.5 (2018).

130. Barocas & Selbst, *supra* note 4, at 692.

unique liability beyond that present in the antidiscrimination provisions of the rest of the regulation.[131]

More importantly, creditors using quantitative methods that do not expressly consider protected-class membership are likely not engaged in intentional discrimination, yet the scoring systems might very well evince a disparate impact. While ECOA does not expressly provide for a disparate impact theory of discrimination, case law suggests that it is very likely available.[132]

The adverse action notice approach has two specific shortcomings for a disparate impact case. First, when reviewing such a notice, the consumer only has access to her own specific outcome. Her single point of reference does not provide any understanding of the frequency of denials along protected-class lines, so she cannot observe disparate impact. Absent understanding of the logic of the system—for example, how different inputs are weighted—she cannot even look at the decision-making to try to guess whether it is discriminatory; the notice simply provides no basis to bring a suit.

Second, disparate impact has a different relationship to reasons behind decisions than does intentional discrimination. While for intentional discrimination, a consumer only needs to know that the decision was not made for an improper reason, knowing the specific reasons for which it *was* made becomes important for a disparate impact case.[133] That is to say, it is not only important to understand how a statistical system converts inputs to specific outputs, but also why the system was set up that way.

As discussed in Part I, one avenue to ensure the existence of an explanation of why the rules are the way they are is to require that the rules be based on intuitive relationships between input and output variables. This is the approach advocated by several scholars, particularly those focused on discrimination.[134] As is discussed in Part IV, it is not the only way, but this inability to engage with the normative purposes of the statute is a clear shortcoming of explanations based solely on the outcome of a single case, which provides neither the logic of the system nor any information about its normative elements.

131. John H. Matheson, *The Equal Credit Opportunity Act: A Functional Failure*, 21 HARV. J. ON LEGIS. 371, 388 (1984).

132. The Supreme Court has not ruled that it is available, but most circuit courts that have considered it have permitted it. *See* Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 193 (2016) (citing Golden v. City of Columbus, 404 F.3d 950, 963 (6th Cir. 2005)). In addition, the Supreme Court ruled in 2015 that disparate impact theory was cognizable in the Fair Housing Act, which also does not expressly provide for it. Texas Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2518 (2015).

133. Barocas & Selbst, *supra* note 4, at 702.

134. *See infra* Part III.A.3.

2. GDPR

In 2016, the European Union (EU) passed the GDPR, which took effect on May 25, 2018, and replaced the 1995 Data Protection Directive.[135] Both laws regulate automated decision-making,[136] but in the twenty-three years of the Directive's existence, little jurisprudence developed around that particular aspect of the law. The GDPR has created renewed interest in these provisions.[137]

The GDPR's discussion of automated decisions is contained in Articles 22, 13(2)(f), 14(2)(g), and 15(1)(h). Article 22 is the primary provision and states, in relevant part, the following:

> 1. The data subject shall have the 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.
> 2. Paragraph 1 shall not apply if the decision:
> (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
> (b) . . .
> (c) is based on the data subject's explicit consent.
> 3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.[138]

Articles 13–15 spell out a data subject's right to be informed about the information that data controllers have about her.[139] Articles 13 and 14 describe the obligations of data controllers to affirmatively notify data subjects about the uses of their information,[140] and Article 15 delineates the access rights that data subjects have to information about how their own data is used.[141] All three demand that the following information be available to data subjects: "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."[142]

---

135. GDPR, *supra* note 12, art. 99.

136. *Id.* art. 22(1) ("The data subject shall have the 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."); Data Protection Directive, *supra* note 84, art. 15.

137. Isak Mendoza & Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, *in* EU INTERNET LAW 77, 80–81 (2017).

138. GDPR, *supra* note 12, art. 22. Article 22(4) is omitted because it is not relevant to this discussion.

139. Wachter et al., *supra* note 23, at 89.

140. *See* GDPR, *supra* note 12, arts. 13–14.

141. *See id.* art. 15.

142. *Id.* arts. 13(2)(f), 14(2)(g), 15(1)(h).

Since passage of the GDPR, scholars have debated whether these requirements amount to a "right to explanation."[143] As one of us has argued elsewhere, that debate has been bogged down in proxy battles over what the phrase "right to explanation" means, but no matter whether one calls it a right to explanation, requiring that data subjects have meaningful information about the logic must mean something related to explanation.[144] Importantly for this discussion, the Regulation demands that the "meaningful information" must be about the *logic* of the decisions.[145] As we defined it in Part I, a model is inscrutable when it defies practical inspection and resists comprehension. An explanation of the logic therefore appears to precisely target inscrutability. The most important aspect of this type of explanation is that it is concerned with the operation of the model in general, rather than as it pertains to a particular outcome.

The particular type of explanation required by the GDPR will depend on the legal standards developed in the EU by the authorities charged with interpreting that law. The overall purposes of the GDPR are much broader than FCRA and ECOA. The EU treats data protection as a fundamental right,[146] and the GDPR seeks to vindicate the following principles with respect to personal data: lawfulness, fairness, and transparency; purpose limitation; data minimization; accuracy; storage limitation; integrity and confidentiality; and accountability.[147] Several of these principles are a restatement of the FIPs that have shaped privacy policy for decades.[148]

143. *See* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. (forthcoming 2019) (manuscript at 17–24), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3196985 [https://perma.cc/92GH-W6HV] (reviewing the literature); *see also* 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, 44 (2017) (arguing that even if a right to explanation exists, it may not be useful); Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 243, 245, 250 (2017) (arguing that the GDPR creates a right to "legibility" that combines transparency and comprehensibility); Mendoza & Bygrave, *supra* note 137 (arguing that a right to explanation can be derived as a necessary precursor to the right to contest the decision); Selbst & Powles, *supra* note 90 (arguing that a right to meaningful information is a right to explanation); Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH 841 (2018) (arguing that a legal right to explanations of automated decisions does not exist); Wachter et al., *supra* note 23 (arguing that there is no legal right to explanation of specific automated decisions); Goodman & Flaxman, *supra* note 20, at 2 (arguing that a right to explanation exists); Maja Brkan, Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond 15 (Aug. 1, 2017) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3124901 [https://perma.cc/C9PN-4PL6] (arguing that "information about the logic involved" and the right to contest decisions imply a right to explanation).

144. *See* Selbst & Powles, *supra* note 90, at 233.

145. GDPR, *supra* note 12, arts. 13(2)(f), 14(2)(g), 15(1)(h).

146. *Id.* art. 1.

147. *Id.* art. 5.

148. Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 106–07 (2014). While different lists of FIPs conflict, one prominent example is the Organisation for Economic Co-Operation and Development's (OECD) list: Collection Limitation Principle, Data Quality Principle,

Considered as a whole, they begin to sound like the general idea of due process in all its expansiveness.

Satisfying this requirement may in some cases involve disclosing the full set of rules behind all decision-making—that is, the entire model.[149] But in some cases, it will not involve such radical disclosure. Depending on the specific goals at issue, the types of rules disclosed can be narrower, or the explanation can perhaps be met interactively by providing data subjects with the tools to examine how changes in their information relate to changes in outcome. One of us has argued that the GDPR's meaningful information requirement applies "to the data subject herself"[150] and "should be interpreted functionally and flexibly," and that the legal standard should be that the explanation "at a minimum, enable[s] a data subject to exercise his or her rights under the GDPR and human rights law."[151]

Although the GDPR's goals are broader than those of ECOA and FCRA, evaluating the ability of logic-based explanations to vindicate the goals of those statutes can demonstrate how explanations of the logic of decision-making can improve upon the shortcomings of the outcome-based approach. The three reasons were awareness, consumer (here, data subject) education, and antidiscrimination.[152] Like in the credit domain, awareness is straightforward and encapsulated by the requirement that a data subject be made aware of the "existence" of automated decision-making. The other two rationales operate differently when logic-based explanations are provided.

Data subject education becomes more straightforward as a legal matter, if not a technical one. Absent inscrutability, a data subject would be told the rules of the model and would be able to comprehend his situation and how to achieve any particular outcome. This solves both problems that Taylor identified.[153] Consider the system where, after the creditor totaled the point values from eight factors, a person missed on her credit application by one point. While it might be impossible to point to four factors that were "principal reasons," the explanation of the logic—what the eight factors were, that they were all assigned point values, and that the hypothetical applicant just missed by a point—would be much more useful to that

---

Purpose Specification Principle, Use Limitation Principle, Security Safeguards Principle, Openness Principle, Individual Participation Principle, and Accountability Principle. Org. for Econ. Co-operation & Dev. [OECD], *The OECD Privacy Framework*, at 14–15 (2013), http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf [https://perma.cc/RWM2-EUD4].

149. The guidelines issued by the Article 29 Working Party, a body tasked with giving official interpretations of EU law, states that the full model is not required. *See* Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, at 25, WP 251 (Feb. 6, 2018) ("The GDPR requires the controller to provide meaningful information about the logic involved, not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm."). As a matter of positive law, then, this is likely to be the outcome, but in some cases it may fall short of something actually meaningful to the data subject.

150. *See* Selbst & Powles, *supra* note 90, at 236.

151. *Id.* at 233.

152. *See supra* notes 108–10 and accompanying text.

153. *See supra* notes 112–16 and accompanying text.

particular rejected applicant.[154] In Taylor's real nonlinear, nonmonotonic, discontinuous, and multidimensional example, the full complexity can be appreciated in the paragraph-long description, where a reason code would in many cases be totally unhelpful. Once machine learning enters the picture, and models become more complex, the limits on technical ability to solve inscrutability may prevent these explanations from coming to fruition. But at least in theory, explanations of the logic are sufficient for data subject education.

Turning to discrimination—which serves as a stand-in for broader normative questions about model justification—while logic-based explanations do fare better than outcome-based ones, they do not completely address the shortcomings. Any rule that is manifestly objectionable becomes visible under logic-based explanations, making them an improvement over outcome-only explanations, which shed no light on rules. This disclosure might enable one to speculate if facially neutral rules will nevertheless have a disparate impact, based on the different rates at which certain input features are held across the population. But this is ultimately little more than guesswork.[155] Although there might not be anything about a rule that appears likely to generate a disparate impact, it still could. Alternatively, a set of rules could appear objectionable or discriminatory, but ultimately be justified. It will often be impossible to tell without more information, and the possibility of happening on a set of rules that lend themselves to intuitive normative assessment is only a matter of chance.

B. Interpretability in Machine Learning

The overriding question that has prompted fierce debates about explanation and machine learning has been whether machine learning can be made to comply with the law. As discussed in Part I, machine learning poses unique challenges for explanation and understanding—and thus challenges for meeting the apparent requirements of the law. Part II.A further demonstrated that even meeting the requirements of the law does not automatically provide the types of explanations that would be necessary to assess whether decisions are well justified. Nevertheless, addressing the potential inscrutability of machine learning models remains a fundamental step in meeting this goal.

As it happens, machine learning has a well-developed toolkit to deal with calls for explanation. There is an extensive literature on "interpretability."[156] Early research recognized and grappled with the challenge of explaining the decisions of machine learning models such that people using these systems

154. The Article 29 Working Party has, however, suggested that this approach is central to the "meaningful information" requirement. *See* Article 29 Data Protection Working Party, *supra* note 149, at 25.

155. *See infra* Part III.A.3.

156. *See generally, e.g.*, Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUTING SURVEYS, Aug. 2018, at 1; Lipton, *supra* note 66.

would feel comfortable acting upon them.[157] Practitioners and researchers have developed a wide variety of strategies and techniques to ensure that they can produce interpretable models from data—many of which may be useful for complying with existing law, such as FCRA, ECOA, and the GDPR.

Interpretability has received considerable attention in research and practice due to the widely held belief that there is a tension between how well a model will perform and how well humans will be able to interpret it.[158] This view reflects the reasonable idea that models that consider a larger number of variables, a larger number of relationships between these variables, and a more diverse set of potential relationships is likely to be *both* more accurate and more complex.[159] This will certainly be the case when the phenomenon that machine learning seeks to model is itself complex. This intuition suggests that practitioners may face a difficult choice: favor simplicity for the sake of interpretability or accept complexity to maximize performance.[160]

While such views seem to be widely held,[161] over the past decade, methods have emerged that attempt to sidestep these difficult choices altogether, promising to increase interpretability while retaining performance.[162] Researchers have developed at least three different ways to respond to the demand for explanations: (1) purposefully orchestrating the machine learning process such that the resulting model is interpretable;[163] (2) applying special techniques after model creation to approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions;[164] and (3) providing tools that allow people to interact with the model and get a sense of its operation.[165]

1. Purposefully Building Interpretable Models

Practitioners have a number of different levers at their disposal to purposefully design simpler models. First, they may choose to consider only a limited set of all possible variables.[166] By limiting the analysis to a smaller set of variables, the total number of relationships uncovered in the learning process might be sufficiently limited to be intelligible to a human.[167] It is

157. van Melle et al., *supra* note 85, at 302.

158. *See, e.g.*, Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 STAT. SCI. 199, 206 (2001); Lou et al., *supra* note 54, at 150.

159. *See* Breiman, *supra* note 158, at 208.

160. *See generally id.*

161. *See* DEF. ADVANCED RESEARCH PROJECTS AGENCY, BROAD AGENCY ANNOUNCEMENT: EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) (2016), https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf [https://perma.cc/3FZV-TZGA]; Henrik Brink & Joshua Bloom, *Overcoming the Barriers to Production-Ready Machine-Learning Workflows*, STRATA (Feb. 11, 2014), https://conferences.oreilly.com/strata/strata2014/public/schedule/detail/32314 [https://perma.cc/2GBV-2QRR].

162. For a recent survey, see Michael Gleicher, *A Framework for Considering Comprehensibility in Modeling*, 4 BIG DATA 75 (2016).

163. *See, e.g.*, *id.* at 81–82.

164. *See, e.g.*, *id.* at 82–83.

165. *See, e.g.*, *id.* at 83.

166. *See id.* at 81.

167. Zeng et al., *supra* note 82, at 690–91.

very likely that a model with five features, for example, will be more interpretable than a model with five hundred.

Second, practitioners might elect to use a learning method that outputs a model that can be more easily parsed than the output of other learning methods.[168] For example, decision tree algorithms are perceived as likely to produce interpretable models because they learn nested rules that can be represented visually as a tree with subdividing branches. To understand how the model would process any particular case, practitioners need only walk through the relevant branches of the tree; to understand the model overall, practitioners can explore all the branches to develop a sense of how the model would determine all possible cases.

The experience of applying machine learning to real-world problems has led to common beliefs among practitioners about the relative interpretability of models that result from different learning methods and how well they perform. Conventional wisdom suggests that there is a trade-off between interpretability and accuracy.[169] Methods like linear regression[170] generate models perceived as highly interpretable, but relatively low performing, while methods like deep learning[171] result in high-performing models that are exceedingly difficult to interpret.[172] While researchers have pointed out that such comparisons do not rest on a rigorous definition of interpretability or empirical studies,[173] such beliefs routinely guide practitioners' decisions when applying machine learning to different kinds of problems.[174]

Another method is to set the parameters of the learning process to ensure that the resulting model is not so complex that it defies human comprehension. For example, even decision trees will become unwieldy for humans if they involve an exceedingly large number of branches and leaves.[175] Practitioners routinely set an upper bound on the number of leaves to constrain the complexity of the model.[176] For decades, practitioners in regulated industries like credit and insurance have purposefully limited themselves to a relatively small set of features and less sophisticated learning methods.[177] In so doing, they have been able to generate models that lend themselves to sensible explanation, but they may have forgone the increased accuracy that would result from a richer and more advanced analysis.[178]

---

168. *See* Lehr & Ohm, *supra* note 51, at 688–95.
169. *See, e.g.*, Breiman, *supra* note 158, at 208.
170. *See Regression*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).
171. *See generally* Jürgen Schmidhuber, *Deep Learning in Neural Networks: An Overview*, 61 NEURAL NETWORKS 85 (2015) (providing an explanation of deep learning in artificial intelligence).
172. Breiman, *supra* note 158, at 206.
173. Alex A. Freitas, *Comprehensible Classification Models—a Position Paper*, 15 SIGKDD EXPLORATIONS, June 2013, at 1.
174. *See* Lipton, *supra* note 66, at 99.
175. *Id.* at 98.
176. *See id.* at 99.
177. Hall et al., *supra* note 120.
178. *Id.*

Linear models remain common in industry because they allow companies to much more readily comply with the law.[179] When they involve a sufficiently small set of features, linear models are concise enough for a human to grasp the relevant statistical relationships and to simulate different scenarios.[180] They are simple enough that a full description of the model may amount to the kind of meaningful information about the logic of automated decisions required by the GDPR. At the same time, linear models also immediately highlight the relative importance of different features by assigning a specific numerical weight to each feature, which allows companies to quickly extract the principal factors for an adverse action notice under ECOA.

Beyond the choice of features, learning method, or learning parameters, there are techniques that can make simplicity an additional and explicit optimization criterion in the learning process. The most common such method is regularization.[181] Much like setting an upper limit on the number of branches in a decision tree, regularization allows the learning process to factor in model complexity by assigning a cost to excess complexity.[182] In doing so, model simplicity becomes an additional objective alongside model performance, and the learning process can be set up to find the optimal trade-off between these sometimes-competing objectives.[183]

Finally, the learning process can also be constrained such that all features exhibit monotonicity.[184] Monotonicity constraints are widespread in credit scoring because they make it easier to reason about how scores will change when the value of specific variables change, thereby allowing creditors to automate the process of generating the reason codes required by FCRA and ECOA.[185] As a result of these legal requirements, creditors and other data-

179. *Id.*

180. *See* Lipton, *supra* note 66, at 98.

181. *See* Gleicher, *supra* note 162, at 81–82.

182. *See id.* at 81. One commonly used version of this method is Lasso. *See generally* Robert Tibshirani, *Regression Shrinkage and Selection via the Lasso*, 58 J. ROYAL STAT. SOC'Y 267 (1996). It was originally designed to increase accuracy by avoiding overfitting, which occurs when a model assigns significance to too many features and thus accidentally learns patterns that are peculiar to the training data and not representative of real-world patterns. *See id.* at 267. Machine learning is only effective in practice when it successfully identifies robust patterns while also ignoring patterns that are specific to the training data. *See* David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STAT. SCI. 1, 2 (2006). Lasso increases accuracy by forcing the learning process to ignore relationships that are relatively weak, and therefore more likely to be artifacts of the training data. *See* Tibshirani, *supra*, at 268. Because Lasso works by strategically removing unnecessary features, the technique can simultaneously improve interpretability (by reducing complexity) in many real-world applications and increase performance (by avoiding overfitting). *See id.* at 267. As such, improved interpretability need not always decrease performance. But where potential overfitting is not a danger, regularization methods may result in degradations in performance. *See* Gleicher, *supra* note 162, at 81–82.

183. Gleicher, *supra* note 162, at 81.

184. Recall that monotonicity implies that an increase in an input variable can only result in either an increase or decrease in the output; it can never change from one to the other. *See supra* notes 57–58 and accompanying text.

185. *See, e.g.*, Hall et al., *supra* note 120. Monotonicity allows creditors to rank order variables according to how much the value of each variable in an applicant's file differs from

driven decision makers often have incentives to ensure their models are interpretable by design.

2. Post Hoc Methods

There exists an entirely different set of techniques for improved interpretability that does not place any constraints on the model-building process. Instead, these techniques begin with models learned with more complex methods and attempt to approximate them with simpler and more readily interpretable methods. Most methods in this camp generate what can be understood as a model of the model.

These methods attempt to overcome the fact that simpler learning methods cannot always reliably discover as many useful relationships in the data. For example, the learning process involved in decision trees is what is known as a "greedy algorithm."[186] Once the learning process introduces a particular branch, the method does not permit walking back up the branch.[187] Therefore, relationships between items on two different branches will not be discovered.[188] Despite lacking the same limitation, more complex learning methods, such as deep learning, do not result in models as interpretable as decision trees. Nonetheless, rules that cannot be *learned* with simpler methods can often be *represented* effectively by simpler models.[189] Techniques like rule extraction[190] allow simple models to "cheat" because the answers that simpler learning methods would otherwise miss are known ahead of time.[191]

This approach can be costly and it does not have universal success.[192] Despite practitioners' best efforts, replicating the performance of more complex models in a simple enough form might not be possible where the phenomena are particularly complex. For example, using a decision tree to approximate a model developed with deep learning might require too large a number of branches and leaves to be understandable in practice.[193]

When these methods work well, they ensure that the entire set of relationships learned by the model can be expressed concisely, without

---

the corresponding value of each variable for the ideal customer—the top four variables can function as reason codes. *Id.*

186. STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 92–93 (3d ed. 2014).

187. *Id.*

188. *Id.* at 93 (noting that, although the greedy algorithm may find a nonoptimal solution, it will not discover relationships between unrelated branches).

189. Gleicher, *supra* note 162, at 82.

190. Rule extraction is the name for a set of techniques used to create a simplified model of a model. The technical details of their operation are beyond the scope of this paper. *See generally* Nahla Barakat & Andrew P. Bradley, *Rule Extraction from Support Vector Machines: A Review*, 74 NEUROCOMPUTING 178 (2010); David Martens et al., *Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines*, 183 EUR. J. OPERATIONAL RES. 1466 (2007).

191. Gleicher, *supra* note 162, at 82.

192. *Id.*

193. *See* Lipton, *supra* note 66, at 98.

giving up much performance. Accordingly, they serve a similar role to the interpretability-driven design constraints discussed above.[194] When they do not work as well, arriving at an interpretable model might necessitate sacrificing some of the performance gained by using the more complex model. But even when these methods involve a notable loss in performance, the resulting models frequently perform far better than simple methods alone.[195]

Other tools have also emerged that attack the problem of interpretability from a different direction. Rather than attempting to ensure that machine learning generates an intelligible model overall, these new tools furnish more limited explanations that only account for the relative importance of different features in particular outcomes—similar to the reason codes required by FCRA and ECOA.[196] At a high level, most of these methods adopt a similar approach: they attempt to establish the importance of any feature to a particular decision by iteratively varying the value of that feature while holding the value of other features constant.[197]

These tools seem well suited for the task set by ECOA, FCRA, or other possible outcome-oriented approaches: explaining the principal reasons that account for the specific adverse decision.[198] As we further discuss in the next section, there are several reasonable ways to explain the same specific outcome. These methods are useful for two of the most common: (1) determining the relative contribution of different features, or (2) identifying the features whose values would have to change the most to change the outcome.[199] One could imagine applying these methods to models that consider an enormous range of features and map out an exceedingly complex set of relationships. While such methods will never make these relationships completely sensible to a human, they can provide a list of reasons that might help provide reason codes for a specific decision.

---

194. *See supra* Part II.B.1.

195. Johan Huysmans et al., *Using Rule Extraction to Improve the Comprehensibility of Predictive Models* (Katholieke Universiteit Leuven Dep't of Decision Scis. & Info. Mgmt., Working Paper No. 0612, 2006), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=961358 [https://perma.cc/8AKQ-LXVE].

196. *See supra* note 106 and accompanying text.

197. *See generally* Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE & INFO. SYSTEMS 95 (2018); David Baehrens et al., *How to Explain Individual Classification Decisions*, 11 J. MACHINE LEARNING RES. 1803 (2010); Anupam Datta et al., *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, *in* PROCEEDINGS OF THE 2016 IEEE SYMPOSIUM ON SECURITY & PRIVACY 598 (2016); Andreas Henelius et al., *A Peek into the Black Box: Exploring Classifiers by Randomization*, 28 DATA MINING & KNOWLEDGE DISCOVERY 1503 (2014); Marco Tulio Ribeiro et al., *"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 (2016).

198. *See supra* note 88 and accompanying text.

199. These methods are generally sensitive to interactions among variables and can measure indirect as well as direct influence. *See, e.g.*, Adler et al., *supra* note 197; Datta et al., *supra* note 197; Julius Adebayo, *FairML: Auditing Black-Box Predictive Models*, CLOUDERA FAST FORWARD LABS (Mar. 9, 2017), http://blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html [https://perma.cc/S5PK-K6GQ].

Unfortunately, these methods may not work well in cases where models take a much larger set of features into account. Should many features each contribute a small amount to a particular determination, listing each feature in an explanation is not likely to be helpful. This is the machine learning version of Taylor's hypothetical eight-factor credit example.[200] The number of features identified as influential might be sufficiently large that the explanation would simply reproduce the problem of inscrutability that it aims to address. The only alternative in these cases—arbitrarily listing fewer reasons than the correct number—is also unsatisfying when all features are equivalently, or nearly equivalently, important. As it happens, post hoc explanations for credit and other similarly important decisions are likely to be most attractive precisely when they do not seem to work well—that is, when the only way to achieve a certain level of performance is to vastly expand the range of features under consideration.

These methods are also unlikely to generate explanations that satisfy logic-like approaches like the GDPR. Indeed, such techniques pose a unique danger of misleading people into believing that the reasons that account for specific decisions must also apply in the same way for others—that the reasons for a specific decision illustrate a general rule. Understandably, humans tend to extrapolate from explanations of specific decisions to similar cases, but the model—especially a complex one—may have a very different basis for identifying similar-seeming cases.[201] These methods offer explanations that apply only to the case at hand and cannot be extrapolated to decisions based on other input data.[202]

3. Interactive Approaches

One final set of approaches is interactive rather than explanatory. Practitioners can allow people to get a feel for their models by producing interactive interfaces that resemble the methods described in the previous sections. This can take two quite different forms. One is the type proposed by Danielle Citron and Frank Pasquale[203] and implemented, for example, by Credit Karma.[204] Beginning with a person's baseline credit information, Credit Karma offers a menu of potential changes, such as opening new credit cards, obtaining a new loan, or going into foreclosure.[205] A person using the interface can see how each action would affect his credit score.[206] This does

200. *See supra* notes 114–15 and accompanying text.

201. *See* Finale Doshi-Velez & Mason Kortz, *Accountability of AI Under the Law: The Role of Explanation* 3 (Harvard Univ. Berkman Klein Ctr. Working Grp. on Explanation & the Law, Working Paper No. 18-07, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064761 [https://perma.cc/SJ5S-HJ3T] (discussing the problem of cases where similar situations lead to differing outcomes and vice versa).

202. *See id.*

203. *See* Citron & Pasquale, *supra* note 7, at 28–30 (discussing "interactive modeling").

204. *See Credit Score Simulator*, CREDIT KARMA, https://www.creditkarma.com/tools/credit-score-simulator [https://perma.cc/XQ2S-GYUE] (last visited Nov. 15, 2018).

205. *Id.*

206. *Id.*

not amount to a full explanation because a person at a different starting point could make similar moves with different outcomes, but it gives the individual user a partial functional feel for the logic of the system as it applies to him specifically.

The second is more complicated and abstract. Mireille Hildebrandt has proposed something she terms "transparency-enhancing technologies."[207] Such technologies would implement an interface that would allow people to simultaneously adjust the value of multiple features in a model with the goal of providing a loose sense of the relationship between these features and a specific outcome, as well as the connection between the features themselves.[208] The goal of this type of technology is not to tell the user what changes in his results specifically but to allow him to get a feel from an arbitrary starting point.[209]

Where models are simple enough, these approaches seem to achieve the educational goals of both ECOA and the GDPR by allowing data subjects to gain an intuitive feel for the system. Ironically, this would be accomplished by complying with neither law because a person will not know a specific reason for denial or have an account of a model's logic after playing with it, even if they feel that they understand the model better afterward.

While regulators have expressed interest in this idea,[210] however, it poses a technical challenge. The statistical relationships at work in these models may be sufficiently complex that no consistent rule may become evident by tinkering with adjustable sliders. Models might involve a very large number of inputs with complex and shifting interdependencies such that even the most systematic tinkering would generate outcomes that would be difficult for a person to explain in a principled way.

One danger of this approach, then, is that it could do more to placate than elucidate. People could try to make sense of variations in the observed outputs by favoring the simplest possible explanation that accounts for the limited set of examples generated by playing with the system. Such an explanation is likely to take the form of a rule that incorrectly assigns a small set of specific variables unique significance and treats their effect on the outcome as linear, monotonic, and independent. Thus, for already simple models that *can* be explained, interactive approaches may be useful for giving people a feel without disclosing the algorithm, but for truly inscrutable systems, they could well be dangerous.

---

207. Mireille Hildebrandt, *Profiling: From Data to Knowledge*, 30 DATENSCHUTZ UND DATENSICHERHEIT 548, 552 (2006); *see also* Mireille Hildebrandt & Bert-Jaap Koops, *The Challenges of Ambient Law and Legal Protection in the Profiling Era*, 73 MODERN L. REV. 428, 449 (2010). *See generally* NICHOLAS DIAKOPOULOS, ALGORITHMIC ACCOUNTABILITY REPORTING: ON THE INVESTIGATION OF BLACK BOXES (2013), http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf [https://perma.cc/H9UU-WK6V].

208. *See* Hildebrandt & Koops, *supra* note 207, at 450.

209. *See id.*

210. *See* INFO. COMM'R'S OFFICE, BIG DATA, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA PROTECTION 87–88 (2017), https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf [https://perma.cc/J97E-N5NV].

* * *

Remarkably, the techniques available within machine learning for ensuring interpretability correspond well to the different types of explanation required by existing law. There are, on the one hand, varied strategies and techniques available to practitioners that can deliver models whose inner workings can be expressed succinctly and sensibly to a human observer, whether an expert (e.g., a regulator) or lay person (e.g., an affected consumer). Laws like the GDPR that seek logic-like explanations would be well served by these methods. On the other hand, outcome-focused laws like ECOA that care only about principal reasons—and not the set of rules that govern all decisions—have an obvious partner in tools that furnish post hoc accounts of the factors that influenced any particular determination.

Where they succeed, these methods can be used to meet the demands of regulatory regimes that demand outcome- and logic-like explanations. Both techniques have their limitations, however. If highly sophisticated machine learning tools continue to be used, interpretability may be difficult to achieve in some instances, especially when the phenomena at issue are themselves complex. Post hoc accounts that list the factors most relevant to a specific decision may not work well when the number of relevant factors grows beyond a handful—a situation that is most likely to occur when such methods would be most attractive.

Notably, neither the techniques nor the laws go beyond describing the operation of the model. Though they may help to explain why a decision was reached or how decisions are made, they cannot address why decisions happen to be made that way. As a result, standard approaches to explanation might not help determine whether the particular way of making decisions is normatively justified.

III. FROM EXPLANATION TO INTUITION

So far, the majority of discourse around understanding machine learning models has seen the proper task as opening the black box and explaining what is inside.[211] Where Part II.A discussed legal requirements and Part II.B discussed technical approaches, here we discuss the motivations for both. Based on a review of the literature, scholars, technologists, and policymakers seem to have three different beliefs about the value of opening the black box.[212] The first is a fundamental question of autonomy, dignity, and

211. *See supra* note 16 and accompanying text.

212. These three rationales seem to track the rationales for ECOA's adverse action notices as described in Part II.A.1. There is also scholarship that offers a fourth rationale, which includes due process and rule-of-law concerns. We set these concerns aside because they pertain to government use of algorithms, while this Article focuses on regulation of the private sector. *See* Brennan-Marquez, *supra* note 19, at 1288–94 (discussing "rule-of-law" principles with respect to police and judicial actions); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1184–90, 1206–09 (2017) (discussing due process and reason-giving in administrative law); ECLT Seminars, *[HUML16] 03: Katherine Strandburg, Decision-Making, Machine Learning and the Value of Explanation*, YOUTUBE (Jan. 23, 2017), https://www.youtube.com/

personhood. The second is a more instrumental value: educating the subjects of automated decisions about how to achieve different results. The third is a more normative question—the idea that explaining the model will allow people to debate whether the model's rules are justifiable.

The black-box-only approach is limited for the purposes of justifying decision-making. The first two beliefs are not about justifying decisions at all, and therefore serve a different purpose. The third is explicitly about justification, so our critique is directed not at its intent, but its operation. For those concerned with the justification for decision-making, the goal of explanation should be to find a way to bring intuition to bear in deciding whether the model is well justified. This Part explains both the power and limitations of such an approach.

A. The Value of Opening the Black Box

This Part identifies and elaborates the three rationales that apparently underlie most of the popular and scholarly calls for explanation.

1. Explanation as Inherent Good

There are several reasons to view explanation as a good unto itself, and perhaps a necessary part of a system constrained by law, including a respect for autonomy, dignity, and personhood.[213] There is a fundamental difference between wanting an explanation for its own sake and wanting an explanation for the purpose of vindicating certain specific empowerment or accountability goals. Fears about a system that lacks explanation are visceral. This fear is best exemplified in popular consciousness by Franz Kafka's *The Trial*,[214] a story about a faceless bureaucracy that makes consequential decisions without input or understanding from those affected.[215]

This concern certainly motivates some lawmakers and scholars. In his article, "Privacy and Power," Daniel Solove refers to this as a "dehumanizing" state of affairs characterized by the "powerlessness and vulnerability created by people's lack of any meaningful form of participation" in the decision.[216] David Luban, Alan Strudler, and David Wasserman argue that "one central aspect of the common good"—which they argue forms the basis of law's legitimacy—"lies in what we might call the *moral intelligibility* of our lives" and that the "horror of the bureaucratic process lies not in officials' mechanical adherence to duty, but rather in the

watch?v=LQj3nbfSkrU [https://perma.cc/CX7S-GCUG] (discussing procedural due process and explanations).

213. *See* Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231, 1238–39 (1992) (explaining that while "person" usually means human being in the law, "personhood" is a question of the attendant "bundle of rights and duties").

214. FRANZ KAFKA, DER PROCESS (1925).

215. *See* Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1397–98 (2001) (arguing that Kafka's *The Trial* is a better metaphor than George Orwell's *1984* for modern anxieties over data).

216. *Id.* at 1423.

individual's ignorance of what the fulfillment of his or her duty may entail."[217] The concerns of dignity and personhood certainly motivate the data protection regime in Europe,[218] if less directly the law in the United States.[219]

We lack the space (and the expertise) to do proper justice to the personhood argument for explanation. Accordingly, our goal here is to flag it and set it aside as a concern parallel to our broader concerns about enabling justifications for automated decisions.

To the extent that the personhood rationale can be converted to a more actionable legal issue, it is reflected in the concept of "procedural justice," which was most famously championed by Tom Tyler. Procedural justice is the essential quality of a legal system that shows respect for its participants, which might entail transparency, consistency, or even politeness.[220] Tyler and others have shown that people care deeply about procedural justice, to the point that they might find a proceeding more tolerable and fair if their procedural-justice concerns are satisfied even if they do not obtain their preferred outcome in the proceeding.[221] Procedural justice, Tyler argues, is necessary on a large scale because it allows people to buy into the legal system and voluntarily comply with the law, both of which are essential parts of a working and legitimate legal system.[222] Presumably, to the extent that automated decisions can be legally or morally justified, people must accept them rather than have them imposed, and as a result, the personhood rationale for model explanation also implicates procedural justice.

Ultimately, that there is inherent value in explanation is clear. But as a practical matter, those concerns are difficult to administer, quantify, and compare to other concerns. Where there are genuine trade-offs between explanation and other normative values such as accuracy or fairness, the inherent value of explanation neither automatically trumps competing considerations nor provides much guidance as to the type of explanation required. Therefore, while inherent value cannot be ignored, other rationales remain important.

---

217. David Luban, Alan Strudler & David Wasserman, *Moral Responsibility in the Age of Bureaucracy*, 90 MICH. L. REV. 2348, 2354 (1992).

218. Lee A. Bygrave, *Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling*, 17 COMPUTER L. & SECURITY REP. 17, 19 (2001); Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216, 223–24 (2017).

219. *See* James Q. Whitman, *The Two Western Cultures of Privacy: Dignity Versus Liberty*, 113 YALE L.J. 1151, 1214–15 (2004).

220. Tom R. Tyler, *What Is Procedural Justice?: Criteria Used by Citizens to Assess the Fairness of Procedures*, 22 LAW & SOC'Y REV. 103, 132 (1988).

221. *See, e.g.*, Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME & JUST. 283, 291 (2003); Tyler, *supra* note 220, at 128.

222. TOM R. TYLER, WHY PEOPLE OBEY THE LAW 6–7 (2006).

2. Explanation as Enabling Action

For others, the purpose of explanation extends to providing actionable information about the rendering of decisions, such that affected parties can learn if and how they might achieve a different outcome. Explanations are valuable, on this account, because they empower people to effectively navigate the decision-making process. Such beliefs are evident in the adverse action notice requirements of credit-scoring regulations,[223] but they have come to dominate more recent debates about the regulatory function of requiring explanations of model-driven decisions more generally.

Across a series of recent papers, the debate has coalesced around two distinct, but related, questions. The first is whether and when the GDPR requires explanations of the logic or outcome of decision-making. The second is how to best explain outcomes in an actionable way.

The first question, whether to focus on outcome- or logic-based explanations, originates with an article by Sandra Wachter, Brent Mittelstadt, and Luciano Floridi.[224] These scholars split explanations between "system functionality" and "specific decisions"—a distinction functionally similar to our outcome- and logic-based framework.[225] This mirrors the debate in the technical community about the best way to understand the meaning of interpretability. As described in Part II.B, the main split is whether to aim for interpretable models or to account for specific decisions. Drawing together the legal and machine learning literature, Lilian Edwards and Michael Veale have created a similar, but slightly altered distinction between "model-centric" and "subject-centric" explanations.[226] While not identical, subject-centric explanations are another way to explain specific outcomes to individuals.[227]

As the discussion has evolved in both the legal and computer science scholarship, new work has converged on the belief that explaining specific outcomes is the right approach. The debate has therefore shifted to the

223. *See supra* Part II.A.1.

224. Wachter et al., *supra* note 23.

225. *Id.* at 78. As Wachter and colleagues define it, system functionality is "the logic, significance, envisaged consequences, and general functionality of an automated decision-making system," and explanations of specific decisions are "the rationale, reasons, and individual circumstances of a specific automated decision." *Id.* While the distinction is broadly useful, our definitions differ from theirs and we believe the line between outcome- and logic-based explanations is less clear than they suggest. *See* Selbst & Powles, *supra* note 90, at 239 (arguing that, given the input data, a description of the logic will provide a data subject with the means to determine any particular outcome, and thus, explanations of the logic will often also explain individual outcomes).

226. Edwards & Veale, *supra* note 143, at 55–56. They define these terms as follows: "Model-centric explanations (MCEs) provide broad information about a [machine learning] model which is not decision or input-data specific," while "[s]ubject-centric explanations (SCEs) are built on and around the basis of an input record."

227. Ultimately, Edwards and Veale argue, as we do, that the explanation debate had been restricted to this question. *Id.* Recognizing that explanations are no panacea, the rest of their paper argues that the GDPR provides tools other than a right to explanation that could be more useful for algorithmic accountability.

second question, which focuses on the many different methods by which outcomes can be explained.

An interdisciplinary working group at the Berkman Klein Center for Internet and Society begin by recognizing that explanations are infinitely variable in concept, but claim that "[w]hen we talk about an explanation for a decision, . . . we generally mean the reasons or justifications for that particular outcome, rather than a description of the decision-making process in general."[228] They propose three ways to examine a specific decision: (1) the main factors in a decision, (2) the minimum change required to switch the outcome of a decision, and (3) the explanations for similar cases with divergent outcomes or divergent cases with similar outcomes.[229] Wachter, Mittelstadt, and Chris Russell have a still narrower focus, writing about counterfactual explanations that represent "the smallest change to the world" that would result in a different answer.[230] They envision a distance metric where, if one were to plot all *n* features in an *n*-dimensional space, the counterfactual is the shortest "distance" from the data subject's point in the space (defined by the values of the features she possesses) to the surface that makes up the outer edge of a desirable outcome.[231]

Accordingly, counterfactual explanations are seen as fulfilling the three goals of explanations discussed in this Part: (1) to help an individual understand a decision, (2) to enable that individual to take steps to achieve a better outcome, and (3) to provide a basis for contesting the decision.[232] When applying the strategy of counterfactual explanations, however, it is clear that most of the value comes from the second rationale: actionable explanations. Wachter and colleagues assert that counterfactual explanations are an improvement over the existing requirements of the GDPR because, as a matter of positive law, the Regulation requires almost nothing except a "meaningful overview," which can be encapsulated via pictorial "icons" depicting the type of data processing in question.[233] Counterfactual explanations, in contrast, offer something specific to the data subject and will thus be more useful in informing an effective response. But if their interpretation of the law is correct—that the GDPR requires no

228. Doshi-Velez & Kortz, *supra* note 201, at 2.

229. *Id.* at 3.

230. Wachter et al., *supra* note 143, at 845.

231. *Id.* at 850–54. Distance metrics are a way to solve this problem. Hall and colleagues describe another distance metric that is used in practice. Hall et al., *supra* note 120. They employ a distance metric to identify the features that need to change the *most* to turn a credit applicant into the ideal applicant. *Id.* Alternatively, other methods could be identifying the features over which a consumer has the most control, the features that would cost a consumer the least to change, or the features least coupled to other life outcomes and thus easier to isolate. The main point is that the law provides no formal guidance as to the proper metric for determining what reasons are most salient, and this part of the debate attempts to resolve this question. *See* 12 C.F.R. § 1002.9 supp. I (2018).

232. Wachter et al., *supra* note 143, at 843.

233. *Id.* at 865.

explanation[234]—then their claim is that counterfactuals offer more than literally nothing, which is not saying much. On contestability, Wachter, Mittelstadt, and Russell ultimately concede that to contest a decision, it is likely necessary to understand the logic of decision-making rather than to just have a counterfactual explanation of a specific decision.[235] The real value, then, of their intervention and others like it, is to better allow data subjects to alter their behavior when a counterfactual suggests that a decision is based on alterable characteristics.[236]

Empowering people to navigate the algorithms that affect their lives is an important goal and has genuine value. This is a pragmatic response to a difficult problem, but it casts the goal of explanations as something quite limited: ensuring people know the rules of the game so they can play it better. This approach is not oriented around asking if the basis of decisions is well justified; rather it takes decisions as a given and seeks to allow those affected by them to avoid or work around bad outcomes.[237] Rather than using explanations to ask about the justifications for decision-making, this approach shifts responsibility for bad outcomes from the designers of automated decisions to those affected by them.[238]

3. Explanation as Exposing a Basis for Evaluation

The final value ascribed to explanation is that it forces the basis of decision-making into the open and thus provides a way to question the validity and justifiability of making decisions on these grounds. As Pauline Kim has observed:

---

234. The positive law debate about the right to explanation is not the subject of this Article, but suffice it to say, there is a healthy debate about it in the literature. See *supra* note 143 and accompanying text for a discussion.

235. Wachter et al., *supra* note 143, at 878. Their one example where a counterfactual can lead to the ability to contest a decision is based on data being inaccurate or missing rather than based on the inferences made. Thus, it is actually the rare situation specifically envisioned by FCRA, where the adverse action notice reveals that a decision took inaccurate information into account. Because of the deficiencies of the FCRA approach, discussed *supra* in Part II.A, this will not solve the general problem.

236. As Berk Ustun and colleagues point out, an explanation generated by counterfactual techniques will not necessarily be actionable unless intentionally structured to be so. Berk Ustun et al., Actionable Recourse in Linear Classification 2 (Sept. 18, 2018) (unpublished manuscript), https://arxiv.org/abs/1809.06514 [https://perma.cc/RPJ4-P4AP].

237. Mireille Hildebrandt, *Primitives of Legal Protection in the Era of Data-Driven Platforms*, 2 GEO. L. TECH. REV. 252, 271 (2018) ("Though it is important that decisions of automated systems can be explained (whether ex ante or ex post; whether individually or at a generic level), we must keep in mind that in the end what counts is whether such decisions can be justified.").

238. This is remarkably similar to the longstanding privacy and data protection debate around notice and consent, where the goal of notice is to better inform consumers and data subjects, and the assumption is that better information will lead to preferable results. *See generally* Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013). In reality, this often fails to protect privacy because it construes privacy as a matter of individual decision-making that a person can choose to protect rather than something that can be affected by others with more power. *See, e.g.*, Roger Ford, *Unilateral Invasions of Privacy*, 91 NOTRE DAME L. REV 1075 (2016).

> When a model is interpretable, debate may ensue over whether its use is justified, but it is at least possible to have a conversation about whether relying on the behaviors or attributes that drive the outcomes is normatively acceptable. When a model is not interpretable, however, it is not even possible to have the conversation.[239]

But what does it mean to have a conversation based on what an interpretable model reveals?

In a seminal study, Rich Caruana and colleagues provide an answer to that question.[240] They discovered that a model trained to predict complications from pneumonia had learned to associate asthma with a reduced risk of death.[241] To anyone with a passing knowledge of asthma and pneumonia, this result was obviously wrong. The model was trained on clinical data from past pneumonia patients, and it turns out that patients who suffer from asthma truly did end up with better outcomes.[242] What the model missed was that these patients regularly monitored their breathing, causing them to go to the hospital earlier.[243] Then, once at the hospital, they were considered higher risk, so they received more immediate and focused treatment.[244] Caruana and colleagues drew a general lesson from this experience: to avoid learning artifacts in the data, the model should be sufficiently simple that experts can inspect the relationships uncovered to determine if they correspond with domain knowledge. Thus, on this account, the purpose of explanation is to permit experts to check the model against their intuition.

This approach assumes that when a model is made intelligible, experts can assess whether the relationships uncovered by the model seem appropriate, given their background knowledge of the phenomenon being modeled. This was indeed the case for asthma, but this is not the general case. Often, rather than assigning significance to features in a way that is obviously right or wrong, a model will uncover a relationship that is simply perceived as strange. For example, if the hospital's data did not reveal a dependence on an asthma diagnosis—which is clearly linked to pneumonia through breathing—but rather revealed a dependence on skin cancer, it would be less obvious what to make of that fact. It would be wrong to simply dismiss it as an artifact of the data, but it also does not fit with any intuitive story even a domain expert could tell.

Another example of this view of explanation is the approach to interpretability known as Local Interpretable Model-Agnostic Explanations ("LIME").[245] It has generated one of the canonical examples of the value of

239. Kim, *supra* note 4, at 922–23.

240. Rich Caruana et al., *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, *in* PROCEEDINGS OF THE 21TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1721, 1721 (2015).

241. *Id.*

242. *Id.*

243. *Id.*

244. *Id.*

245. Ribeiro et al., *supra* note 197. This is one of the methods described *supra* in Part II.B.2.

interpretability in machine learning. Marco Ribeiro and colleagues used LIME to investigate a deep-learning model trained to distinguish images of wolves from huskies. The authors discovered that the model did not rely primarily on the animals' features, but on whether snow appeared in the background of a photo.[246]

There are three reasons this is such a compelling example. First, what LIME identified as the distinguishing feature—snow—is legible to humans. Second, this feature is obviously not a property of the category "wolf." Third, humans can tell a story about why this mistake occurred: wolves are more likely to be found in an environment with snow on the ground. Although this story may not actually be true, the important point is that we can convince ourselves it is.[247] Like the asthma example, the ability to determine that the model has overfit the training data relies on the inherent legibility of the relevant feature, the existence of background knowledge about that feature, and our ability to use the background knowledge to tell a story about why the feature is important. In this example, the realization relies on something closer to common sense than to specialized expertise, but the explanation serves the same function—to allow observers to bring their intuition to bear in evaluating the model.

The final examples come from James Grimmelmann and Daniel Westreich,[248] as well as Kim, whose work was discussed earlier.[249] Grimmelmann and Westreich imagine a scenario in which a model learns to distinguish between job applicants on the basis of a feature—musical taste—that is both correlated with job performance and membership in a protected class.[250] They further stipulate that job performance varies by class membership.[251] As they see it, this poses the challenge of determining whether the model, by relying on musical tastes, is in fact relying on protected-class membership.[252]

Grimmelmann and Westreich then argue that if one cannot tell a story about why musical taste correlates with job performance, the model must be learning something else.[253] They propose a default rule that the "something else" be considered membership in a protected class unless it can be shown

246. Ribeiro et al., *supra* note 197, at 1142–43. This is a textbook example of overfitting the training data.

247. In fact, while writing this section, we remembered the finding, but until we consulted the original source we disagreed with each other about whether the wolves or huskies were the ones pictured in snow. This suggests that the story would have been equally compelling if the error had been reversed.

248. Grimmelmann & Westreich, *supra* note 75.

249. Kim, *supra* note 4.

250. Grimmelmann & Westreich, *supra* note 75, at 166–67.

251. *Id.* at 167.

252. The only reason a model would learn to do this is if: (1) class membership accounts for all the variance in the outcome of interest or (2) class membership accounts for more of the variance than the input features. In the second case, the easy fix would be to include a richer set of features until class membership no longer communicates any useful information. The only way that adding features could have this effect, though, is if the original model was necessarily less than perfectly accurate, in which case a better model should have been used.

253. Grimmelmann & Westreich, *supra* note 75, at 174.

otherwise, specifically by the defendant.[254] The problem with this reasoning is that the model might not be learning protected-class membership, but a different latent variable that explains the relationship between musical taste and job performance—an unobserved or unknown characteristic that affects both musical taste and job performance. By assuming that it should be possible to tell a story about such a variable if it exists, they—as in the examples above—fail to account for the possibility of a strange, but legitimate, result. They use the ability to tell a story as a proxy for the legitimacy of the decision-making, but that only works if a justification, or lack thereof, immediately falls out of the description, as it did in the asthma and snow examples.

Kim uses a real example to make a similar point. She cites a study stating that employees who installed web browsers that did not come with their computers stay longer on their job.[255] She then speculates that either there is an unobserved variable that would explain the relationship or it is "entirely coincidental."[256] To Kim, what determines whether the relationship is "substantively meaningful" rather than a mere statistical coincidence is whether we can successfully tell ourselves such stories.[257] Like Grimmelmann and Westreich, for Kim, if no such story can be told, and the model has a disparate impact, it should be illegal.[258] What these examples demonstrate is that, whether one seeks to adjudicate model validity or normative justifications, intuition actually plays the same role.

Unlike the first two values of explanation, this approach has the ultimate goal of evaluating whether the basis of decision-making is well justified. It does not, however, ask the question: "Why are these the rules?" Instead, it makes two moves. The first two examples answered the question, "What are the rules?" and expected that intuition will furnish an answer for both why the rules are what they are and whether they are justified. The latter two examples instead argued that decisions should be legally restricted to intuitive relationships. Such a restriction short-circuits the need to *ask* why the rules are what they are by guaranteeing up front that an answer will be available.[259]

---

254. *Id.* at 173.

255. Kim, *supra* note 4, at 922.

256. *Id.* So too did the chief analytics officer in the company involved, in an interview. Joe Pinsker, *People Who Use Firefox or Chrome Are Better Employees*, ATLANTIC (Mar. 16, 2015), https://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/ [https://perma.cc/3MYM-SXAQ] ("'I think that the fact that you took the time to install Firefox on your computer shows us something about you. It shows that you're someone who is an informed consumer,' he told Freakonomics Radio. 'You've made an active choice to do something that wasn't default.'").

257. Kim, *supra* note 4, at 917.

258. *Id.*

259. This might also explain the frequent turn to causality as a solution. Restricting the model to causal relationships also short-circuits the need to ask the "why" question because the causal mechanism is the answer. Ironically, a causal model need not be intuitive, so it may not satisfy the same normative desires as intuition seems to. *See supra* note 78.

These two approaches are similar, but differ in the default rule they apply to strange cases. In the case of the two technical examples, the assumption is that obviously *flawed* relationships will present themselves and should be overruled; relationships for which there is no intuitive explanation may remain. The two legal examples, by contrast, are more conservative. They presume that obviously *correct* relationships will show themselves, so that everything else should be discarded by default, while allowing for the possibility of defeating such a presumption. Both are forced to rely on default rules to handle strange, but potentially legitimate, cases because the fundamental reliance on intuition does not give them tools to evaluate these cases.

B. Evaluating Intuition

Much of the anxiety around inscrutable models comes from the legal world's demands for justifiable decision-making. That decisions based on machine learning reflect the particular patterns in the training data cannot be a sufficient explanation for why a decision is made the way it is. Evaluating whether some basis for decision-making is fair, for example, will require tools that go beyond standard technical tests of validity that would already have been applied to the model during its development.[260] While the law gives these tests some credence, reliance on accuracy is not normatively adequate with respect to machine learning.[261]

For many, the presumed solution is requiring machine learning models to be intelligible.[262] What the prior discussion demonstrates, though, is that this presumption works on a very specific line of reasoning that is based on the idea that with enough explanation, we can bring intuition to bear in evaluating decision-making. As Kim observes:

> Even when a model is interpretable, its *meaning* may not be clear. Two variables may be strongly correlated in the data, but the existence of a statistical relationship does not tell us if the variables are causally related, or are influenced by some common unobservable factor, or are completely unrelated.[263]

---

260. Even among practitioners, the interest in interpretability stems from warranted suspicion of the power of validation; there are countless reasons why assessing the likely performance of a model against an out-of-sample test set will fail to accurately predict a model's real-world performance. Yet even with these deep suspicions, practitioners still believe in validation as the primary method by which the use of models can and should be justified. *See* Hand, *supra* note 182, at 12–13. In contrast, the law has concerns that are broader than real-world performance, which demand very different justifications for the basis of decision-making encoded in machine learning models.

261. Barocas & Selbst, *supra* note 4, at 673 ("[T]he process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.").

262. *See* Brennan-Marquez, *supra* note 19, at 1253; Grimmelmann & Westreich, *supra* note 75, at 173; Kim, *supra* note 4, at 921–22.

263. Kim, *supra* note 4, at 922.

Her response is to constrain the model to features that bear an intuitive relationship to the outcome.[264]

This way of thinking originates in disparate impact doctrine, which—among several ways of describing the requirement—calls for an employment test to have a "manifest relationship" to future job performance.[265] But there is a difference between a manifest relationship of a model to job performance and a manifest relationship of a particular *feature* to job performance. Models can be shown to have a manifest relationship to job performance if the *target variable* is manifestly related to job performance and the model is statistically valid. This is true even if none of the individual *features* are manifestly related.[266] People who advocate for a nexus between features and the outcome are dissatisfied with a purely statistical test and want some other basis to subject a model to normative assessment. Models must be restricted to intuitive relationships, the logic goes, so that such a basis will exist.

Regulatory guidance evinces similar reasoning. In 2011, the Federal Reserve issued formal guidance on model risk management.[267] The purpose of the document was to expand on prior guidance that was limited to model validation.[268] The guidance notes that models "may be used incorrectly or inappropriately" and that banks need diverse methods to evaluate them beyond statistical validation.[269] Among other recommendations discussed in Part IV, the guidance recommends "outcomes analysis," which calls for "expert judgment to check the intuition behind the outcomes and confirm that the results make sense."[270]

In an advisory bulletin about new financial technology, the Federal Reserve Board recommended that individual features have a "nexus" with creditworthiness to avoid discriminating in violation of fair lending laws.[271] In their view, a nexus enables a "careful analysis" about the features assigned

264. *Id.*; *cf.* Nick Seaver, *Algorithms as Culture*, BIG DATA & SOC'Y, July–Dec. 2017, at 6 ("To make something [accountable] means giving it qualities that make it legible to groups of people in specific contexts. An accountable algorithm is thus literally different from an unaccountable one—transparency changes the practices that constitute it. For some critics, this is precisely the point: the changes that transparency necessitates are changes that we want to have.").

265. Barocas & Selbst, *supra* note 4, at 702 ("A challenged employment practice must be 'shown to be related to job performance,' have a 'manifest relationship to the employment in question,' be 'demonstrably a reasonable measure of job performance, bear some relationship to job-performance ability,' []or 'must measure the person for the job and not the person in the abstract.'" (quoting Linda Lye, Comment, *Title VII's Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defense*, 19 BERKELEY J. EMP. & LAB. L. 315, 321 (1998) (footnotes omitted) (quoting Griggs v. Duke Power Co., 401 U.S. 424 (1971)))).

266. *Id.* at 708.

267. BD. OF GOVERNORS OF THE FED. RESERVE SYS., OFFICE OF THE COMPTROLLER OF THE CURRENCY, SR LETTER 11-7, SUPERVISORY GUIDANCE ON MODEL RISK MANAGEMENT (2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf [https://perma.cc/AR85-AASR].

268. *Id*. at 2.

269. *Id.* at 4.

270. *Id.* at 13–14.

271. Evans, *supra* note 121, at 4.

significance in a model predicting creditworthiness.[272] Here, intuitiveness is read into ECOA as a natural requirement of having to justify decision-making that generates a disparate impact via the "business-necessity" defense.[273] The business-necessity defense asks whether the particular decision-making mechanism has a tight enough fit with the legitimate trait being predicted[274] and whether there were equally effective but less discriminatory ways to accomplish the same task. With a model that lacks intuitive relationships, a plaintiff could argue that the model is indirectly—and thus poorly—measuring some latent and more sensible variable that should serve as the actual basis of decision-making. The Federal Reserve Board guidance suggests that one way to avoid an uncertain result in such litigation is to limit decision-making to features that bear an intuitive—and therefore justifiable—relationship to the outcome of interest. While it is not clear that relying on proxies for an unrecognized latent variable presents problems under current disparate impact doctrine,[275] the guidance treats an intuition requirement as a prophylactic. This reasoning seems to underlie the recommendations of Kim as well as Grimmelmann and Westreich.

What should be clear by now is that intuition is the typical bridge from explanation to normative assessment. This can be a good thing. Intuition is powerful. It is a ready mechanism by which considerable knowledge can be brought to bear in evaluating machine learning models. Such models are myopic, having visibility into only the data upon which they were trained.[276] Humans, in contrast, have a wealth of insights accumulated through a broad range of experiences, typically described as "common sense." This knowledge allows us to immediately identify and discount patterns that violate our well-honed expectations and to recognize and affirm discoveries that align with experience. In fact, intuition is so powerful that humans cannot resist speculating about latent variables or causal mechanisms when confronted by unexplained phenomena.

Intuition can also take the form of domain expertise, which further strengthens the capacity to see where models may have gone awry. The social sciences have a long history of relying on face validity to determine whether a model is measuring what it purports to measure.[277] A model that assigns significance to variables that seem facially irrelevant is given little credence or is subject to greater scrutiny. Such a practice might seem ad hoc, but questioning face validity is a fundamental part of the social-scientific

---

272. *Id.*

273. It is interesting that the demand for intuitiveness, on this account, comes not from the procedural requirements of the adverse action notices—the part of ECOA most obviously concerned with explanations—but from the substantive concerns of disparate impact doctrine.

274. *See, e.g.*, Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 1010 (1988) (Blackmun, J., concurring) (explaining that a business-necessity defense must be carefully tailored to objective, relevant job qualifications).

275. *See* Barocas & Selbst, *supra* note 4, at 709–10 (discussing the problems with the "fix-the-model" approach to alternative practice claims).

276. Andrew D. Selbst, *A Mild Defense of Our New Machine Overlords*, 70 VAND. L. REV. EN BANC 87, 101 (2017).

277. *See supra* note 73 and accompanying text.

process. Crucially, intuition allows us to generate competing explanations that account for the observed facts and to debate their plausibility.[278]

Importantly, however, intuition has its downsides. Most immediately, it can be wrong. It can lead us to discount valid models because they are unexpected or unfamiliar, or to endorse false discoveries because they align with existing beliefs.[279] Intuition encourages us to generate "just so" stories that appear to make good sense of the presented facts. Such stories may feel coherent but are actually unreliable. In fact, the rich literature on cognitive biases—including the "narrative fallacy"—is really an account of the dangers of intuition.[280] While intuition is helpful for assessing evidently good and bad results, it is less useful when dealing with findings that do not comport with or even run counter to experience. The overriding power of intuition means that strange results will stand out, but intuition may not point in a productive direction for making these any more sensible.

This is a particularly pronounced problem in the case of machine learning, as its value lies largely in finding patterns that go well beyond human intuition. The problem in such cases is not only that machine learning models might depart from intuition, but that they might not even lend themselves to *hypotheses* about what accounts for the models' discoveries. Parsimonious models lend themselves to more intuitive reasoning, but they have limits—a complex world may require complex models. In some cases, machine learning will have the power to detect the subtle patterns and intricate dependencies that can better account for reality.

If the interest in explanation stems from its intrinsic or pragmatic value, then addressing inscrutability is worthwhile for its own sake. But if we are interested in whether models are well justified, then addressing inscrutability only gets us part of the way. We should consider how else to justify models. We should think outside the black box and return to the question: Why are these the rules?

IV. DOCUMENTATION AS EXPLANATION

Limiting explanation of a model to its internal mechanics forces us to rely on intuition to guess at why the model's rules are what they are. But what would it look like for regulation to directly seek an answer to that question? By now, it is well understood that data are human constructs[281] and that subjective decisions pervade the modeling and decision-making process.[282]

278. *See, e.g.*, Brennan-Marquez, *supra* note 19; Michael Pardo & Ronald J. Allen, *Juridical Proof and the Best Explanation*, 27 LAW & PHIL. 223, 230 (2008).

279. Raymond S. Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, 2 REV. GEN. PSYCHOL. 175, 175 (1998).

280. *See generally* KAHNEMAN, *supra* note 77.

281. Lisa Gitelman & Virginia Jackson, *Introduction* to RAW DATA IS AN OXYMORON 1, 3 (Lisa Gitelman ed., 2013); *see also* danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO., COMM. & SOC'Y 662, 666–68 (2012).

282. Barocas & Selbst, *supra* note 4, at 673; *see also* Seaver, *supra* note 264, at 5.

Explaining why the model works as it does requires accounting for these decisions.

Furnishing such answers will require process, documentation, and access to that documentation. This can be done in a public format, with impact assessments, or companies can do it privately, with access triggered on some basis, like discovery in litigation.

A. The Information Needed to Evaluate Models

When we seek to evaluate the justifications for decision-making that relies on a machine learning model, we are actually asking about the institutional and subjective process behind its development. The Federal Reserve Board guidance discussed in Part III.B moves in this direction by recommending documentation, but its approach appears to be mostly about validation—how to validate well, thoroughly, on an ongoing basis, and in preparation for a future legal challenge.[283] Careful validation is essential and nontrivial,[284] but it is also not enough. Normatively evaluating decision-making requires, at least, an understanding of: (1) the values and constraints that shape the conceptualization of the problem, (2) how these values and constraints inform the development of machine learning models and are ultimately reflected in them, and (3) how the outputs of models inform final decisions.

To illustrate how each of these components work, consider credit scoring. What are the values embedded in credit-scoring models and under what constraints do developers operate? Lenders could attempt to achieve different objectives with credit scoring at the outset: Credit scoring could aim to ensure that all credit is ultimately repaid, thus minimizing default. Lenders could use credit scoring to maximize profit. Lenders could also seek to find ways to offer credit specifically to otherwise overlooked applicants, as many firms engaged in alternative credit scoring seek to do. Each of these different goals reflects different core values, but other value judgments might be buried in the projects as well. For example, a creditor could be morally committed to offering credit as widely as possible, while for others that does not factor into the decision. Or a creditor's approach to regulation could be to either get away with as much as possible or steer far clear of regulatory scrutiny. Each of these subjective judgments will ultimately inform the way a project of credit scoring is conceived.

The developers of credit-scoring models will also face constraints and trade-offs. For example, there might be limits on available talent with both domain expertise and the necessary technical skills to build models. Models might be better informed if there were much more data available, even though

283. BD. OF GOVERNORS OF THE FED. RESERVE SYS., *supra* note 267. The guidance wants developers to consider where the data comes from, whether it suffers from bias, whether the model is robust to new situations, whether due care has been taken with respect to potential limitations and outright faults with the model, and so on. *Id.* at 5–16; *see also* Edwards & Veale, *supra* note 143, at 55–56; Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 196 (2017).

284. Barocas & Selbst, *supra* note 4, at 680–92.

there are practical challenges to collecting so much data. Ultimately, both trade-offs are issues of cost,[285] but they include more practical realities as well, such as limitations on talent in the geographical area of the firm or privacy concerns that limit the collection of more data. How to deal with these trade-offs is a judgment call every firm will have to make.[286]

Another cost-related trade-off is competition. Before credit scoring was popular, creditors used to work with borrowers over the lifetime of the loan to ensure repayment; credit scores first took hold in banks as a way to reduce the cost of this practice.[287] Creditors today *could* return to that model, but it would likely involve offering higher interest rates across the board to account for increased operating costs, perhaps pushing such a firm out of the market. As a result, competition operates as a constraint that ultimately changes the decision process.

The values of and constraints faced by a firm will lead to certain choices about how to build and use models. As we have discussed in prior work, the subjective choices a developer makes include choosing target variables, collecting training data, labeling examples, and choosing features.[288] Developers must also make choices about other parts of the process, such as how to treat outliers, how to partition their data for testing, what learning algorithms to choose, and how and how much to tune the model, among other things.[289] The act of developing models is quite complex and involves many subjective decisions by the developers.

In the credit example, the values discussed above may manifest in the model in several ways. For example, consider the different project objectives discussed above. If a firm seeks to maximize profit, it may employ a model with a different target variable than a firm that seeks to minimize defaults. The target variable is often the outcome that the model developers want to maximize or minimize, so in the profit-seeking case, it would be *expected profit per applicant*, and in the risk-based case, it could be *likelihood of default*. While the alternative credit-scoring model hypothesized above might rely on the same likelihood-of-default target variable, firms' values are likely to influence the type of data they collect; they might seek alternative data sources, for example, because they are trying to reach underserved populations. In addition to the values embedded a priori, the values of the firms dictate how they resolve the different constraints they face—for example, cost and competition. The traditional credit scorers tend to not

285. *See* FREDERICK SCHAUER, PROFILES, PROBABILITIES, AND STEREOTYPES 124–26 (2003).

286. Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 998 (1988) (plurality opinion) (considering costs and other burdens relevant to a discrimination case).

287. Martha Ann Poon, What Lenders See—a History of the Fair Isaac Scorecard 109, 120 (Jan. 1, 2012) (unpublished Ph.D. dissertation, University of California, San Diego), https://cloudfront.escholarship.org/dist/prd/content/qt7n1369x2/qt7n1369x2.pdf?t=o94tcd [https://perma.cc/V24B-8G3M].

288. Barocas & Selbst, *supra* note 4, at 677–92.

289. Lehr & Ohm, *supra* note 51, at 683–700; *see also* Brian d'Alessandro, Cathy O'Neil & Tom LaGatta, *Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification*, 5 BIG DATA 120, 125 (2017).

make the extra effort or spend the extra money to obtain the data needed to make predictions about people on the margins of society.[290] There is also regulatory uncertainty regarding the permissibility of new types of credit data.[291] Therefore, their models reflect the fact that the developers are more sensitive to cost and regulatory penalty than inclusion.

Models are not self-executing; an additional layer of decisions concerns the institutional process that surrounds the model. Are the model outputs automatically accepted as the ultimate decisions?[292] If not, how central is the model to the decision? How do decision makers integrate the model into their larger decision frameworks? How are they trained to do so? What role does discretion play?

These questions are all external to the model, but they directly impact the model's importance and normative valence. For example, certain creditors may automatically reject applicants with a predicted likelihood of default that exceeds 50 percent.[293] Others, however, may opt to be more inclusive. Perhaps a local credit union that is more familiar with its members and has a community-service mission might decide that human review is necessary for applicants whose likelihood of default sits between 40 percent and 60 percent, leaving the final decision to individual loan officers. A similar creditor might adopt a policy where applicants that the model is not able to score with confidence are subject to human review, especially where the outcome would otherwise be an automatic rejection of members of legally protected classes.

Many of these high-level questions about justifying models or particular uses of models are not about models at all, but whether certain policies are

290. Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, 82 Fed. Reg. 11,183, 11,185 (Feb. 21, 2017).

291. *Id.* at 11,187–88.

292. The distinction between models and ultimate decisions is the focus of the GDPR's prohibition on "decision[s] based solely on *automated* processing." Article 29 Data Protection Working Party, *supra* note 149, at 19–22 (emphasis added).

293. This is not how credit typically works in the real world, but for demonstrative purposes, we decided to work with a single hypothetical. In reality, the best examples of this divergence between model and use come from policing and criminal justice. For example, the predictive-policing measure in Chicago, known as the Strategic Subject List, was used to predict the 400 likeliest people in a year to be involved in violent crime. Monica Davey, *Chicago Police Try to Predict Who May Shoot or Be Shot*, N.Y. TIMES (May 23, 2016), http://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html [https://perma.cc/TZ2T-NMEJ]. When Chicago sought funding for the initiative, the city premised it on the idea of providing increased social services to those 400 people, but in the end only targeted them for surveillance. DAVID ROBINSON & LOGAN KOEPKE, STUCK IN A PATTERN: EARLY EVIDENCE ON "PREDICTIVE POLICING" AND CIVIL RIGHTS 9 (2016). The fairness concerns are clearly different between those use cases. *See* Selbst, *supra* note 4, at 142–44. Similarly, COMPAS, the now-infamous recidivism risk score, was originally designed to figure out who would need greater access to social services upon reentry to reduce the likelihood of rearrest but is now commonly used to decide whom to detain pending trial. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/F9CK-Z995].

acceptable independent of whether they use machine learning.[294] Questions about justifying a model are often just questions about policy in disguise.[295] For example, a predatory lender could use the exact same prediction of default to find prime candidates in underserved communities and offer them higher interest rates than they might otherwise receive. This will create more profit because the underserved loan candidates will be more willing to pay a higher rate, but it is clearly predation: interest rates are not being used to offset risk, but to extract maximum profit from vulnerable consumers.[296] Most importantly, that this practice is predatory can be judged with no reference to the credit-scoring model.

Evaluating models in a justificatory sense means comparing the reasoning behind the choices made by the developers against society's broader normative priorities, as expressed in law and policy. In order to perform this evaluation, then, documentation about the decisions that lie behind and become part of models must exist and be made available for scrutiny. With an understanding of what that information looks like, the next section begins to explore how to ensure access.

B. Providing the Necessary Information

Assuming the documentation exists, there are numerous ways it can become open to scrutiny. For purposes of demonstration, two are discussed here, although many more are possible: (1) the possibility that documentation is made publicly available from the start and (2) that it becomes accessible upon some trigger, like litigation. The former is essentially an algorithmic impact statement (AIS),[297] a proposed variant of the original impact statements required by the National Environmental Policy Act.[298] The most common trigger of the latter is a lawsuit, in which documents can be obtained and scrutinized and witnesses can be deposed or examined on the stand, but auditing requirements are another possibility. In both approaches, the coupling of existing documentation with a way to access it create answers to the question of what happened in the design process, with the goal of allowing overseers to determine whether those choices were justifiable. Like FCRA and ECOA, these examples have no inherent

294. *See* VIRGINIA EUBANKS, AUTOMATING INEQUALITY 37 (2018) ("[W]hen we focus on programs specifically targeted at poor and working-class people, the new regime of data analytics is more evolution than revolution. It is simply an expansion and continuation of moralistic and punitive poverty management strategies that have been with us since the 1820s.").

295. *See, e.g.*, *id.* at 38; Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 99–101 (2017); Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633 (2017); Sonia Katyal, *Algorithmic Civil Rights*, 104 IOWA L. REV. (forthcoming 2018) (draft on file with authors); Sandra G. Mayson, *Dangerous Defendants*, 127 YALE L.J. 490, 507–18 (2018).

296. According to sociologist Jacob Faber, this is actually what happened in the subprime crisis to people of color. Jacob W. Faber, *Racial Dynamics of Subprime Mortgage Lending at the Peak*, 28 HOUSING POL'Y DEBATE 328, 343 (2013).

297. Selbst, *supra* note 4, at 169–93.

298. *See* 42 U.S.C. § 4332(C) (2012).

connection to machine learning, but the methods can be easily applied in this context.

An impact statement is a document designed to explain the process of decision-making and the anticipated effects of that decision in such a way as to open the process up to the public. Generally, the requirement is designed to ensure that developers do their homework, create a public record, and include public comments.[299] Impact statements are an idea that originated in 1970 with the National Environmental Policy Act[300] and have since been emulated repeatedly at all levels of government, in many substantive areas of policy.[301] Aside from environmental law, the federal government requires privacy impact assessments "when developing or procuring information technology systems that include personally identifiable information."[302] Individual states not only have their own legislation requiring environmental impact statements,[303] but also racial impact statements for sentencing policy, among other requirements.[304] Recently, led by the ACLU's Community Control Over Police Surveillance (CCOPS) initiative,[305] counties and cities have begun requiring impact statements that apply to police purchases of new technology.[306]

One of us has argued that a future AIS requirement should be expressly modeled on the environmental impact statement (EIS): the original and most thorough version, with the fullest explanation requirements. Such an impact statement would require thoroughly explaining the types of choices discussed above. This includes direct choices about the model, such as target variables, whether and how new data was collected, and what features were considered. It also requires a discussion of the options that were considered but not chosen, and the reasons for both.[307] Those reasons would—either explicitly or implicitly—include discussion of the practical constraints faced by the developers and the values that drove decisions. The AIS must also discuss the predicted impacts of both the chosen and unchosen paths, including the

299. Selbst, *supra* note 4, at 169.

300. 42 U.S.C. §§ 4321–4347.

301. Bradley C. Karkkainen, *Toward a Smarter NEPA: Monitoring and Managing Government's Environmental Performance*, 102 COLUM. L. REV. 903, 905 (2002).

302. Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy Decision-Making in Administrative Agencies*, 75 U. CHI. L. REV. 75, 76 (2008).

303. *E.g.*, California Environmental Quality Act (CEQA), CAL. PUB. RES. CODE §§ 21000–21178 (2018).

304. Jessica Erickson, Comment, *Racial Impact Statements: Considering the Consequences of Racial Disproportionalities in the Criminal Justice System*, 89 WASH. L. REV. 1425, 1445 (2014).

305. AN ACT TO PROMOTE TRANSPARENCY AND PROTECT CIVIL RIGHTS AND CIVIL LIBERTIES WITH RESPECT TO SURVEILLANCE TECHNOLOGY § 2(B) (ACLU Jan. 2017), https://www.aclu.org/files/communitycontrol/ACLU-Local-Surveillance-Technology-Model-City-Council-Bill-January-2017.pdf [https://perma.cc/AQ8T-3NKM] (ACLU CCOPS Model Bill).

306. *See, e.g.*, SANTA CLARA COUNTY, CAL., CODE OF ORDINANCES § A40-3 (2016).

307. Selbst, *supra* note 4, at 172–75.

possibility of no action, and the effects of any potential mitigation procedures.[308]

The typical American example of an impact statement is a public document. Thus, a law requiring them would also require that the developers publish the document and allow for comments between the draft and final impact statements.[309] Of course, such an idea is more palatable in the case of regulation of public agencies. While disclosure of the kinds of information we describe does not actually imply disclosure of the model itself—obviating the need for a discussion of trade secrets and gaming—firms may still be reluctant to publish an AIS that reveals operating strategy, perceived constraints, or even embedded values. Thus, it is also useful to consider a documentation requirement that allows the prepared documents to remain private but available as needed for accountability.[310]

A provision of the GDPR actually does just this. Article 35 requires "data protection impact assessments" (DPIAs) whenever data processing "is likely to result in a high risk to the rights and freedoms of natural persons."[311] As Edwards and Veale discuss, the DPIA requirement is very likely to apply to machine learning,[312] and the assessments require "appropriate technical and organizational measures" to protect data subject rights.[313] In Europe, DPIAs are private documents, though making summaries public is officially encouraged.[314] The European solution to making this private document available is to require consultation with the member state data protection authorities whenever the DPIA indicates a high risk of interference with data subject rights.[315]

One could imagine another way of making an essentially private impact assessment accessible, initiated by private litigation. Interrogatories, depositions, document subpoenas, and trial testimony are all tools that enable litigation parties to question human witnesses and examine documents. These are all chances to directly ask model developers what choices they made and why they made them.

A hypothetical will help clarify how these opportunities, coupled with documentation—whether a DPIA or something similar—differ from the use of intuition as a method of justification. Imagine a new alternative credit-scoring system that relies on social media data.[316] This model assigns

---

308. *Id.*

309. *Id.* at 177.

310. *See* W. Nicholson Price, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 435–37 (2017).

311. GDPR, *supra* note 12, art. 35.

312. Edwards & Veale, *supra* note 143, at 77–78.

313. GDPR, *supra* note 12, art. 35.

314. Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679*, at 18, WP 248 (Apr. 4, 2017).

315. Edwards & Veale, *supra* note 143, at 78.

316. *See, e.g.*, Astra Taylor & Jathan Sadowski, *How Companies Turn Your Facebook Activity into a Credit Score*, NATION (May 27, 2015), https://www.thenation.com/article/how-companies-turn-your-facebook-activity-credit-score/ [https://perma.cc/P9FW-DSTN].

significance to data points that are unintuitive but reliably predict default. Suppose the model also evinces a disparate impact along racial lines, as revealed by investigative journalists.

Black applicants denied credit then bring suit under the substantive nondiscrimination provisions of ECOA. Assuming, reasonably, that the judge agrees that disparate impact is a viable theory under ECOA,[317] the case will turn on the business-necessity defense. Thus, in order to determine whether there was a legal violation, it is necessary to know why the designer of the model proceeded in using the particular features from social media and whether there were equally effective alternatives with less disparate impact.

Under an intuition-driven regime, such as that proposed by either Kim or Grimmelmann and Westreich, the case would begin with a finding of prima facie disparate impact, and then, to evaluate the business-necessity defense, the plaintiffs might put the lead engineer on the stand. The attorney would ask why social media data was related to the ultimate judgment of creditworthiness. The engineer would respond that the model showed they were related: "the data says so." She is not able to give a better answer because the social media data has no intuitive link to creditworthiness.[318] Under their proposed regime, the inquiry would end. The defendant has not satisfied its burden and would be held liable.[319]

Under a regime of mandated documentation and looking beyond the logic of the model, other explanations could be used in the model's defense. Rather than be required to intuitively link the social media data to the creditworthiness, the engineer would be permitted to answer why the model relies on the social media data in the first place. The documentation might show, or the engineer might testify, that her team tested the model with and without the social media data and found that using the data reduced the disproportionate impact of the model.[320] Alternatively, the documentation might demonstrate that the team considered more intuitive features that guaranteed similar model performance but discovered that such features were exceedingly difficult or costly to measure. The company then used social media data because it improved performance and reduced disparate impact under the practical constraints faced by the company.

---

317. *See* CONSUMER FIN. PROT. BUREAU, CFPB BULL. 2012-04 (FAIR LENDING), LENDING DISCRIMINATION 2 (2012), https://files.consumerfinance.gov/f/201404_cfpb_bulletin_lending_discrimination.pdf [https://perma.cc/M42R-W9J7].

318. The engineer might have been able to come up with a story for why social media relates to credit—perhaps many of the applicant's friends have low credit scores and the operating theory is that people associate with others who have similar qualities—and under this regime, such a story might have satisfied the defense. But the engineer knows this is a post hoc explanation that may bear little relationship to the actual dynamic that explains the model.

319. Grimmelmann & Westreich, *supra* note 75, at 170.

320. In fact, a recent Request for Information by the Consumer Financial Protection Bureau seems to anticipate such a claim. Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, 82 Fed. Reg. 11,183, 11,185–86 (Feb. 21, 2017).

These justifications are not self-evidently sufficient to approve of the credit model in this hypothetical. Certainly, reducing disparate impact seems like a worthwhile goal. In fact, prohibiting or discouraging decision makers from using unintuitive models that exhibit any disparate impact may have the perverse effect of maintaining a disparate impact. Cost is a more difficult normative line[321] and would likely require a case-by-case analysis. While intuition-based evaluation—and its reliance on default rules—would forbid the consideration of either of these motivations for using social media data, both rationales should at least enter into the discussion.[322]

Having to account for all the decisions made in the process of project inception and model development should reveal subjective judgments that can and should be evaluated. This kind of explanation is particularly useful where intuition fails. In most cases, these decisions would not be immediately readable from the model.[323] Recall that intuition is most useful where explanations of a model reveal obviously good or bad reasons for decision-making but will often offer no help to evaluate a strange result. Documentation will help because it provides a different way of connecting the model to normative concerns. In cases where the individual features are not intuitively related to the outcome of interest but there is an obviously good or bad reason to use them anyway, documentation will reveal those reasons where explanation of the model will not. Accordingly, these high-level explanations are a necessary complement to any explanation of the internals of the model.

Documentation will not, however, solve every problem. Even with documentation, some models will both defy intuition and resist normative clarity. Regardless, a regime of documentation leaves open the possibility of developing other ways of asking whether this was a well-executed project, including future understanding of what constitutes best practice. As common flaws become known, checking for them becomes simply a matter of being responsible. A safe harbor or negligence-based oversight regime may emerge or become attractive as the types of choices faced by firms become known and standardized.[324] Documentation of the decisions made will be necessary to developing such a regime.

---

321. *See generally* Ernest F. Lidge III, *Financial Costs as a Defense to an Employment Discrimination Claim*, 58 ARK. L. REV. 1 (2006).

322. Documentation provides a further benefit unrelated to explanation. If the requirement for an intuitive link is satisfied, then the case moves to the alternative practice prong, which looks to determine whether there was another model the creditor "refuses" to use. *Cf.* 42 U.S.C. § 2000e-2(k)(1)(A)(ii) (2012). Normally, a "fix-the-model" response will not be persuasive because it is difficult to tell exactly how it went wrong, and what alternatives the developers had. Barocas & Selbst, *supra* note 4, at 705. With documentation, the alternatives will be plainly visible because that is exactly what has been documented.

323. Barocas & Selbst, *supra* note 4, at 715.

324. *See generally* William Smart, Cindy Grimm & Woody Hartzog, An Education Theory of Fault for Autonomous Systems (Mar. 22, 2017) (unpublished manuscript), http://www.werobot2017.com/wp-content/uploads/2017/03/Smart-Grimm-Hartzog-Education-We-Robot.pdf [https://perma.cc/6WJM-4ZQH].

While there will certainly still be strange results for which neither intuition nor documentation works today, the overall set of cases we cannot evaluate will shrink considerably with documentation available.

CONCLUSION

Daniel Kahneman has referred to the human mind as a "machine for jumping to conclusions."[325] Intuition is a basic component of human reasoning, and reasoning about the law is no different. It should therefore not be surprising that we are suspicious of strange relationships in models that admit no intuitive explanation at all. The natural inclination at this point is to regulate machine learning such that its outputs comport with intuition.

This has led to calls for regulation by explanation. Inscrutability is the property of machine learning models that is seen as the problem, and the target of the majority of proposed remedies. The legal and technical work addressing the problem of inscrutability has been motivated by different beliefs about the utility of explanations: inherent value, enabling action, and providing a way to evaluate the basis of decision-making. While the first two rationales may have their own merits, the law has more substantial and concrete concerns that must be addressed. Those who believe solving inscrutability provides a path to normative evaluation also fall short because they fail to recognize the role of intuition.

Solving inscrutability is a necessary step, but the limitations of intuition will prevent normative assessment in many cases. Where intuition fails, the task should be to find new ways to regulate machine learning so that it remains accountable. Otherwise, maintaining an affirmative requirement for intuitive relationships will potentially impede discoveries and opportunities that machine learning can offer, including those that would reduce bias and discrimination.

Just as restricting evaluation to intuition will be costly, so would abandoning it entirely. Intuition serves as an important check that cannot be provided by quantitative modes of validation. But while there will always be a role for intuition, we will not always be able to use it to bypass the question of why the rules are the rules. We need the developers to show their work.

Documentation can relate the subjective choices involved in applying machine learning to the normative goals of substantive law. Much of the discussion surrounding models implicates important policy discussions, but does so indirectly. Often, when models are employed to change a way of making decisions, too much focus is placed on the technology itself instead of the policy changes that either led to the adoption of the technology or were wrought by its adoption.[326] Quite aside from correcting one failure mode of intuition, documentation has a separate worth in laying bare the kinds of value judgments that go into designing these systems and allowing society to engage in a clearer normative debate in the future.

---

325. KAHNEMAN, *supra* note 77, at 185.
326. *See generally* EUBANKS, *supra* note 294.

We cannot and should not abandon intuition. But only by recognizing the role intuition plays in our normative reasoning can we recognize that there are other ways. To complement intuition, we need to ask whether people have made reasonable judgments about competing values under their real-world constraints. Only humans can answer these questions.